

Modelling the Dynamics of the Limit Order Book in Financial Markets

By:

Kylie-Anne RICHARDS

B.Sc., The University of Melbourne
B.Com., The University of Melbourne
M.Fin., The University of Hong Kong

A thesis in fulfillment of the requirements for the degree of
Doctor of Philosophy



UNSW
SYDNEY

School of Mathematics and Statistics
Faculty of Science

February 11, 2019

ORIGINALITY STATEMENT

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

Signed *Richard*

Date *31-August-2018*

COPYRIGHT STATEMENT

'I hereby grant the University of New South Wales or its agents the right to archive and to make available my thesis or dissertation in whole or part in the University libraries in all forms of media, now or here after known, subject to the provisions of the Copyright Act 1968. I retain all proprietary rights, such as patent rights. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

I also authorise University Microfilms to use the 350 word abstract of my thesis in Dissertation Abstract International (this is applicable to doctoral theses only).

I have either used no substantial portions of copyright material in my thesis or I have obtained permission to use copyright material; where permission has not been granted I have applied/will apply for a partial restriction of the digital copy of my thesis or dissertation.'

Signed *Richard*

Date *31-August-2018*

AUTHENTICITY STATEMENT

'I certify that the Library deposit digital copy is a direct equivalent of the final officially approved version of my thesis. No emendation of content has occurred and if there are any minor variations in formatting, they are the result of the conversion to digital format.'

Signed *Richard*

Date *31-August-2018*

To Thomas and Ella

Contents

Acknowledgements	X
Abstract	XII
List of figures	XXV
List of tables	XXX
1 Introduction and literature review	1
1.1 Context	1
1.2 Derivative asset classes and market specificities	5
1.2.1 Exchange venues	5
1.2.2 Futures	7
1.3 The limit order book	9
1.4 Models for the limit order book	14
1.4.1 Point processes for the limit order book	15
1.5 Unmarked Hawkes processes	17
1.5.1 Event times	19
1.5.2 Hawkes processes for macroeconomic applications	20
1.5.3 Hawkes processes for market reflexivity applications	22
1.5.4 Hawkes processes for optimal execution, price impact and order-flow applications	25
1.5.5 Hawkes processes for market microstructure applications	29
1.5.6 Summary	32
1.6 Marked Hawkes processes	33
1.7 Foundation for marks identification	38
1.7.1 Summary	41
1.8 Theoretical properties of the Hawkes process	43
1.9 Contributions	44
1.10 Organization of the thesis	45
2 Statistical properties of the limit order book volume process	47
2.1 Data specifics	49
2.1.1 Liquid market hours	49
2.2 Empirical analysis of limit order book volumes	51

2.2.1	Shape of limit order book volume profiles	52
2.2.2	Assessing long range dependence	54
2.2.3	Empirical evidence for heavy tails in limit order book volumes	61
2.3	Statistical estimation methods for models of the limit order book volume process	65
2.3.1	α -stable distribution parameter estimation	66
2.3.2	Generalized extreme value parameter estimation	68
2.3.3	Generalized Pareto distribution parameter estimation	70
2.3.4	Goodness-of-fit testing for heavy tailed models of the limit order book volume process	75
2.4	Results and discussions on model estimation for the limit order book volume process	76
2.4.1	α -stable model estimation results	77
2.4.2	Generalized extreme value model estimation results	79
2.4.3	Generalized Pareto distribution model estimation results	80
2.5	Goodness-of-Fit via variance weighted modified Kolmogorov-Smirnov test	82
2.6	Conclusion	84
3	Construction of the observed limit order book and event process	86
3.1	Physical, reported and observed data	86
3.1.1	The physical limit order book	86
3.1.2	The reported limit order book	87
3.1.3	The observed limit order book	89
3.1.4	Assets and data attributes	98
3.2	Event processes	99
3.2.1	Empirical studies of event processes	100
3.3	Conclusion	109
4	Defining the Hawkes process	111
4.1	Defining a univariate Hawkes self-exciting marked point process	111
4.1.1	Copula models for distributions of the marks	113
4.1.2	Quasi-likelihood	114
4.1.3	Residual process	116
4.2	Numerical algorithms	116
4.2.1	Estimation of the Hawkes process	116
4.2.2	Simulation algorithm for the univariate Hawkes process	121
4.2.3	Goodness-of-fit algorithms	123
4.2.4	Computational considerations	123
4.3	Software for modelling the Hawkes process	125
4.4	Optimization: Hessian and bootstrapping	127
4.5	Simulation Experiments	129
4.5.1	Case Study 1: Hawkes process with a constant boost function	131

4.5.2	Case Study 2: Hawkes process with a univariate i.i.d. mark with a range of parametric distributions	134
4.5.3	Case Study 3: Hawkes process with independent bivariate marks with a range of parametric distributions	139
4.5.4	Case Study 4: Hawkes process with marks in higher dimension . . .	144
4.5.5	Case Study 5: Hawkes process with jointly coupled marks	145
4.6	Conclusion	154
5	Modelling the mark random vector	157
5.1	A discussion of potential marks	158
5.2	Defining the mark vector	162
5.2.1	Volume based marks	162
5.2.2	Volume imbalance marks	164
5.2.3	Price based marks	164
5.2.4	Count based marks	165
5.3	Properties of the mark vector	166
5.3.1	Statistical properties	166
5.3.2	Methods of analysis	168
5.3.3	Simulating a Hawkes process with serial dependences	170
5.3.4	Case studies and categorization	173
5.3.5	Extended studies across time	184
5.4	Extensions to boost functions for partially observed marks and other variants	190
5.5	Conclusion	191
6	The score test for detecting marks	193
6.1	Log-likelihood, score and information	195
6.2	The score test	197
6.2.1	Boost function details	197
6.2.2	Definition of the score test	198
6.2.3	Implementation of the score statistic	200
6.3	Asymptotic distribution of the score statistic	202
6.4	Stationary serially dependent marks	204
6.5	Simulation methodology	207
6.6	Simulation Experiments with i.i.d. Marks	209
6.6.1	Size and power of the score test	209
6.6.2	Comparison of empirical and estimated theoretical moments of the marks distribution	213
6.6.3	Robustness against moments not existing for generalized Pareto distributed marks	215
6.7	Simulation Experiments with Dependence	217
6.7.1	Impact of ignored serial dependence in marks	218
6.7.2	Impact of ignored joint dependence in marks	219
6.8	Simulation Experiments with Extensions	221

6.8.1	Impact of increasing serial dependence with heavy tailed marks . . .	222
6.8.2	A breakdown in power	225
6.8.3	Extended robustness tests of moments with increasing serial dependence	228
6.8.4	The discrete case with increasing serial dependence	231
6.8.5	Power properties of the score test for high dimensional marks	233
6.8.6	Impact of increasing the joint dependency between bivariate i.i.d. marks	234
6.8.7	Impact of increasing the joint dependency between bivariate serially dependent marks	237
6.9	Conclusion	239
7	Score test application and the decoupled approximate likelihood method	241
7.1	Score test application to real data	242
7.1.1	Selection of intra-day time segments	243
7.1.2	Score test of pairwise marks	248
7.1.3	Score test of higher dimensional marks	249
7.2	Fitting marked Hawkes process models	250
7.2.1	Modelling a Hawkes process with a one dimensional mark	251
7.2.2	Modelling a Hawkes process with two dimensional marks	259
7.2.3	A decoupled approximate likelihood method	266
7.3	Conclusion	272
8	Conclusion	274
8.1	Summary and contributions	274
8.1.1	The limit order book volume process	274
8.1.2	Accurately describing the limit order book and identifying marks . .	275
8.1.3	A score test for the detection of marks	275
8.1.4	Hawkes process with multivariate marks for the limit order book . .	276
8.2	Future work	276
	Bibliography	278
A	Chapter 4	292
A.1	Copula models for distributions of marks	292
A.1.1	Gaussian Copula	292
A.1.2	Archimedean Copula	293
A.2	Moments	296
A.3	Simulation algorithm for the univariate Hawkes process for both recursive and non-recursive procedures	297
A.3.1	Simulation study to verify that the recursive method matches the non-recursive method in the case of exponential decay	300

B Chapter 6	302
B.1 Score Test Derivation Details	302
B.2 Properties of decay functions	305

Acknowledgements

I would like to express my deepest gratitude to my supervisors, Professor William Dunsmuir and Professor Gareth Peters, who have been tremendous mentors, providing the support and encouragement to complete this work. It has been a great privilege to study under the guidance of two exceptionally talented academics, who inspire a genuine pursuit of knowledge. Professor William Dunsmuir taught me the joys of mastery, possessing a patience that truly knows no end. This thesis would not have been possible without his excellent intellectual inputs, scientific rigour, enthusiasm, pragmatism, generosity of time, and the care with which he oversaw the work. Professor Gareth Peters' unmatched breadth and depth of knowledge provided both financial and statistical insights into advancing this research. He provided unflinching belief and encouragement, which was the catalyst in pursuing higher standards. I am grateful for the many hours of detailed review that they both committed during the final stages of writing. I will always be profoundly grateful to have been afforded this opportunity, and I could not have wished for more talented and friendly supervisors.

This thesis has been supported by *Boronia Capital*. I would like to thank Boronia for their financial support, high quality financial data sets and a place to research for a number of years. A special thanks to Dr Chris Mellon and Mr Russell Grew for their enthusiasm and willingness to answer my many questions.

The School of Mathematics and Statistics has provided an excellent academic environment. Some faculty members of the school who have extended their time during various phases of my research are, Dr Diana Combe, Professor David Warton and Dr Feng Chen. I was proud to attend the postgraduate conferences each year and bear witness to the talented researchers I had the privilege to be amongst.

This thesis would not have been possible without my family and many friends who have supported me on this journey. My dear friend and mentor David Harris, I would never have embarked on such an ambitious undertaking without your support. I will always be humbly grateful for your kindness. Ian Heddle, your belief in me far exceeds what I deserve. David Jame, an all encompassing thank you - who would have thought first year 'Statistics for Life Science' would end up here! I am humbled by the support and encouragement (and humour) in the final six months from David J., Jude, Maya, Laura T., Naomi, Jess, and Lisa - what a lead out team to the finish! A particular thanks to: Karen, Carrie, Chris, Heath, Elke, Pia, Andrew, Laura A., Damian W., Manzoor and the Athletics East crew.

To my brothers Steve and Paul, thanks for toughening me up! Mum, you lit the fire to strive for more. To the kindest and most selfless person I know, my dad, you epitomise integrity, hard work and patience. Thanks for teaching me to ‘get-back-on-the-horse’, metaphorically and literally.

Dear Thomas and Ella, Thomas, I started this mammoth undertaking when you were four months old, Ella you arrived less than two years later. Throughout my own personal journey I have had the privilege of observing your own determined learning, your curiosity of all things, and your mastery of new skills. These elements have reminded me daily of what is crucial for success in research (- and life). Thomas, even at seven years old, you are the only person outside of my academic environment that can articulate what my thesis is actually about, despite sometimes calling it an ‘Eagle process’. Ella you have added vibrant colour to my life - and my charts, asking me to print them off and colouring them with crayons. Thank you my beautiful children for giving meaning and joy, above all else, to my life.

To my darling husband Brendan, this was never going to be possible without you. I can’t say how profoundly grateful I am for your untiring support and giving me the space to be my own person. For a man of few words, your actions have spoken volumes. And yes, I am cooking dinner for the next 10 years.

Through the unconditional love and support of the people in my life, I have been able to stumble my way forward to places I never thought possible.

Abstract

This thesis develops models and methods for the statistical properties of the limit order book for financial markets, a complex dynamical system of orders and cancellations, in continuous time and at multiple price levels. Initially, the heavy tailed features of limit order book volumes, aggregated to short, evenly spaced time intervals are investigated. These are found to require heavy tailed distributional models to adequately capture their statistical features.

A novel process to transform the physically operating order book into data suitable for analysis is presented. Limitations for point process modelling, such as events frequently occurring at the same time, are established. The marked Hawkes process is identified as a suitable model for event clustering. This overcomes many data constraints by aggregating events, allowing additional information potentially impacting the intensity, to be incorporated in marks attached to these events. Marks are identified by empirical research, which is guided by available literature.

A detailed description, methods of simulation, and parameter estimation of the univariate Hawkes process with multivariate marks is presented. This incorporates dependence features via copula models, with heavy tailed marginal distributions and requires substantial MATLAB implementation. Joint estimation via maximum likelihood, with the number and complexity of identified marks, necessitates the development of a method for screening marks that is computationally straightforward to implement. This new approach is based on the score test, which only requires the single fitting of the unboosted Hawkes process to the sequence of observed event times, together with the estimates of the moments of the functions of marks under assessment. The moments can be obtained parametrically, or non-parametrically. The test has an asymptotic chi-squared distribution under the null hypothesis that the marks do not impact the intensity. Extensive simulations confirm the power and utility of the test under realistic models and sample sizes. Application of the score test is made to futures data, and the identified serial dependence of the marks, leads to the new decoupled approximate method of likelihood estimation. This reduces model assumptions on statistical properties of the marks and leads to good performance of Hawkes process parameter estimation.

List of Figures

1.1	Example of a futures contract specification: COMEX Gold Futures.	9
1.2	Example of the limit order book.	11
2.1	5YTN: Heat maps of the volume for the first five levels of the LOB and the median volume on each level of the LOB on the bid and ask for 2010. The black lines represent the 9 U.S. public holidays that the CME observes.	53
2.2	BOBL: Heat maps of the volume for the first five levels of the LOB and the median volume on each level of the LOB on the bid and ask for 2010.	53
2.3	Boxplots of daily Hurst exponent for 2010. <i>Top Row - Left to Right</i> , level 1 to level 5 bid volumes; <i>Bottom Row - Left to Right</i> , level 1 to level 5 ask volumes; <i>Each Subplot</i> , assets left to right (1 to 6) are, 5YTN, BOBL, SP500, NIKKEI, GOLD and SILVER.	55
2.4	BOBL: extremogram heat map for 250 trading days, using <i>10 second</i> sub-sampled data, $a_m = 80$ th percentile and geometric distribution p-value= $1/200$	57
2.5	BOBL: extremogram for volume level 1 of the LOB, using 10 second sub-sampled data, 10,000 bootstrapped samples and for 100 lags. The upper chart shows the bid-side and the lower chart shows the ask-side. The solid horizontal line of height 0.20, represents the extremogram under independence. The bounds are found using the 0.025 and 0.975 quantiles from the empirical distribution of the bootstrapped replicates.	59
2.6	5YTN: Quantiles for exponential distribution model versus sample order statistics, for intra-daily volume data, every 25th trading day of 2010. <i>Top Row Bid</i> from left to right is level 1 to level 5 of LOB, and <i>Bottom Row Ask</i> from left to right is level 1 to level 5 of LOB.	62
2.7	GOLD: Quantiles for exponential distribution model versus sample order statistics, for intra-daily volume data, every 25th trading day of 2010. <i>Top Row Bid</i> from left to right is level 1 to level 5 of LOB, and <i>Bottom Row Ask</i> from left to right is level 1 to level 5 of LOB.	63
2.8	5YTN: Mean Excess plot versus the threshold u , for intra-daily volume data, every 25th trading day of 2010. <i>Top Row Bid</i> from left to right is level 1 to level 5 of LOB, and <i>Bottom Row Ask</i> from left to right is level 1 to level 5 of LOB.	64

2.9	GOLD: Mean Excess plot versus the threshold u , for intra-daily volume data, every 25th trading day of 2010. <i>Top Row Bid</i> from left to right is level 1 to level 5 of LOB, and <i>Bottom Row Ask</i> from left to right is level 1 to level 5 of LOB.	65
2.10	α -stable 10-day moving average of the daily parameter estimation, for the year 2010, using McCullochs method for <i>BOBL</i> , at a time resolution of 10 seconds. <i>Top Left Plot</i> : tail index parameter α daily estimates. <i>Top Right Plot</i> : asymmetry parameter β daily estimates. <i>Bottom Left Plot</i> : scale parameter δ daily estimates. <i>Bottom Right Plot</i> : location parameter μ daily estimates. The blue line is the bid level 1 and the red line is ask level 1.	78
2.11	Generalized extreme value intra-day 10-day moving average of the parameter estimation on each trading day of the year, for <i>BOBL</i> , bid and ask side, at a time resolution of 10 seconds. The blue line is the bid level 1 and the red line is ask level 1.	80
2.12	Generalized Pareto distribution intra-day 10-day moving average of the parameter estimation for each trading day of 2010, for <i>BOBL</i> , bid and ask side, at a time resolution of 10 seconds. The blue line is the bid level 1 and the red line is ask level 1.	81
2.13	Modified KS-Test statistics for the 5YTN, for every trading day, bid and ask side, using generalized Pareto distribution, MLE and Pickands' method.	83
3.1	Flow chart of the various stages of data processing.	89
3.2	Percentage of matched trades, as a function of total number of reported events, across 10 trading days (July 2015), for each futures asset considered.	93
3.3	An example of the matched market depth and trade data, with an example of multiple records on a single time-stamp, and the case of fill-down, whereby the trade data does not match the market depth data.	94
3.4	Count of events on the bid and ask side, LO, MO, C, for <i>SILVER</i> within 1 minute time bins, for the trading date 20-July-2015.	103
3.5	Correlation matrix of the count of events in each level 1:10, on the bid and ask side, LO, MO, C, for <i>SILVER</i> , within 1 minute time bins, for the trading date 20-July-2015.	104
3.6	Count of events on the bid and ask side, LO, MO, C, for <i>NIKKEI</i> within 1 minute time bins, for the trading date 20-July-2015.	105
3.7	Correlation matrix of the count of events in each level 1:10, on the bid and ask side, LO, MO, C, for <i>NIKKEI</i> , within 1 minute time bins, for the trading date 20-July-2015.	105
3.8	Sequence of reversion times r_z , when the count of events in a $z = 1$ second time bin have spiked above the mean level. We consider 10 trading days (20-31 July 2015), for <i>SILVER</i> , bid side only, 5 levels and for events, LO, MO, C.	108

3.9	Sequence of reversion times r_z , when the count of events in a $z = 1$ second time bin have spiked above the mean level. We consider 10 trading days (20-31 July 2015), for SILVER, bid side only, 5 levels and for events, LO, MO, C.	108
3.10	Moving average of a sequence of mean reversion time factors $\mathbb{E}[r_z]/z$, for increasing time bins, $z \in \{1, 2, \dots, 360\}$ seconds. We consider 10 trading days (20-31 July 2015), for SILVER and NIKKEI, bid side only, 5 levels and for events, LO, MO, C.	109
4.1	Boxplot of parameter estimates for a Hawkes process, with a constant boost function, a sample size $n = 1,000$, for 1,000 replicates. The solid red line represents the true value of each parameter.	131
4.2	Parameter estimate bias and the variability ratio for a Hawkes process, with a constant boost function, as the sample size of the simulation (1,000 replicates each) increases $n \in \{200, 400, \dots, 2,000\}$	133
4.3	Boxplots of the standard errors of each estimated parameter for a Hawkes process, with a constant boost function, using the inverse of the Hessian obtained from the optimization procedure and the bootstrapping method. The bootstrapped standard error estimates each use 200 replicates. Each simulation contains 1,000 replicates.	134
4.4	Boxplot of parameter estimates for a Hawkes process, with a sample size of $n = 1,000$, for 1,000 replicates. The mark is exponentially distributed $X \sim \text{Exp}(\lambda = 1)$, with a linear boost and with estimated theoretical moments. The solid red line represents the true value of each parameter.	135
4.5	Parameter estimate bias and variability ratio for a Hawkes process, with a linear boost function, with an exponentially distributed mark $X \sim \text{Exp}(\lambda = 1)$, as the sample size of the simulation (1,000 replicates each) increases $n \in \{200, 400, \dots, 2,000\}$	136
4.6	Boxplot of parameter estimates for a Hawkes process, with a sample size of $n = 1,000$, for 1,000 replicates. The mark is GPD $X \sim \text{GPD}(\zeta = 0.05, \delta = 1.00)$, with a linear boost and with estimated theoretical moments. The solid red line represents the true value of each parameter.	136
4.7	Parameter estimate bias and variability ratio for a Hawkes process, with a linear boost function, with a GPD mark $X \sim \text{GPD}(\zeta = 0.05, \delta = 1.00)$, as the sample size of the simulation (1,000 replicates each) increases $n \in \{200, 400, \dots, 2,000\}$	137
4.8	Boxplot of parameter estimates for a Hawkes process, with a sample size of $n = 1,000$, for 1,000 replicates. The mark is Poisson distributed $X \sim \text{Pois}(\mu = 1.50)$, with a linear boost and with estimated theoretical moments. The solid red line represents the true value of each parameter.	138

4.9	Parameter estimate bias and variability ratio for a Hawkes process, with a linear boost function, with a Poisson distributed mark $X \sim \text{Pois}(\mu = 1.50)$, as the sample size of the simulation (1,000 replicates each) increases $n \in \{200, 400, \dots, 2,000\}$	139
4.10	Boxplot of parameter estimates for a Hawkes process, with linearly boosted, bivariate exponentially distributed marks, where $X_i \in \mathbb{R}^2$ and $X_i \sim \text{Exp}(\lambda = 1.00)$, with a sample size of $n = 1,000$, for 1,000 replicates. The solid red line represents the true value of each parameter.	140
4.11	Boxplot of parameter estimates for a Hawkes process, with linearly boosted, bivariate GPD marks, where $X_i \in \mathbb{R}^2$ and $X_i \sim \text{GPD}(\zeta = 0.05, \delta = 1.00)$, with a sample size of $n = 1,000$, for 1,000 replicates. The solid red line represents the true value of each parameter.	141
4.12	Boxplot of parameter estimates for a Hawkes process, with linearly boosted, bivariate Poisson distributed marks, where $X_i \in \mathbb{R}^2$ and $X_i \sim \text{Pois}(\mu = 1.50)$, with a sample size of $n = 1,000$, for 1,000 replicates. The solid red line represents the true value of each parameter.	141
4.13	Kernel densities of the boost functions for a Hawkes process, with linearly boosted, bivariate marks $X_i \in \mathbb{R}^2$, distributed $X_i \sim \text{GPD}(\zeta = 0.05, \delta = 1.00)$. Each replicate combines the boost function multiplicatively, additively and jointly additively. The sample size of each replicate is $n = 1,000$	142
4.14	Boxplot of parameter estimates for a Hawkes process, with linearly boosted, bivariate GPD marks, where $X_i \in \mathbb{R}^2$ and $X_i \sim \text{GPD}(\zeta = 0.05, \delta = 1.00)$, combined with an <i>additive</i> boost function. The sample size is $n = 1,000$, for 1,000 replicates. The solid red line represents the true value of each parameter.	143
4.15	Boxplot of parameter estimates for a Hawkes process, with linearly boosted, bivariate GPD marks, where $X_i \in \mathbb{R}^2$ and $X_i \sim \text{GPD}(\zeta = 0.05, \delta = 1.00)$, combined with a <i>jointly additive</i> boost function. The sample size is $n = 1,000$, for 1,000 replicates. The solid red line represents the true value of each parameter.	144
4.16	Boxplot of parameter estimates for a Hawkes process, with linearly boosted, four dimensional GPD marks, where $X_i \sim \text{GPD}(\zeta = 0.05, \delta = 1.00)$, with a sample size of $n = 1,000$, for 1,000 replicates. The solid red line represents the true value of each parameter.	145
4.17	Boxplot of parameter estimates for a Hawkes process, with linearly boosted, bivariate exponentially distributed marks $\text{Exp}(\lambda = 1.00)$, with joint dependence modelled by a <i>Gaussian copula</i> , where $\rho_s = 0.50$. The sample size is $n = 1,000$, for 1,000 replicates. The solid red line represents the true value of each parameter.	147

4.18	Boxplot of parameter estimates for a Hawkes process, with linearly boosted, bivariate GPD marks $\text{GPD}(\zeta = 0.05, \delta = 1.00)$, with joint dependence modelled by a <i>Gaussian copula</i> , where $\rho_s = 0.50$. The sample size is $n = 1,000$, for 1,000 replicates. The solid red line represents the true value of each parameter.	147
4.19	Kernel densities of the boost functions from a Hawkes process, with linearly boosted marks $X_i \in \mathbb{R}^2$ distributed $X_i \sim \text{GPD}(\zeta = 0.05, \delta = 1.00)$. The first case considers independent marks, the second case considers jointly coupled marks via a <i>Gaussian copula</i> , where $\rho_s = 0.50$. The densities have a sample size of $n = 1,000$	148
4.20	Normalization (denominator) of the boost function, assuming GPD marks, for combinations of shape $\zeta \in \{0, 0.04, \dots, 0.48\}$ and scale $\delta \in \{0.01, 0.06, \dots, 0.80\}$, for a fixed $\psi = 0.5$. The first plot presents the case of independent marks and the second plot presents the normalization for jointly dependent marks, where $\rho_s = 0.50$	149
4.21	<i>Theoretical moments.</i> Boxplot of parameter estimates for a Hawkes process, with linearly boosted, bivariate GPD marks $\text{GPD}(\zeta = 0.49, \delta = 1.00)$, with joint dependence modelled by a <i>Gaussian copula</i> , where $\rho_s = 0.50$. The sample size is $n = 1,000$, for 1,000 replicates. The moments for the boost normalization are estimated theoretically. The solid red line represents the true value of each parameter.	150
4.22	<i>Imposed upper bound.</i> Boxplot of parameter estimates for a Hawkes process, with linearly boosted, bivariate GPD marks $\text{GPD}(\zeta = 0.40, \delta = 1.00)$, with joint dependence modelled by a <i>Gaussian copula</i> , where $\rho_s = 0.50$. The sample size is $n = 1,000$, for 1,000 replicates. The moments for the boost normalization are estimated theoretically. The optimization has an upper bound of 0.4999 for the shape parameters. The solid red line represents the true value of each parameter.	151
4.23	<i>Empirical moments.</i> Boxplot of parameter estimates for a Hawkes process with linearly boosted, bivariate GPD marks $\text{GPD}(\zeta = 0.40, \delta = 1.00)$, with joint dependence modelled by a <i>Gaussian copula</i> , where $\rho_s = 0.500$. The sample size is $n = 1,000$, for 1,000 replicates. The moments for the boost normalization are estimated empirically. The solid red line represents the true value of each parameter.	151
4.24	Boxplot of parameter estimates for a Hawkes process, with linearly boosted, bivariate GPD marks $\text{GPD}(\zeta = 0.05, \delta = 1.00)$, with joint dependence modelled by a <i>Gumbel copula</i> , where $\rho_s = 0.50$. The sample size is $n = 1,000$, for 1,000 replicates. The solid red line represents the true value of each parameter.	153

4.25	Boxplot of parameter estimates for a Hawkes process, with linearly boosted, bivariate GPD marks $\text{GPD}(\zeta = 0.05, \delta = 1.00)$, with joint dependence modelled by a <i>Clayton copula</i> , where $\rho_s = 0.50$. The sample size is $n = 1,000$, for 1,000 replicates. The solid red line represents the true value of each parameter.	153
5.1	Time series plots, ACF and histograms of a sample of 10,000 events, for Bid depth, with levels 1:5, for SILVER, on trading date 31-July-2015. . . .	175
5.2	Probability integral transform (PIT) histogram, for a sample of 10,000 events, for Bid depth, with levels 1:5, for SILVER, on trading date 31-July-2015.	176
5.3	QQ-plots, histograms and ACF plots. <i>Top Left</i> : GPD simulated data without serial dependence; <i>Right Column</i> : simulated data with serial dependence; and <i>Mid and Bottom Left</i> : a sample of 10,000 events for Bid depth ('real data') associated with levels 1:5, for SILVER, on trading date 31-July-2015. Theoretical quantiles correspond to a simulated sample from the model.	177
5.4	QQ-plots, histograms and ACF plots. <i>Top Left</i> : negative binomial distributed simulated data without serial dependence; <i>Right Column</i> : simulated data with serial dependence; and <i>Mid and Bottom Left</i> : a sample of 10,000 events for Bid depth ('real data') associated with levels 1:5, for SILVER, on trading date 31-July-2015. Theoretical quantiles correspond to a simulated sample from the model.	178
5.5	Time series plots, ACF and histograms of a sample of 10,000 events for Bid vol MOLOC, with levels 1:5, for SILVER, on trading date 31-July-2015. . . .	179
5.6	QQ-plots, histograms and ACF plots. <i>Top Left</i> : GPD simulated data without serial dependence; <i>Right Column</i> : simulated data with serial dependence; and <i>Mid and Bottom Left</i> : a sample of 10,000 events for Bid vol MOLOC ('real data') associated with levels 1:5, for SILVER, on trading date 31-July-2015. Theoretical quantiles correspond to a simulated sample from the model.	180
5.7	QQ-plots, histograms and ACF plots. <i>Top Left</i> : negative binomial distributed simulated data without serial dependence; <i>Right Column</i> : simulated data with serial dependence; and <i>Mid and Bottom Left</i> : a sample of 10,000 events for Bid vol MOLOC ('real data') associated with levels 1:5 for SILVER, on trading date 31-July-2015. Theoretical quantiles correspond to a simulated sample from the model.	181
5.8	Time series plots, ACF and histograms of a sample of 10,000 events for Bid vol C, with levels 1:5, for SILVER, on trading date 31-July-2015.	182

5.9	QQ-plots, histograms and ACF plots. <i>Top Left</i> : negative binomial distributed simulated data without serial dependence; <i>Right Column</i> : simulated data with serial dependence; and <i>Mid and Bottom Left</i> : a sample of 10,000 events for Bid vol C ('real data') associated with levels 1:5, for SILVER, on trading date 31-July-2015. Theoretical quantiles correspond to a simulated sample from the model.	183
5.10	Histogram, empirical CDF plots and the theoretical CDFs for the generalized Pareto distribution, negative binomial distribution and Poisson distributions, for the mark, Bid vol MOLOC associated with levels 1:5, for SILVER on 31st July 2015.	186
5.11	Boxplots of intra-day Hurst exponent across 10 trading days in July 2015, for bid side marks, for SILVER	188
5.12	Boxplots of intra-day Hurst exponent across 10 trading days in July 2015, for ask side marks, for SILVER	188
5.13	Pearson's linear correlation of the mark vector for events associated with levels 1:5, for SILVER, bid side, for the trading date 31-July-2015.	189
5.14	Spearman rank correlation of the mark vector for events associated with levels 1:5, for SILVER, bid side, for the trading date 31-July-2015.	190
6.1	Simulated upper tail probabilities of the score test under a range of data generating mechanisms. The light blue bands reflect 95% ranges around the nominal chi-squared upper 1%, 5%, 10% type I errors.	210
6.2	Comparison of sampling distribution of the score statistic against the asymptotic chi-squared distribution for several cases, which have varying degrees of freedom, $r \in \{1, 2, 4\}$: $X_i \sim \text{Exp}(\lambda = 1.00)$, with linear and quadratic boost; $X_i \sim \text{GPD}(\zeta = 0.25, \delta = 1.00)$, with linear boost; and $X_i \sim \text{GPD}(\zeta = 0.24, \delta = 1.00)$, with quadratic boost. We consider a small sample size of $T = 130$	211
6.3	Power curves for two sample sizes of $T = 130$ and $T = 1,000$ for a selection of cases: $X_i \sim \text{Pois}(\mu = 1.00)$, with linear boost; $X_i \sim \text{Exp}(\lambda = 1.00)$, with linear boost for i.i.d. marks and jointly dependent marks (Gaussian copula model, $\rho_s = 0.8$) and quadratic boost; $X_i \sim \text{GPD}(\zeta = 0.25, \delta = 1.00)$, with linear boost; and $X_i \sim \text{GPD}(\zeta = 0.24, \delta = 1.00)$, with quadratic boost.	212
6.4	Histograms of linear and quadratic boost functions for a single replicate of sample size $T = 1,000$ for: $X_i \sim \text{Exp}(\lambda = 1)$, with linear and quadratic boosts; $X_i \sim \text{GPD}(\zeta = 0.25, \delta = 1.00)$, with linear boost; and $X_i \sim \text{GPD}(\zeta = 0.24, \delta = 1.00)$, with quadratic boost, where $\psi_1 = \psi_2 = 0.5$	212
6.5	Power of the score test statistic for a linear boost and marks with an exponential distribution $X \sim \text{Exp}(\lambda = 1.00)$. This compares the use of theoretical moments with empirical moments in the score statistic, as well as true parameters and estimated parameters in the intensity function and the marks distribution.	214

6.6	Robustness of $\chi_{(r)}^2$ distribution for the sampling distribution, comparing when moments of the GPD exist, marginally exist and do not exist. The cases considered are, linear boost with $r=1$: $X_i \sim \text{GPD}(\zeta = 0.25, \delta = 1.00)$; $X_i \sim \text{GPD}(\zeta = 0.49, \delta = 1.00)$; and $X_i \sim \text{GPD}(\zeta = 0.50, \delta = 1.00)$, and quadratic boost with $r=2$: $X_i \sim \text{GPD}(\zeta = 0.10, \delta = 1.00)$; $X_i \sim \text{GPD}(\zeta = 0.24, \delta = 1.00)$; and $X_i \sim \text{GPD}(\zeta = 0.25, \delta = 1.00)$. The sample size is $T = 1,000$	216
6.7	Power of the score test for various combinations of GPD distributed marks and boost functions, illustrating the impact of shape parameter varying close to values, for which the moments required to define the score statistic, don't exist or only marginally exist. The sample size is $T = 1,000$	217
6.8	Impact on the sampling distribution and power curves of the score test statistic, by adjusting and not adjusting for serial dependence, with a sample size $T = 1,000$. The marks are conditionally GPD, $X_i \sim \text{GPD}(\zeta = 0.10, \delta_i)$, with a linear boost and empirically estimated moments.	218
6.9	Power of the score test for a linear boost, marks with a GPD and joint dependence modelled by a Gaussian copula, with Spearman's rank correlation, $\rho_s = \{0, 0.4, 0.8\}$. Comparing the use of theoretical moments and empirical moments, with calculation of cross terms and imposed block diagonalize of the covariance matrix in the score statistic. The sample size is $T = 1,000$	220
6.10	Simulated upper tail probabilities of the score test using theoretical and empirical moments, with calculation of cross terms and imposed block diagonalize of the covariance matrix. We consider the score test for a linear boost, marks with a GPD and joint dependence modelled by a Gaussian copula, with Spearman's rank correlation $\rho_s = \{0, 0.2, 0.6, 0.8\}$. The sample size is $T = 1,000$	221
6.11	Power of the score test, with a linear boost, increasing serial dependence, for $X \sim \text{GPD}(\zeta = 0.10, \delta_i)$, with a sample size $n = 1,000$. For later comparison, note that the immigration intensity is $\eta = 0.0010$ in this study.	223
6.12	Histograms of the linear boost function for a single simulant, for i.i.d. marks and those with serial dependence $a_1 \in \{0.4, 0.9\}$, for $X \sim \text{GPD}(\zeta = 0.10, \delta_i)$, with a sample size $n = 1,000$	224
6.13	Power of the score test, with a linear boost, increasing serial dependence, for $X \sim \text{GPD}(\zeta = 0.10, \delta_i)$, with a sample size $n = 1,000$. The immigration intensity is $\eta = 0.0020$	224
6.14	Power of the score test for a linear boost, with a boost parameter fixed at $\psi = 0.5$, increasing serial dependence, for $X \sim \text{GPD}(\zeta = 0.10, \delta_i)$, and an increasing immigration intensity. The sample size is $n = 1,000$	226
6.15	Long run intensity $\mathbb{E}[\lambda]$ for various combinations of immigration intensity η and branching coefficient ϑ	227

6.16	The drop in power of the score test for different combinations of immigration intensity and branching coefficient, for $\psi = 0.5$, across 1,000 simulations each. The measure of model stability IB for different combinations of immigration intensity and branching coefficient.	228
6.17	Power of the score test for a linearly boosted i.i.d. mark distributed $X_i \sim \text{GPD}(\zeta, \delta = 1.00)$ and increasing shape parameter ζ . The sample size is $n = 1,000$	229
6.18	Power of the score test for a linearly boosted, serially dependent mark distributed $X_i \sim \text{GPD}(\zeta, \delta_i)$, where $\delta_i = 0.9\delta_{i-1} + \epsilon_i$ and an increasing shape parameter ζ . The sample size is $n = 1,000$	230
6.19	Histograms of a linear boost function with a mark distributed $X_i \sim \text{GPD}(\zeta, \delta)$, for a single simulant, for each level of the shape parameter $\zeta \in \{0.05, 0.20, \dots, 0.80\}$, assuming both i.i.d. marks, where $\delta = 1.00$ and serially dependent marks, where $\delta_i = 0.9\delta_{i-1} + \epsilon_i$. The sample size is $n = 1,000$	230
6.20	Power of the score test, with a linear boost, increasing serial dependence, for $X \sim \text{NB}(r_i, p = 0.50)$, with a sample size $n = 300$	232
6.21	Histograms of the linear boost function, for a single simulant, i.i.d. marks and those with serial dependence $a_1 \in \{0.4, 0.9\}$, for $X \sim \text{NB}(r_i, p = 0.5)$, with a sample size $n = 300$	232
6.22	Power of the score test, with a linear boost, increasing marks dimension, and assuming i.i.d. GPD marks $X_i \sim \text{GPD}(\zeta = 0.10, \delta = 1.00)$. The sample size is $n = 1,000$	233
6.23	Histograms of the linear boost function for a single simulant, i.i.d. marks with dimension $d \in \{2, 6, 11, 16, 21\}$, for $X_i \sim \text{GPD}(\zeta = 0.10, \delta = 1.00)$. The sample size is $n = 1,000$	234
6.24	Power of the score test, with a linear boost, bivariate marks with Gaussian copula dependence, assuming i.i.d. GPD marks, $X_i \sim \text{GPD}(\zeta = 0.10, \delta = 1.00)$. The sample size is $n = 1,000$	236
6.25	Power of the score test, with a linear boost, bivariate marks with Gumbel copula dependence, assuming i.i.d. GPD marks $X_i \sim \text{GPD}(\zeta = 0.10, \delta = 1.00)$. The sample size is $n = 1,000$	236
6.26	Histograms of the linear boost function for a single simulant, jointly dependent i.i.d. marks $X_i \sim \text{GPD}(\zeta = 0.10, \delta = 1.00)$, with Spearman's rank, $\rho_s \in \{0, 0.4, 0.9\}$. The joint dependence is modelled by a Gaussian and a Gumbel copula, respectively. The sample size is $n = 1,000$	237
6.27	Power of the score test, with a linear boost, bivariate marks with Gaussian copula dependence, assuming serially dependent marks with a conditional GPD $X \sim \text{GPD}(\zeta = 0.10, \delta_i)$, where $\delta_i = 0.9\delta_{i-1} + \epsilon_i$. The sample size is $n = 1,000$	238

6.28	Histograms of the linear boost function for a single simulant, jointly dependent marks with Spearman's rank $\rho_s \in \{0, 0.4, 0.9\}$. The joint dependence is modelled by a Gaussian copula. The marks are serially dependent, with a conditional GPD $X \sim \text{GPD}(\zeta = 0.10, \delta_i)$, where $\delta_i = 0.9\delta_{i-1} + \epsilon_i$. The sample size is $n = 1,000$	239
7.1	Counts of events across even time intervals of 6 minutes for the entire day. The four plots show the counts of events across even time intervals of 1 second, for a duration of 6 minutes. The asset is SILVER, bid side and trading date 31-Jul-2015. Event types correspond to events $e \in \{LO, MO, C\}$ and on levels $l \in \{1, \dots, 5\}$	245
7.2	Counts of events across even time intervals of 6 minutes for the entire day. The four plots show the counts of events across even time intervals of 1 second, for a duration of 40 minutes. The asset is NIKKEI, bid side and trading date 31-Jul-2015. Event types correspond to events $e \in \{LO, MO, C\}$ and on levels $l \in \{1, \dots, 5\}$	245
7.3	<i>Individual marks.</i> The proportion (<i>right hand column scale</i>) of segments a mark is significant by the score test, across all time segments size $n \in \{1,000, 2,000, \dots, 10,000\}$, for 10 trading days 20-Jul-2015 to 31-Jul-2015, bid side, for SILVER and NIKKEI. Marks are evaluated using matched LOB data, with event types $e \in \{LO, MO, C\}$ and levels $l \in \{1, \dots, 5\}$	246
7.4	<i>Pairwise marks.</i> The proportion of time segments, which pairs of marks are significant via the score test, for time segments of 6 minutes, for SILVER and 40 minutes, for NIKKEI, across 10 trading days from 20-July-2015 to 31-July-2015. Combinations of $\{i, j\}$ marks are $i \in \{1, \dots, 20\}$ and $j \in \{2, \dots, 21\}$. Table 7.1 shows the names corresponding to the numerical values i and j . Marks are constructed using matched bid side LOB data, with event types $e \in \{LO, MO, C\}$ and levels $l \in \{1, \dots, 5\}$	248
7.5	<i>Multiple marks.</i> The proportion of significant marks via the score test, for time segment $n = 1,000$, for SILVER and $n = 5,000$, for NIKKEI, across 10 trading days 20-July-2015 to 31-July-2015, for combinations $j:i$ of marks. Marks are constructed using matched bid side LOB data, with event types $e \in \{LO, MO, C\}$ and levels $l \in \{1, \dots, 5\}$. Table 7.1 shows the names corresponding to the numbers in the chart.	249
7.6	Boxplot of parameter estimates for a Hawkes process, with a one dimensional, serially dependent mark, under the incorrect assumption of i.i.d., and with a sample size $n = 1,000$, for 1,000 replicates. The mark is conditionally GPD $X \sim \text{GPD}(\zeta = 0.05, \delta_i)$, with a linear boost and with estimated theoretical moments.	252
7.7	Histogram of the mark distribution, with a shape parameter $\zeta = 0.5529$. Histogram of a bootstrapped simulant of a serially dependent mark, with an MLE estimated shape parameter $\zeta = 0.8583$, under the incorrect assumption of i.i.d..	253

7.8	Barcode plot for each time segment, with an indicator $I = 1$ when the score test is significant, and an indicator $I = 1$ when the Hawkes process, with a linear boost and a univariate mark, is well calibrated ($IB < 0.4$). The asset is SILVER, on 31-July-2015, with event types $e \in \{LO, MO, C\}$, levels $l \in \{1, \dots, 5\}$ and bid side only.	254
7.9	<i>Model 1</i> : Bid depth. Boxplot of the parameter estimates for a Hawkes process, with a linear boost and a single mark. The models are estimated across 47 time segments, for SILVER bid side, on 31-July-2015, with event types $e \in \{LO, MO, C\}$ and levels $l \in \{1, \dots, 5\}$	255
7.10	<i>Model 2</i> : Bid vol MOLOC. Boxplot of the parameter estimates for a Hawkes process, with a linear boost and a single mark. The models are estimated across 47 time segments, for SILVER bid side, on 31-July-2015, with event types $e \in \{LO, MO, C\}$ and levels $l \in \{1, \dots, 5\}$	255
7.11	Boost function parameter estimates across all 47 time intervals for a Hawkes process, with a linear boost, a single mark, for SILVER bid side, on 31-July-2015, with event types $e \in \{LO, MO, C\}$ and levels $l \in \{1, \dots, 5\}$	256
7.12	Counts of events across even 1 minute time intervals. Insert plots are counts across even 1 second time intervals. Counts are for SILVER bid side, on trading day 31-July-2015, event types $e \in \{LO, MO, C\}$ and levels $l \in \{1, \dots, 5\}$. The blue bar represents the time segment to be modelled by a Hawkes process, with linear boost and a single mark.	256
7.13	A subset of the intensity function, with decay versus time, for SILVER. The subset represents 9 seconds of trading during 31-July-2015 and is modelled by a Hawkes process, with a linear boost function (blue) and a Hawkes process, with a constant boost function (red). Model 1 has a mark: Bid depth and <i>Model 2</i> has a mark: Bid vol MOLOC.	257
7.14	QQ-plots of the residual inter-arrival times and the residual counting process, for the Hawkes process, with a linear boost and univariate mark, Bid depth. The asset is SILVER on the trading date 31-July-2015, with event types $e \in \{LO, MO, C\}$, levels $l \in \{1, \dots, 5\}$ and bid side only. The two confidence bands for the Poisson null hypothesis are, 95% (black-dashed) and 99% (green-dashed).	258
7.15	QQ-plots of the residual inter-arrival times and the residual counting process, for the Hawkes process, with a linear boost and univariate mark, Bid vol MOLOC. The asset is SILVER on the trading date 31-July-2015, with event types $e \in \{LO, MO, C\}$, levels $l \in \{1, \dots, 5\}$ and bid side only. The two confidence bands for the Poisson null hypothesis are, 95% (black-dashed) and 99% (green-dashed).	258

7.16	Boxplot of parameter estimates for a Hawkes process, with bivariate serially and jointly dependent marks $X_i \in \mathbb{R}^2$, under the incorrect assumption of i.i.d., and with a sample size $n = 1,000$, for 1,000 replicates. The marks are conditionally GPD $X_i \sim \text{GPD}(\zeta = 0.05, \delta_i)$ and with a linear boost function. The joint dependence is modelled by a Gaussian copula with $\rho_s = 0.5$	261
7.17	Barcode plot for each time segment, with an indicator $I = 1$ when the score test is significant, and an indicator $I = 1$ when the Hawkes process, with a linear boost and bivariate marks is well calibrated ($IB < 0.4$). The asset is SILVER, on 31-July-2015, with event types $e \in \{LO, MO, C\}$, levels $l \in \{1, \dots, 5\}$ and bid side only.	262
7.18	Boxplot of the parameter estimates for a Hawkes process, with a linear boost and bivariate marks. The models are estimated across 47 time segments, for SILVER on 31-July-2015, with event types $e \in \{LO, MO, C\}$, levels $l \in \{1, \dots, 5\}$ and bid side only.	262
7.19	Boost function parameter estimates across all 47 time intervals for a Hawkes process, with a linear boost and bivariate marks, for SILVER bid side, on 31-July-2015, with event types $e \in \{LO, MO, C\}$ and levels $l \in \{1, \dots, 5\}$	263
7.20	Counts of events across even time intervals of one minute. Insert plots are counts of events across even time intervals of one second. The counts of events are for SILVER on a single trading date 31-July-2015. The event process is defined as event types $e \in \{LO, MO, C\}$, levels $l \in \{1, \dots, 5\}$ and bid side only. The blue bars represents the two time segment that will be modelled by a Hawkes process, with linear boost and bivariate marks.	263
7.21	A subset of the intensity function, with decay versus time, for SILVER. The two time segments represent 100 milliseconds and one second on 31-July-2015. The two models are: a Hawkes process with linearly boost bivariate marks, Bid depth and Bid vol MOLOC (blue); and a Hawkes process with a constant boost (red).	265
7.22	QQ-plots of the residual inter-arrival times and the residual counting process, for <i>Time segment one</i> . The Hawkes process has a linear boost and bivariate marks, Bid depth and Bid vol MOLOC. The asset is SILVER on the trading day 31-July-2015, with event types $e \in \{LO, MO, C\}$ and levels $l \in \{1, \dots, 5\}$. The two confidence bands for the Poisson null hypothesis are: 95% (black-dashed) and 99% (green-dashed).	265
7.23	QQ-plots of the residual inter-arrival times and the residual counting process, for <i>Time segment two</i> . The Hawkes process has a linear boost and bivariate marks, Bid depth and Bid vol MOLOC. The asset is SILVER on the trading day 31-July-2015, with event types $e \in \{LO, MO, C\}$ and levels $l \in \{1, \dots, 5\}$. The two confidence bands for the Poisson null hypothesis are: 95% (black-dashed) and 99% (green-dashed).	266

7.24	Boxplot of parameter estimates from 1,000 simulations, with bivariate serially dependent GPD marks that are jointly dependent. The parameters are estimated via: the decoupled approximate likelihood method; and the <i>approximate likelihood method by decoupling the marks parameters</i> , with specifications of a jointly dependent GPD marks, evaluated with empirical moments.	268
7.25	Boxplot of parameter estimates (misspecified) from 1,000 simulations, with bivariate serially dependent GPD marks that are jointly dependent. The replicates are modelled using a Hawkes process, with jointly dependent exponentially distributed marks, evaluated with empirical moments.	269
7.26	Boxplot of the parameter estimates evaluated by the decoupled approximate likelihood method, and the <i>approximate likelihood method by decoupling the marks parameters</i> , that specified the bivariate marks as GPD, with joint dependency modelled via a Gaussian copula. The models are estimated across 47 time segments, for SILVER on 31-July-2015, with event types $e \in \{LO, MO, C\}$, levels $l \in \{1, \dots, 5\}$ and bid side only.	270
7.27	Boxplot of the parameter estimates evaluated by the <i>approximate likelihood method by decoupling the marks parameters</i> , for a Hawkes process with bivariate marks specified as exponentially distributed, with joint dependency modelled via a Gaussian copula. The models are estimated across 47 time segments, for SILVER on 31-July-2015, with event types $e \in \{LO, MO, C\}$, levels $l \in \{1, \dots, 5\}$ and bid side only.	271
A.1	Recursive method parameter estimate minus the non-recursive method parameter estimates in the full joint likelihood of a Hawkes SEPP with linear boost and an Exponentially distributed mark, $X \sim \text{Exp}(\lambda = 1)$. The simulation has 1000 replicates of sample size $n = 1000$ each.	301

List of Tables

1.1	Futures and exchanges studied in this thesis. Average daily volume and open interest as of Q1 2018, for the CME Group listed instruments. Current daily volume and open interest as of August 2018 front contract, for EUREX and SGX listed instruments.	8
2.1	Asset description used in the analysis and modeling. Market hours refer to the liquid market hours in local trading time of the exchange.	50
2.2	Descriptive statistics for volumes on level 1 of the LOB across all trading days, for 2010, using sub-sample data of 10 seconds.	51
2.3	Mean Hurst Exponent across all trading days, using varying sub-sampled data of volumes on <i>level 1</i> of the LOB.	56
2.4	Distributions and methods fit to volume data for six assets, 5YTN, BOBL, SP500, NIKKEI, GOLD and SILVER. The methods include, MLE, method of moments, McCullochs quantile based estimation (McCulloch, 1986), Pickands' estimator (Pickands III, 1975) and empirical percentile method.	66
2.5	Superior method fitted to volume data for six assets, 5YTN, BOBL, SP500, NIKKEI, GOLD and SILVER. The methods include, MLE, method of moments, McCullochs quantile based estimation (McCulloch, 1986), Pickands' estimator (Pickands III, 1975) and empirical percentile method	83
3.1	<i>Market depth data.</i> MD=Market Depth, BP=Bid Price, AP=Ask Price, BV=Bid Volume, AV= Ask Volume, N=number.	90
3.2	<i>Trades and Quotes.</i>	90
3.3	<i>Matched data.</i> MD=Market Depth, LO=Limit Order, B=Bid, A=Ask. . . .	91
3.4	Asset description used in the analysis and modeling. Market hours refer to the liquid market hours in local trading time of the exchange.	99
3.5	<i>Matched data.</i> Mean values of counts of events divide by total events. The tick range within bid and ask, not including the spread, across 10 trading days (July 2015) for each futures asset considered.	99
3.6	<i>Aggregated data.</i> Mean values of counts of events divide by total events. The spreads between the best bid and ask across a 10 trading days (July 2015) for each futures asset considered.	99
3.7	The percentage of simultaneous events occurring on the bid and ask side, and the mean number of events on the bid and ask, for various combinations of event types and levels, across 10 trading days (July 2015), for 5YTN. . .	101

3.8	The percentage of simultaneous events occurring on the bid and ask side, and the mean number of events on the bid and ask, for various combinations of event types and levels, across 10 trading days (July 2015), for BOBL. . .	101
3.9	The percentage of simultaneous events occurring on the bid and ask side, and the mean number of events on the bid and ask, for various combinations of event types and levels, across 10 trading days (July 2015), for GOLD. . .	101
3.10	The percentage of simultaneous events occurring on the bid and ask side, and the mean number of events on the bid and ask, for various combinations of event types and levels, across 10 trading days (July 2015), for SILVER. . .	101
3.11	The percentage of simultaneous events occurring on the bid and ask side, and the mean number of events on the bid and ask, for various combinations of event types and levels, across 10 trading days (July 2015), for SP500. . .	102
3.12	The percentage of simultaneous events occurring on the bid and ask side, and the mean number of events on the bid and ask, for various combinations of event types and levels, across 10 trading days (July 2015), for NIKKEI. . .	102
3.13	The mean summary statistics of counts of events within evenly spaced time intervals of one minute, occurring on the bid and ask side, event types, LO, MO, C, across each level, for 10 trading days (July 2015), for SILVER. . .	102
3.14	The mean summary statistics of counts of events within evenly spaced time intervals of one minute, occurring on the bid and ask side, event types, LO, MO, C, across each level, for 10 trading days (July 2015), for NIKKEI. . .	104
4.1	<i>Simulation.</i> Combinations of continuous and discrete mark distributions, boost functions and dependence structures available for simulation. . . .	126
4.2	<i>Maximum likelihood estimation.</i> Combinations of continuous and discrete mark distributions, boost functions and dependence structures available for likelihood estimation.	127
4.3	Parameter estimate bias (%) and variability ratio for three bivariate Hawkes process models, with a linear boost function and marks $X_i \in \mathbb{R}^2$ distributed, $X_i \sim \text{Exp}(\lambda = 1.00)$, $X_i \sim \text{GPD}(\zeta = 0.05, \delta = 1.00)$ and $X_i \sim \text{Pois}(\mu = 1.50)$, respectively. The simulations each have 1,000 replicates, with a sample size of $n = 1,000$ each.	140
4.4	Parameter estimate bias (%) and variability ratio for three bivariate Hawkes process models, with linear boost functions, with marks $X_i \in \mathbb{R}^2$ distributed $X_i \sim \text{GPD}(\zeta = 0.05, \delta = 1.00)$. The three different models combine the boost functions, multiplicatively, additively and jointly additively. The simulations each have 1,000 replicates, with a sample size of $n = 1,000$ each.	143

4.5	Parameter estimate bias (%) and variability ratio for four bivariate Hawkes process models, with a linear boost function, with marks $X_i \in \mathbb{R}^2$ distributed: $X_i \sim \text{Exp}(\lambda = 1.00)$ (labelled ‘Exp’); marginal $X_i \sim \text{Exp}(\lambda = 1.00)$ and jointly coupled via a <i>Gaussian copula</i> where $\rho_s = 0.50$ (labelled ‘Exp-Cop’); $X_i \sim \text{GPD}(\zeta = 0.05, \delta = 1.00)$ (labelled ‘GPD’); and marginal $X_i \sim \text{GPD}(\zeta = 0.05, \delta = 1.00)$ and jointly coupled via a <i>Gaussian copula</i> where $\rho_s = 0.50$ (labelled ‘GPD-Cop’). The simulations each have 1,000 replicates, with a sample size of $n = 1,000$ each.	146
4.6	Parameter estimate bias (%) and variability ratio for three bivariate Hawkes process models, with linear boost functions, marks $X_i \in \mathbb{R}^2$ distributed $X_i \sim \text{GPD}(\zeta = 0.49, \delta = 1.00)$. The marks are jointly coupled and modelled via a <i>Gaussian copula</i> with $\rho_s = 0.50$. The three different models are, theoretical moments (<i>Theo.</i>), theoretical moments with an upper bound (<i>UB.</i>) of 0.4999 for the optimization procedure and empirical moments (<i>Emp.</i>). The simulations each have 1,000 replicates, with a sample size of $n = 1,000$ each.	150
4.7	<i>Theoretical moments.</i> Parameter estimate bias (%) and variability ratio for three models, each being a bivariate Hawkes process model, with a linear boost function, with marks $X_i \in \mathbb{R}^2$ and marginal distributions $X_i \sim \text{GPD}(\zeta = 0.05, \delta = 1.00)$. The marks for each three models are jointly coupled via a Gaussian, Gumbel and Clayton model, respectively, where $\rho_s = 0.50$. The simulations each have 1,000 replicates, with a sample size of $n = 1,000$ each.	152
4.8	<i>Empirical moments.</i> Parameter estimate bias (%) and variability ratio for three models, each being a bivariate Hawkes process model, with a linear boost function, with marks $X_i \in \mathbb{R}^2$ and marginal distributions $X_i \sim \text{GPD}(\zeta = 0.05, \delta = 1.00)$. The marks for each three models are jointly coupled via a Gaussian, Gumbel and Clayton model, respectively, where $\rho_s = 0.50$. The simulations each have 1,000 replicates with a sample size of $n = 1,000$ each.	154
5.1	Shortened naming convention, with equation reference for the endogenous marks constructed from matched LOB data.	162
5.2	Mean values of the summary statistics of the mark vector, related to both bid and ask side events, for levels 1:5, across a 10 trading days (July 2015), for SILVER.	167
5.3	A summary of the transformations applied to the mark vector. This is related to both bid and ask side (where appropriate), for levels 1:5, across a 10 trading days (July 2015), for SILVER	174
5.4	Summary statistics for Bid depth, for levels 1:5, across a 10 trading days (July 2015), for SILVER and two simulates from a conditional GPD and a conditional negative binomial distribution, with serial dependence.	176

5.5	Summary statistics for Bid vol MOLOC, for levels 1:5, across a 10 trading days (July 2015), for SILVER, and two simulates from a conditional GPD and a conditional negative binomial distribution, respectively.	180
5.6	Summary statistics for Bid vol C, for levels 1:5, across a 10 trading days (July 2015), for SILVER, and a simulate from a conditional negative binomial distribution with serial dependence.	182
5.7	Categorization of the mark vector	184
5.8	Proposed continuous and discrete marginal distribution for each mark, for levels 1:5, across 10 trading days (July 2015), for SILVER. Shape range refers to the GPD shape parameter estimates, across the 10 trading days.	185
5.9	Mean values of the summary statistics of the mark vector, related to both bid and ask side events, for levels 1:5, across a 10 trading days (July 2015), for SILVER.	186
5.10	Ljung Box test, with percentage of times that the null hypothesis is rejected in favour of the data exhibiting serial dependence, for each mark and time segments of size 10,000, across 10 trading days.	187
6.1	Upper tail probabilities for the nominal chi-squared upper 5% type I errors.	221
7.1	Shortened naming convention with equation reference for the endogenous marks constructed from matched LOB data. The numbers following the name prior to the equation reference are used in the charts that follow.	243
7.2	The proportion of significant marks, as defined by the score test, across time segment size, $n \in \{1,000, 2,000\}$, for SILVER and $n \in \{5,000, 6,000, 7,000\}$, for NIKKEI, across 10 trading days from 20-Jul-2015 to 31-Jul-2015. Marks are evaluated using matched LOB data, with event types $e \in \{LO, MO, C\}$, levels $l \in \{1, \dots, 5\}$ and for the bid side only.	247
7.3	Impact of a mark with serial dependence on estimating the parameters for the Hawkes process, whilst incorrectly assuming i.i.d. marks. The sample size is $n = 1,000$, for 1,000 replicates. The mark is conditionally GPD, $X \sim \text{GPD}(\zeta = 0.05, \delta_i)$, with a linear boost and with estimated theoretical moments.	252
7.4	<i>Model 1</i> and <i>Model 2</i> parameter estimates for a Hawkes process with a linear boost function, estimated for a single time segment (6 minutes), for SILVER on 31-July-2015, with event types $e \in \{LO, MO, C\}$, levels $l \in \{1, \dots, 5\}$ and for the bid side only.	257
7.5	Impact of bivariate marks with serial and joint dependence on estimating the parameters for the Hawkes process, whilst incorrectly assuming i.i.d. marks. The sample size is $n = 1,000$, for 1,000 replicates. The marks are conditionally GPD, $X_i \sim \text{GPD}(\zeta = 0.05, \delta_i)$ and with a linear boost function. The joint dependence is modelled by a Gaussian copula with $\rho_s = 0.5$	260

7.6	<i>Time segment 1</i> and <i>Time segment 2</i> parameter estimates for a Hawkes process with a linear boost function and bivariate marks that are jointly dependent, for SILVER on 31-July-2015, with event types $e \in \{LO, MO, C\}$, levels $l \in \{1, \dots, 5\}$ and for the bid side only.	264
7.7	Mean parameter estimates for a Hawkes process with a linear boost function, estimated across all time segments, for SILVER on 31-July-2015, with event types $e \in \{LO, MO, C\}$, levels $l \in \{1, \dots, 5\}$ and for the bid side only. We consider two methods of estimation and two parametric specifications for the Hawkes process, GPD and exponential for the <i>approximate likelihood method by decoupling the marks parameters</i>	271
A.1	Archimedean copula generator functions, inverse generator functions, and generator function d -th derivatives, $\psi^{(d)}$	295
A.2	Recursive method and the non-recursive method for the full joint likelihood parameter estimation of a Hawkes SEPP with linear boost and an Exponentially distributed mark, $X \sim \text{Exp}(\lambda = 1)$. The simulation has 1000 replicates of sample size $n = 1000$ each.	300

Chapter 1

Introduction and literature review

1.1 Context

The relationship between financial development and economic growth was first documented by Goldsmith (1969) 50 years ago, debunking the long held view that financial systems are an inconsequential side show responding passively to economic growth and industrialization (Levine, 1997). In recent times, this area of research has received a great deal of attention, with two contradictory strands of literature about the role financial markets play in economies. The finance-growth nexus literature refers to a positive long-run association between financial development and economic growth. Conflicting literature highlights excessive financial deepening, or too rapid a growth of credit, giving rise to growth-inhibiting financial crises (Rousseau and Wachtel, 2011). Recent work by Breitenlechner et al. (2015) links these two contradicting strands, presenting evidence that financial development is linked positively to GDP growth in normal, non-crisis times, and larger financial sectors lead to significantly negative economic outcomes in times of banking crisis.

The role of financial markets is to: facilitate the transfer of resources; provide borrowers with funds to enable them to carry out investment plans, whilst lenders earn interest or dividends; provide liquidity in the market to facilitate trading; disseminate information across many segments of the market; diversification and risk management; and allocate funds in the economy, based on demand and supply, through a mechanism called the *price discovery process*. The financial market is a generic term used to reflect the marketplace where various financial instruments are traded typically via an electronic system. Each market is defined by its market structure. Its trading rules and trading systems determine who can trade and what instruments are traded (Harris, 2003). Capital markets facilitate the long term flow of funds, providing financing via shares (stock market) and bonds (bond markets). Capital markets can be divided into primary and secondary markets. The issuance of a new security (initial public offerings) is facilitated in the primary markets, and secondary markets facilitate the trading of existing securities. Derivative markets offer products which are used to mitigate risk, but have also been used, paradoxically, to exploit risk. Others, include money markets, commodity markets, foreign exchange markets, spot markets and interbank lending markets. There are many introductory texts on financial markets with a highly referenced resource being Harris (2003).

Research on financial markets and the price discovery process is conducted on different time-scales, with macroeconomic applications, through to understanding market microstructure. Within this section, we consider the evolution of financial markets from a macroeconomic perspective then we lead into the role that electronic market order book structures have played within this evolution. Integrated within these themes is market regulation, which attempts to influence the financial markets into cultivating a more efficient price discovery process. Presenting the backdrop of financial evolution in recent times is intended to orient the reader in the complexity and interconnectedness of these various strands of research at different time scales and the role that these shorter intra-day time scale dynamics of the limit order book, which we study in this thesis, play on price discovery.

In recent years a large body of literature on financialization has emerged. There is no single definition of financialization. Broadly speaking it describes the growth of financial capitalism and encompasses the increasing role of financial motives, financial markets, financial actors, and financial institutions in the operation of domestic and international economies (Epstein, 2005). Studies on financialization aim to understand the implications of financial expansion beginning in the 1980s, after the deregulation of financial markets shifted the basic structure of the economy to favour the financial sector, and the failure to regulate new financial instruments or strategies (Tomaskovic-Devey and Lin, 2013).

If we consider more specifically the financialization of commodities, historically commodity futures have been used as a tool for commodity producers to hedge risk. The inflow of commodity index traders (CITs), also known as index speculators, has propagated the inflow of capital. This has increased the co-movements of commodity prices with other financial assets and equity indexes, and less with the physical process that creates and refines the raw commodities underlying the commodity futures. Financialization of commodities has led to the decoupling of the futures price from the original price structure. During volatile periods, for example the Global Financial Crisis of 2007-2008, several commodities across energy, metal, and agricultural sectors experienced a synchronized boom-bust cycle. This raised concerns about whether financialization of commodity markets had distorted prices (Silvennoinen and Thorp, 2013; Tang and Xiong, 2012; Cheng and Xiong, 2014). For further detail on commodities and commodity derivatives, see Geman (2005) and Geman (2008).

This significant growth in financial markets and the speed at which transactions can take place through electronic systems has led to a multitude of events at great economic cost: the stock market crash known as Black Monday in 1987; the 1997 Asian Financial Crisis; the near collapse of the hedge fund Long Term Capital Management in 1999; the dot.com collapse in 2000; the Global Financial Crisis of 2007-2008; the 2010 Flash Crash; and the 2015-2016 stock market sell-off, that wiped off an estimated 10 trillion dollars from global markets. During times of crisis when markets are highly volatile, markets become highly correlated, with extensive literature addressing the causes and effects of these crisis cross-market linkages (Dungey and Martin, 2007; Longstaff, 2010). Research on crisis contagion shocks typically considers longer time-scales. Crisis contagion shock

could be a shock in a particular asset market affecting other markets or countries, a shock in an asset market that impacts on specific asset class within a group of countries, or a country shock which impacts all asset classes of that country (Dungey and Martin, 2007). Cross market linkages are also studied on shorter and often intra-day time-scales, both in and out of crises periods, for example intra-day volatility spillovers between spot and futures indexes (Kang et al., 2013). These shorter, intra-day time scale applications will benefit from the statistical methods developed within this thesis.

The interrelationship between longer time-scale macroeconomic factors and the shorter time-scale market liquidity are front and centre in the functioning of financial markets and ensuing economic outcomes. Market liquidity (often referred to as *liquidity* in this thesis) is the ease at which an asset can be sold without a loss of value. This is distinct but related to funding liquidity, which is the ease at which a trader can obtain funding (Brunnermeier and Pedersen, 2009). Market liquidity is represented by bid-ask spreads, limit order book depths and volumes, and is influenced by macroeconomic factors such as interest rates, default spreads and market volatility (Chordia et al., 2001). The statistical analysis and modelling within this thesis is applied to elements of market liquidity, such as limit order book depth, volumes, spreads, and more, at high frequency intra-day time scales. As posited by Harris (2003), everyone likes liquidity. Traders like liquidity because it allows them to implement their trading strategies cheaply. Exchanges like liquidity because it attracts traders to their markets. Regulators like liquidity because liquid markets are often less volatile than illiquid ones. It is the most important characteristic of a well-functioning market.

In the years before the Global Financial Crisis, the deregulated global markets had high levels of market liquidity, due to macroeconomic factors such as low interest rates, high rates of savings in Asia, economic growth, and low volatility. This led to low borrowing costs and apparent low risk, with financial institutions becoming highly leverage (*positive liquidity spiral*) Pedersen (2013). Despite the sub-prime mortgage sector being small relative to the broader financial system and the exposure widely dispersed through securitization, the deterioration in the credit quality of these mortgages led to an amplified shock leading to the credit crisis.

The majority of the securities and derivatives involved in the Global Financial Crisis were traded in over-the-counter (OTC) markets. Assets traded OTC, trade via a dealer network rather than a centralized exchange. There is less transparency, lower liquidity, less rigorous regulation and dealers can withdraw from market-making at any time. Without liquid and orderly markets there was no price discovery process, and in turn no easy and definitive way to value the securities¹. The credit crunch saw the decline in the value of sub-prime mortgages and the realization by investors that many would default. With an increase in volatility, market-making became risky and expensive for dealers and led to their withdrawal. This withdrawal meant that investors could not trade out of losing positions nor meet margin calls, marking the beginnings of the liquidity crisis, an acute shortage or *drying up* of liquidity.

¹<http://www.imf.org/external/pubs/ft/fandd/basics/markets.htm>

Spillover catalysts can be: a macroeconomic downturn in a major economy; failure of large financial institutions, such as Lehman Brothers; or in this case, the shock to the sub-prime credit market, which quickly led to the indiscriminate spillover to other markets (cross-market linkages), even the most liquid (Pedersen, 2013). As prices moved away from fundamentals, the markets became illiquid, volatility increased, and the increase in co-movements in the markets induced further liquidity dry-ups and *spirals* (Brunnermeier and Pedersen, 2009). The considerable challenges brought forth by the Global Financial Crisis underscored financialization as a systematic risk to the global economy, necessitating additional market regulation. Whilst the research in this thesis studies electronic exchange listed assets (not assets that trade OTC), the increase in market regulation across the board has had substantial impacts on how electronic exchanges operate, therefore affecting the price formation mechanism on all trading venues.

Market regulation is aimed at: ensuring financial market stability; market integrity; competitive and efficient market trading; and consumer protection by subjecting financial institutions to requirements, restrictions and guidelines. The aftermath of the Global Financial Crisis led to regulatory intervention of unprecedented scale. In November 2007, investment firms and regulated markets in Europe have been closely regulated by the EU Markets in Financial Instrument Directive (MiFID) (Moloney, 2010). Similarly, the U.S. implemented the Regulation National Market System (Reg NMS)². In the US, The Dodd-Frank Wall Street Reform and Consumer Protection Act³ and its various subcomponents, such as the Volcker Rule, were signed in 2010 and included, consolidated regulatory agencies to monitor financial markets for signs of systemic risk, and transparency of derivatives. The Basel III accord, international regulatory framework for banks⁴, was signed in 2010 and was an extension of the Basel II framework, introducing higher capital requirements and new liquidity management rules. However, the international standardized bodies, such as the Basel Committee BCBS, the International Organization of Securities Commission IOSCO and the International Accounting Standards Board IASB, depend on voluntary compliance with the rules developed (Mayntz, 2012). Myriad of other regulatory changes have taken place and the pattern of increasing regulation continues. As recent as the 3 January 2018, the MiFID regime was replaced by MiFID II and whilst MiFID had a reputation of being strict, MiFID II tighten the reins even further after the Global Financial Crisis revealed gaps in the previous legislation (Busch, 2017). There has been much concern about the possibility of permanently inhibiting the market-making capacity of large banks, with dire consequences in terms of under-provisions of market liquidity. However, studies by Trebbi and Xiao (2017) suggest that post crisis, U.S. regulatory intervention did not appear to have produced structural deteriorations in market liquidity. For a more detailed discussion on exchange regulation, refer to Peters and Vishnia (2017, Chapter 12).

The complexity of the interconnected financial markets and the increasingly elaborate trading instruments and strategies that execute on electronic systems, have added exten-

²<https://www.sec.gov/rules/final/34-51808.pdf>

³<https://www.cftc.gov/LawRegulation/DoddFrankAct/index.htm>

⁴<https://www.bis.org/bcbs/basel3.htm>

sive complications in understanding liquidity. However, undisputed is the role liquidity plays for not only market participants transacting at well-formed prices and regulators designing policies for financial markets, but the consequential impacts to the broader economy. A focus of this thesis is to develop advanced statistical methods for the study of liquidity dynamics at the shorter intra-day time scales and the role they play in the price discovery process, which is the cornerstone of well functioning financial markets.

1.2 Derivative asset classes and market specificities

For a reference book on derivative markets, see Gottesman (2016). Broadly speaking, the main participants of the derivatives market segregated by their trading motives are, investors, hedgers, speculators and arbitragers. Investors take a position in the underlying asset with a long-term investment view, which in turn impacts the pricing of the derivative. Speculators are short-term investors who trade the derivative and benefit from either an increase or decrease in the underlying asset. Changes in regulation have led to a decrease in risky speculative trading and associated derivative products. Hedgers require the asset at some future time using the derivative to lock in a price and gain if the underlying instrument declines. Alternatively, they buy the asset and lock in its value for future sale. Arbitragers do not have a view on the direction of the price of the asset, but trade by buying and selling the same or equivalent products at different prices. They achieve essentially a risk-less profit due to mis-pricing of the instrument or discrepancies in efficient price formation on different exchanges or markets. For example, if there is a deviation in the relationship of the spot and futures price, they will buy the cheap asset, sell the expensive, closing the deviation. Arbitragers play an important role in keeping spot and futures prices aligned.

1.2.1 Exchange venues

A detailed overview of exchanges venues can be found in Peters and Vishnia (2017, Chapter 12). The derivatives market can be broken up into three categories, exchange traded derivatives, over-the-counter (OTC) derivatives, and cleared OTC derivatives. *Exchange traded derivatives*, such as futures and options are highly standardized contracts. They trade through primary exchanges, which are regulated markets (also known as primary markets), and other trading venues, such as dark pools. The clearing and settlement of transactions is with a central counter-party (clearing-house). Primary exchanges are non-discretionary execution systems, which are strictly a platform for investors to carry out trades to fulfil their investment decisions (Peters and Vishnia, 2017). Examples of primary exchanges are, the CME, New York Stock Exchange (NYSE), Nasdaq and Tokyo Stock Exchange (TSE). Primary exchanges typically offer ‘Lit’ (visible) limit order books. Exchange traded futures will be the assets studied in this thesis.

Prior to the recent regulations discussed above, there was only a second type of exchange venue available, the OTC market where *OTC derivatives* trade. OTC trading is done directly between two parties, otherwise known as bilateral trading with no central

exchange. Dealers act as market-makers, quoting prices, which may not be the same as prices quoted by other dealers. Historically this information was provided via telephone, email, instant message, and electronic bulletin boards. In more recent times, this process has become automated, creating concerns about the removal of fiduciary duties with respect to best execution obligations (Peters and Vishnia, 2017). This prompted further regulatory requirements in MiFID I.

Cleared OTC derivatives are a bilateral trading of standardized transactions that are privately negotiated, but booked with a central counter-party. They share similar features to exchange traded derivatives, such as the standardization of contracts, however the standardization is not as extensive and the transactions of cleared OTCs are booked through a clearing-house. They remain an OTC product as they are not traded on a central trading venue.

Following on from MiFID I there has been an emergence of other trading venues, such as multilateral trading facilities (MTFs), which are self-regulated trading venues and are an alternative to primary exchanges. In the same way that primary exchanges are non-discretionary, MTFs have no discretion as to how interests of participants may interact, rather execution taking place under system rules and protocols. Similarly to primary exchanges, they offer pre- and post-trade transparency (publishing bids and asks) (Peters and Vishnia, 2017)

Systematic Internalisers, replacing Broker Crossing Networks (BCNs), are privately run crossing networks which match flow of clients in order to execute. Their lack of transparency is a differentiator to a primary exchange. Another example of privately run trading venues are ‘dark pools’ (different to ‘Lit’). Dark pools are private exchanges that systematically match orders and are not accessible by the public. Dark pools originated to facilitate block trades (significantly large sized trade) as the investor is likely to achieve a better price. This is due to the lack of transparency by not being required to disclose their intention to the exchange first and because the dark pool contains other large investors. Transactions in dark pools do not contribute to public price discovery. Under new regulation, the dark pool universe will need to support the same order types and pricing and queue priority rules as a Lit exchange, but will not publish publicly their order book (Peters and Vishnia, 2017).

From a modelling perspective, the availability and structure of the data varies greatly from exchange traded derivatives to OTC derivatives. For example, primary exchanges and MTFs are required to disseminate information about trades, quantities and quotes on a real time basis, so that they are accessible to investors. Primary exchanges and MTFs offer Lit limit order books providing best bid and ask prices, and volumes (first level of the book), which are published for any subscriber (Peters and Vishnia, 2017). Deeper levels of the limit order book are available by subscription. Financial data collected from exchanges by data vendors is largely standardized across assets and exchanges, with these observable data sets being amendable to modelling. We will discuss this further in Section 1.3 and Chapter 3. Whereas, the pricing information for instruments trading on the OTC markets and similarly, systematic internalisers and dark pools are not open to all

participants. The mechanism by which OTC markets, systematic internalisers and dark pools operate is vastly different to exchanges and the data structures do not contribute to the electronic daily volume traded on an asset, thus unobservable. From a modelling perspective, primary exchanges and to a similar extent MTFs, offer an observable data set that contributes to the price discovery mechanism of a traded asset.

1.2.2 Futures

The use of futures dates back to commodities futures trading in 16th century England. The first recognized exchange, traded rice futures in Japan in 1710. The first formalized futures exchange in the U.S. was the Chicago Board of Trade (CBOT), which opened in 1848, with the first futures contract recorded in 1851 (Kettell, 2002). The price discovery of a future was historically determined by open out-cry in a trading pit, i.e. people calling prices to each other in a room. A few exchanges still have physical traders on the floor, for example the London Metal Exchange which was established in 1877, however the days of pit trading have essentially ended. The equivalent now is the electronic limit order book, discussed further in Section 1.3. The shift away from conventional open out-cry markets was to reduce transaction costs, increase speed of execution and reduce trading errors.

Futures are a derivative of an underlying asset and are in fact constructed from simple elements known to financial markets for centuries (Kettell, 2002). Futures contracts are a binding agreement to take, or make delivery of a pre-specified quantity and quality of an asset, on a predetermined date and location. The value of the futures contract is derived from the eventual use of the underlying asset. There are two main types of futures contracts with underlying assets being:

1. Physical commodity: energy, metals, and agriculture (such as gold, silver, corn, soy-bean meal and wheat)
2. Financial security: interest rates, equity indexes, and FX currency (such as Treasury bills, Treasury bonds, S&P 500 Index, Nikkei 225 Index, Euro and Australian dollar)

The futures market instruments are remarkably varied. This is due to the attributes of a commodity or security that are required to establish a futures market. Commodities need to be traded freely, generally without direct government controls over the prices paid for the commodity or security. They need to exhibit a considerable degree of price fluctuation. The commodities quality must be graded according to a universally accepted standard. There needs to be plentiful supply of the commodity or security, or cash settlement is possible (Viney, 2002). The broad range of choices facilitates the diverse selection of futures which we consider in this thesis (Table 1.1), enabling a greater cross-section for analysis.

Our selection represents futures that are highly liquid ensuring the rate of information arrival is fast and reported prices are efficient. This in turn, ensures in an efficient price discovery process. For example, the 5 Year T-Note is the third highest traded interest rate future on CBOT, E-mini S&P 500 is the highest traded equity index future on CME,

Gold and Silver are the first and third highest traded metal future traded on COMEX, respectively.⁵

Table 1.1: Futures and exchanges studied in this thesis. Average daily volume and open interest as of Q1 2018, for the CME Group listed instruments. Current daily volume and open interest as of August 2018 front contract, for EUREX and SGX listed instruments.

Underlying asset name	Exchange	Daily volume	Open interest
Interest rate futures			
5 Year T-Note	Chicago Board of Trade (CBOT)	1,207,634	3,427,082
Euro-BOBL	Eurex Exchange (EUREX)	372,931	1,646,736
Equity index futures			
SGX Nikkei 225	Singapore Exchange (SGX)	60,698	138,818
E-mini S&P 500	Chicago Mercantile Exchange (CME)	1,959,829	2,983,386
Precious metals futures			
Gold	Commodity Exchange, Inc (COMEX)	381,582	499,565
Silver	Commodity Exchange, Inc (COMEX)	101,387	229,131

Source: CME Group⁶; EUREX⁷; SGX⁸

The majority of traders do not take physical delivery, rather they close out their positions prior to expiration. The volume of futures trading and the underlying quantity of a commodity far exceeds the total production. This differential continues to widen with financialization. The liquidity in the futures market is crucial, ensuring there is sufficient market participants generating market turnover and to enable contracts to be bought or sold without price distortion. However, the impact of liquidity is two-fold, large volumes not only influencing futures prices, but ultimately impact the price of the underlying asset.

The standardization of futures contracts is a core requirement for exchange traded futures. The standardization of futures contracts creates homogeneity, whereby the counter parties can always unwind a previous commitment prior to expiration by taking an off-setting contract. Figure 1.1 presents an example of a Gold Futures contract, which is traded on COMEX. The specification presented in Figure 1.1, is typical for a commodity futures contract. Some attributes include the contract size, which is the actual size per contract (100 troy ounces in this example). The price quotation is specific to the commodity. The minimum fluctuation is the tick size or price increments. Unlike stocks, futures have expiry dates (termination of trading), which is typically related to the cycle of the commodity. Refer to Section 2.1 for a discussion on expiry dates. The deliverable grade is the predetermined quality of the commodity that is required for delivery.

Futures exchanges require that both counter parties post collateral, or margin, to protect itself against the possibility of default. These margin accounts are held by the exchanges clearing-house and are marked-to-market daily or intra-daily depending on the exchange, which means there is an adjustment for price movements on a daily basis. Margin requirements are typically between 5% – 15% of the contracts value. To continue to hold the future, the money held by the clearing-house must stay above the maintenance margin requirement, which is the minimum margin required to maintain their account. If the margin requirement of the value of the outstanding futures contract drops below the

⁵<https://www.cmegroup.com/education/files/cme-group-leading-products-2018-q1.pdf>

maintenance margin level, then a margin call requires the investor to deposit an amount that brings the account back to this margin requirement.

Contract Specifications

Gold Futures		
Product Symbol	GC	
Venue and Hours (All Times are New York Time/ET)	CME Globex CME ClearPort	Sunday – Friday 6:00 p.m. – 5:00 p.m. (5:00 p.m. – 4:00 p.m. Chicago Time/CT) with a 45-minute break each day beginning at 5:00 p.m. (4:00 p.m. CT)
Contract Size	100 troy ounces	
Price Quotation	U.S. Dollars and Cents per troy ounce	
Minimum Fluctuation	\$0.10 per troy ounce	
Termination of Trading	Trading terminates on the third last business day of the delivery month.	
Listed Contracts	Trading is conducted for delivery during the current calendar month; the next two calendar months; any February, April, August, and October falling within a 23-month period; and any June and December falling within a 72-month period beginning with the current month.	
Settlement Type	Physical	
Delivery Period	Delivery may take place on any business day beginning on the first business day of the delivery month or any subsequent business day of the delivery month, but not later than the last business day of the current delivery month.	
Trading at Settlement (TAS)	Trading at Settlement is allowed in the active contract month. The active contract months will be February, April, June, August and December. On any given date, TAS transactions will be allowed only in a single contract month. TAS transactions may be executed at the current day's settlement price or at any valid price increment ten ticks higher or lower than the settlement price.	
Grade and Quality Specifications	Gold delivered under this contract shall assay to a minimum of 995 fineness.	
Rulebook Chapter	113	

Source: CME Group

Figure 1.1: Example of a futures contract specification: COMEX Gold Futures.

1.3 The limit order book

It is of little use to draw up grandiose schemes of concepts and to list far-reaching hypotheses unless the concepts can be quantified and the hypotheses tested – Goldsmith (1959)

In 1971 the National Association of Securities Dealers Automation Quotations (NASDAQ) was launched. However, this system was essentially an electronic bulletin board for posting bids and asks. The CME Group's electronic trading platform was fully launched in 1992, followed by CBOT launching the 'E Open Outcry', which was an electronic platform that operated along side the open out-cry pits. Advances in technology, access to personal computers and internet connectivity, resulted in huge growth of electronic trading for both institutional and retail traders throughout the 1990's. By the 2000's, trading algorithms and faster hardware led to the increase in high frequency trading. By 2015, 99% of all futures contracts traded on the CME, were electronic⁹. Along with these developments were changes to regulation, which promoted the transition of exchanges to electronic systems.

In recent times there has been significant investment by firms in increasing their speed-to-market for high frequency trading. The emergence of the limit order book as the primary trading mechanism globally has provided an unusually rich, detailed and high-quality historic data source. The demand for limit order book data and at smaller and smaller time granularities has grown substantially. Thompson Reuters Tick History (TRTH)¹⁰ started

⁹<https://www.fxcm.com/insights/>

¹⁰<https://financial.thomsonreuters.com/en/products/data-analytics/market-data.html>

collection of tick history in 1996. Bloomberg¹¹ commenced collection of tick data in the mid 2000's. With advancements in technology, and algorithmic and high frequency trading demands, many other data providers have emerged. Companies such as TickData,¹² LOBSTER,¹³ Olsen,¹⁴ Tickstory,¹⁵ Tickdatamarket,¹⁶ and more offer varying tick history, quality of data, and minimum time granularity, across various global markets. TRTH has been a major source of tick data for both industry and academic research over the years. They have the longest history of normalized data globally, which makes it suitable for academic research across a variety of assets and exchanges.

The limit order book data has facilitated suitable testing ground for theories about well-established statistical regularities common to a wide range of markets, which ultimately provides insight into the price formation mechanism (Gould et al., 2013). However, this growth has led to significant data complexity in the observable scale and detail of financial data. The data structures go well beyond the classical time series econometric data, to ultra-high frequency event based data. Compounding this, is the changing market regulations that impact the trading mechanism and rules. These factors make the need to understand and model aspects of the price formation process increasingly important for both market participants and regulatory policy makers alike. Research of the dynamics of the limit order book is largely uncharted, providing great opportunity to explore the stochastic structures of such processes for enhanced modelling and simulation.

The limit order book can be considered a queue of limit orders that have not traded, with a specified buy or sell price. There are two sides to a limit order book, the *bid* (limit order to buy an asset) and the *ask* (limit order to sell an asset). In broad terms we define three classical order (event) types of the limit order book as, limit order, market order and cancellation. The limit orders are displayed at different *price levels* in the limit order book. If a trader wants to immediately buy an asset at the best ask price, they submit a *market order*, which means buyer and seller match via submission of a buy market order matching a sell limit order. The result is a *trade*. Throughout this thesis we refer to a market order and trade interchangeably. In this context, the limit order book can be described as a multi-level stochastic process of orders at different price points, on the bid and ask side. Incorporated within this complex process are elements, such as *cancellation* of orders, price and volume amendments and the dynamics of market orders on limit orders. Outside of the scope of this thesis are conditional limit orders, such as stop limit order, stop market order, market-if-touched order (MIT), tick-sensitive order, and market-not-held order. Refer to Harris (2003, Chapter 4) for a detailed overview of order types.

How orders arrive in the limit order book is determined by the classification for that particular market. The markets in this study are classified as order-driven markets which are the most common form of market. Order-driven markets operate according to specified

¹¹<https://www.bloomberg.com/professional/product/market-data/>

¹²<https://www.tickdata.com/>

¹³<https://lobsterdata.com/>

¹⁴<http://www.olsendata.com/>

¹⁵<https://tickstory.com/>

¹⁶<http://www.tickdatamarket.com/>

trading rules and orders typically have a price and time priority, where orders of the same price are prioritized by the time that they arrive. It is worth noting that another priority mechanism used in futures markets is pro-rata. If a buy market order arrived in the limit order book, it would be matched with the limit orders at that price proportional to the volume of each order. For this thesis we consider the almost universal price-time priority.

A *limit order* is defined as any order with a specified maximum buy price or a specified minimum sell price. When a limit order is submitted it enters a queue where priority is given to the highest priced limit order for buy side and lowest priced limit order for sell side. When two orders are identically priced, the order submitted first (older order) is given priority. A number of parameters must be specified when submitting a limit order, the limit price, buy (or sell) and order volume.

Figure 1.2 presents a simplistic example of the limit order book. The bid side (blue) represents the limit orders for market participants who wish to buy the assets. The ask side (red) represents the limit orders for those wishing to sell. The limit orders are arranged in price-time priority. The difference between the bid and ask is the spread.

In the event that a market order matches against multiple limit orders on more than one price level $j \in \{1, \dots, m\}$, the market order is priced as the volume weighted average price (VWAP). In the case of benchmarking, it may also represent j market orders across some time interval (for example, the day), or some combination of the aforementioned. VWAP is defined as

$$P^{(VWAP)} = \frac{1}{\sum_{j=1}^m V_j} \times \sum_{j=1}^m (V_j \times P_j), \quad (1.1)$$

where V_j is the volume and P_j is the price.

Consider Figure 1.2, if a sell market order of 925 shares, matches against volume on price level 1 (38.8000) and price level 2 (38.7900), the resulting traded VWAP price is 38.0903.



Figure 1.2: Example of the limit order book.

The limit order book data required for this study would ideally incorporate the entire history of the limit order book with full depth at each price level. It would include all transaction data, inclusive of both limit order executions and market order executions. However, the *full* order book is generally inaccessible. The models proposed in this thesis will also be valid for order book data that is consolidated at the price level, consolidated by time (ie multiple orders occurring within the same millisecond) and data which contains depth that is a subset of the depth of the full order book. For example, most vendors will supply Level-II data that contains up to price level 5 or price level 10 depending on the market and asset, see Chapter 3 for further detail.

A significant body of research within this thesis is dedicated to defining the limit order book and outlining a rigorous approach to construct the observed data sets. This will be used for developing appropriate models to study of order flows in limit order books across multiple exchanges. In addition, this approach employed in the construction of the data sets, affords flexibility in constructing sub-sets of data to study various financial phenomena.

The process by which the physically operating limit order book is converted to data provided by vendors and subsequently used for analysis, is complex and a description of this whole process is non-trivial. There is no one method for consolidating the data in the limit order book, and this is confounded by the different exchange mechanism across assets. The large volume of data of the limit order book makes it difficult to assess both the validity of its construction and then to conduct rigorous statistical analysis on the empirical features. Whilst there are vast amounts of empirical limit order book research, inconsistencies in the underlying data and the absence of a ‘correct’ method for constructing the observable data set, make the findings of the research difficult to incorporate into a statistical model of limit order book dynamics. Adding to this, the empirical research and current limit order book models have been built on statistical regularities that have been observed in ‘old’ data, often on a single asset and over small time frames (Gould et al., 2013). There has been limited guidance on the underlying assumptions behind the construction of the data sets used, with exception of research by Hautsch (2012) and Toke (2016). The changing market place and technology advancements are strong motivators to clearly define data sets studied, ensuring studies have the flexibility to incorporate multiple assets, across a broad but recent time horizon.

Hautsch (2012) outlined methods of handling high frequency data, including descriptions on matching trades and quotes, data cleaning, dealing with split transactions and identifying buyer and seller initiated trades. Hautsch (2012) discussed various properties of financial duration data and properties of trading characteristics, such as returns, trading volumes, spreads and market depth. Since Hautsch (2012) was published, a significant increase in vendors supplying varying quality of high frequency data necessitates further consideration of data handling methods, which should go beyond recommendations made by Hautsch (2012).

Toke (2016) introduced the limit order book and the reported, tick-by-tick data as provided by Thompson Reuters (TRTH) database. Toke (2016) outlined the algorithm

used to match the trade data with the quote data and some of the challenges this process presents. The description provided by Toke (2016) is one of the most comprehensive discussions we have found on the matching process and presenting the often undervalued, but extremely important choices one must make in the data preparation stages of researching the limit order book.

Toke (2016) presented and compared three matching algorithms, the perfect match, grouping trades to match a single limit order book change on the same time stamp, and grouping trades that do not have exactly the same time stamp to match the limit order book change. Toke (2016) presented a discussion on the third case when reported transactions are executed against hidden liquidity and may result in a single update to the market, despite executing against more shares than displayed. Also discussed was the potential for ‘off the book’ trades that may not affect the quote file, but are reflected as trades in the trade file. Toke (2016) studied the accuracy of the Lee-Ready procedure (Lee and Ready, 1991) in determining whether a trade is buy or sell side initiated and the impact of choices in the matching processes. Toke (2016) demonstrated a toy model of the shape of the ‘trade signature’, which presents optimal lags when using the Lee-Ready procedure. Finally, Toke (2016) presented a simple Hawkes process and investigated the impact of using trade data versus matched trade data, as defined by the matching algorithm, with the latter proving a superior data set to model.

Filimonov and Sornette (2015) is one of the only papers to present the biases created in microstructure research of high frequency financial data, due to the challenges of data integrity. Filimonov and Sornette (2015) provided a detailed description of the FIX/FAST protocol, which bundles up multiple messages from the exchange, then sent to a data provider such as TRTH. The millisecond time stamps of events in the limit order book and trade data occur at the vendor (TRTH) and ‘stamping’ occurs when the vendor receives the data, not at the exchange. Filimonov and Sornette (2015) reported the biases caused in the preparation of this data, such as: the impact of overnight trading; intra-day seasonality and de-trending the data; and the vulnerability of analysis due to regime shifts. As stated by Filimonov and Sornette (2015), calibrating a model is akin to an excursion within a minefield that requires expert and careful testing before any conclusive step can be taken.

The complexity of high frequency financial data sets should not be underestimated. The development of robust methods to construct, and ultimately aggregate data, make a material difference to the choice of analysis methods and models used to replicate the stylized features of the limit order book. This leads to a better understanding of price formation and the implications of liquidity on market stability. This challenge is magnified when attempting to establish the statistical regularities across markets and assets. The constant enhancements in technology, the quality of the data and frequency which it is reported, provides a shifting landscape. It is imperative that methods of data handling are revisited frequently and with each new study.

1.4 Models for the limit order book

The *price formation mechanism* relates the dynamics of supply and demand for an asset and is revealed through financial data, via subsequent variations in its market price (Sirignano and Cont, 2018). The *exogenous* components of the price formation mechanism vary between assets. For example, FOREX prices are influenced by international trade and investment flows, whereas equity, bond and derivatives markets are influenced by economic and political conditions. The way prices are formed in markets for physical commodities and futures contracts is the result of complex interactions between factors, such as: product characteristics (quality, store-ability or substitutability, etc.); supply and demand factors (capital intensity, industry concentration, production facilities, etc.); access to finance, public subsidies and interventions; and the weather (Valiante, 2013). The classical scenario suggests that the exogenous flow of information drives the prices towards a fundamental value (Bacry et al., 2015). The *endogenous* component is the internal feedback mechanism, readily studied in literature on stylized features of financial data (Cont, 2001), with supporting evidence of universal statistical regularities as the main source of price variations (Cont and Tankov, 2004).

The relationship of regulatory rules and market behaviour is not one of linear causality in either direction, but is instead a complex dance in which market behaviour and regulatory action shadow, anticipate, and react to each others' moves in turn. –Black (2010)

The study of the impact of algorithms and high frequency trading on market stability and price formation is important for both the market practitioner finding ways to optimally execute trading strategies and from a market regulatory perspective. The growth of electronic exchanges has extended the field of study to understanding the impact of market abuse, such as insider trading and market manipulation on market microstructure. This has contributed to increased financial regulation such as Reg NMS in the U.S., and MiFID II (Busch, 2017) in Europe, as discussed in Section 1.1. A regulators understanding of price formation is critical for designing policies and procedures aimed at improving the price formation mechanism further, or to curtail degradation in price formation due to financial crises or market malpractice. For example, the “Order Protection Rule” of Reg NMS requires trading centres to establish polices and procedures designed to prevent the execution of trades at prices inferior to protected quotations displayed by other trading centres.¹⁷ The relationship of policy change and ensuing market microstructure change is often unpredictable and driven by dynamics that are not fully understood. This motivates detailed modelling of the limit order book process for simulation purposes, which can be beneficial to test regulation and test price formation under different conditions and scenarios.

Empirical studies of the limit order book identify stylized facts, which can be thought of as non-trivial statistical regularities (Gould et al., 2013). The identification of the stylized features provides evidence that there is non-equilibrium behaviour in the limit order book.

¹⁷<https://www.sec.gov/rules/final/34-51808.pdf>

This promotes the development of theories about the existence of these stylized features, which appear to be common for a range of assets and markets. Statistical models are often employed to build upon the understanding of the dynamics of the relationships of the stylized features and the key determinants of the price formation mechanism, such as order flow, price impact, volatility, market stability and resiliency, accounting for systematic risk contagion, deriving optimal execution strategies and capturing the dynamics of the full limit order book, to name a few. However, the development of statistical models often employs trade data only, and/or best bid and ask. As discussed by Sirignano and Cont (2018), the TeraBytes of high frequency data on transactions, order flow and limit order book dynamics in listed markets, which is a map representing the relationship between market price and variables such as price history and order flow, can help to better understand the nature of the price formation mechanism.

Prior to the application of point processes in finance the majority of limit order book dynamic models consolidated limit order book events into regularly spaced time intervals. Information is lost due to the aggregation of data and applications, such as volatility measurement and optimal execution strategies which require that the models relate in some way to real time (Bowsher (2007)).

1.4.1 Point processes for the limit order book

Point processes and queuing analogies have long been associated with modelling order flows for financial markets (Cont, 2011). A limit order book can be considered a high-dimensional queuing system, whereby queues of buy and sell limit orders are executed against market orders or cancelled before execution. The underlying assumption is that the order flows follow independent Poisson processes (Huang et al., 2015). This was first considered by Smith et al. (2003) who introduced the zero-intelligence limit order book models. The concept of zero intelligence is to accurately model the market mechanism, whilst assuming participants have no strategy and behave at random. The aim is to separate out the effects of the market mechanism and trader strategy, to determine the driving forces within the market (Ladley, 2012). With the available financial data, it is possible to reconstruct the limit order book and make comparisons between the dynamics of the zero-intelligence models and historical markets. These models were further enriched with more realistic conditions by Cont et al. (2010), Huang et al. (2015) and Toke (2015), with theoretical studies by Abergel and Jedidi (2013). In reality, the order flows cluster in time and this results in autocorrelation in durations and cross-correlation of arrival rates. These features are not captured in models based on Poisson processes (Cont, 2011).

There has been a proliferation of the use of point processes in financial literature (refer to Daley and Vere-Jones (2007) for a comprehensive textbook on point processes). Point processes describe the arrival of events of some kind. Financial data sets contain many possibilities for how one might define an ‘event’ for a specific financial application. The first type of point process that was proposed for high frequency financial data was the autoregressive conditional duration models by Engle and Russell (1998), where the durations are a measure of time between events. A review and mathematical descriptions

of the ACD model and extensions can be found in Bauwens and Hautsch (2009). We present the general ACD model, as described in Engle and Russell (1998). Let $\tau_i = t_i - t_{i-1}$ be the interval between two arrival times called durations. Let ψ_i be the expectation of the i th duration given by

$$\mathbb{E}[\tau_i | \tau_{i-1}, \dots, \tau_i] = \psi_i(\tau_{i-1}, \dots, \tau_i; \theta) = \psi_i. \quad (1.2)$$

Let the ACD class of models consist of parametrization of (1.2) and the assumption that

$$\tau_i = \psi_i \epsilon_i, \quad (1.3)$$

where ϵ_i defines an i.i.d. random variable for which it is assumed that $\mathbb{E}[\epsilon_i] = 1$. The general model, which is called an ACD(m, q) is given by

$$\psi_i = \alpha_0 + \sum_{j=1}^m \alpha_j \tau_{i-j} + \sum_{j=1}^q \beta_j \psi_{i-j}, \quad (1.4)$$

where the m and q refer to the orders of the lags, $\alpha_j > 0$, $\beta_j > 0$, and $j > 0$.

Multivariate extensions of models of durations were considered by Engle and Lunde (2003), who modelled a bivariate process of trades and mid-point changes. Bowsher (2007) presented the limitations of this and other multivariate extensions within a durations setting. The model proposed by Engle and Lunde (2003) was not a full bivariate point process as it does not imply an intensity in continuous time for mid-price change events. Further to this, the mid-price change during an inter-trade duration cannot influence the trade intensity during that duration, leading to an implicit loss of information (Bowsher, 2007). Another approach proposed by Russell (1999) is the Autoregressive Conditional Intensity model, which is an alternative to the Hawkes process (Hawkes, 1971). Bowsher (2007) highlights that whilst this method provides an alternative to the Hawkes process, the properties of the model have not been proven in the bivariate case.

Following on from autoregressive conditional duration and intensity models, the more flexible Hawkes point process came to the fore. Mathematical descriptions and variations of this model are given in Sections 1.5 and 1.6. The distinguishing feature of this approach, is that rather than directly specifying a model of durations, the model is specified via the stochastic intensity. This is advantageous as it allows the specification of a multivariate point process model, or a model for a univariate point process conditional on some other continuous time process (Bowsher, 2007). Since the development of the Hawkes process, the vast applications have extended from seismology, finance, sociology, genome analysis, neuroscience and criminology. Zhu (2013) provides a summary of applications of the Hawkes process across various fields of research. The Hawkes process was first applied to financial data by Bowsher (2007), Chavez-Demoulin et al. (2005), Hewlett (2006) and Large (2007) who collectively recognize the great flexibility and versatility for modelling high frequency data. Since the introduction of the Hawkes process for the modelling of financial data in 2007, there has been a proliferation of research in this field with many extensions and variations of the model proposed.

The simplest form of a Hawkes process is the unmarked process. Trades (market orders) and mid-prices are examples of data sets that are relatively easy to acquire and require limited processing prior to modelling, relative to the full set of limit order book data. It therefore forms a logical starting point for research, and represents the vast majority of applications of the Hawkes process. As the data sets become more complex, for example considering depth in the limit order book, the research becomes sparse. However, to accurately characterise the statistical properties of market data, limit order book dynamics cannot be ignored. This presents additional challenges as the limit order book events are far more frequent than mid-prices moves and the underlying data sets are hugely complex.

Substantial information can be garnered from the limit order book to potentially enhance the fitting and prediction of the Hawkes process. This is done via the incorporation of a vector of additional *marks* associated with that event. An example of a marked point process is a set of event times associated with trade arrivals and the mark being the volume of each trade. The information contained in the marks is likely to be of significant importance to most financial applications of the Hawkes process. However, the literature on marked Hawkes processes is very limited, and no literature exists on marked Hawkes processes for limit order book depth data. Whilst the benefits of fitting a Hawkes process with marks are easily recognisable, the construction of marks and fitting challenges are non-trivial and which we aim to address within this thesis.

1.5 Unmarked Hawkes processes

The objective of this research is to develop sophisticated statistical techniques to effectively capture the dynamics of the limit order book. Throughout this section, we present an introduction that motivates the study of high frequency financial data and we explain how Hawkes processes have been applied to address the modelling or understanding of these research motives. It is not the aim of this research to address the ideas proposed below, but it is meaningful to highlight the relevance of the current work and the research opportunities that exist with a sophisticated ‘tool-set’ that facilitates a more rigorous investigation of the limit order book. In this section we establish important findings that can be applied to our research and research gaps that need to be addressed.

Definitions and properties of the univariate Hawkes process can be found in Chapter 4. A brief overview of the multivariate, univariate and renewal Hawkes processes are outlined below to aid in the discussions that follow. In practice, the point process is only observed over the interval of time $[0, T]$ and for computation a truncated version must be used, as presented in Liniger (2009), which calculates the likelihood over available event times. Consider a multivariate Hawkes self-exciting point process N , observed over the interval $t \in [0, T]$. The observed points of this process are $\{t_i, i = 1, \dots, n\}$ $N_t \in \mathbb{N}$, with event times $0 = t_0 < t_1 < t_2 < \dots < t_n \leq T$. The multivariate ‘unmarked’ Hawkes intensity process has the j th component $j \in \{1, \dots, m\}$ intensity defined as

$$\lambda_j(t) := \eta_j + \sum_{k=1}^m \vartheta_{j,k} \int_{[0,t)} w_j(t-s; \alpha) N_k(ds), \quad t \in \mathbb{R}, \quad (1.5)$$

where immigration rate is denoted by $\eta_j > 0$, the branching coefficient is denoted by $\vartheta_j > 0$ and the decay function is denoted by w_j .

The univariate ‘unmarked’ Hawkes intensity process is defined as

$$\lambda(t) := \eta + \vartheta \int_{[0,t)} w(t-s; \alpha) N(ds), \quad t \in \mathbb{R}, \quad (1.6)$$

where immigration rate is denoted by $\eta > 0$, the branching coefficient is denoted by $\vartheta > 0$ and the decay function is denoted by w . The univariate Hawkes intensity process has only one intensity process, whereas the multivariate version in (1.5) has m intensity processes.

For both models the immigration intensity, in the modelling context of this thesis, will be interpreted as the frequency with which new events arrive in the limit order book. It can be described as the background intensity representing the arrival of exogenous events. Extensions to the Hawkes process incorporate an inhomogeneous immigration intensity $\eta(t)$, which is a deterministic function of time (Bowsher, 2007; Toke and Pomponio, 2012; Gao et al., 2017).

Another extension is to specify the immigration intensity as a general renewal process in (1.7) (Stindl and Chen, 2018), where the immigration intensity will renew at each immigration point. This allows the number of immigrant events of the same type in non-overlapping time intervals to have serial correlation and to be over- or under-dispersed relative to the Poisson distribution (Stindl and Chen, 2018). Denote the unobservable immigrant status indicator as M_i , where $M_i = 0$ if the event is an immigrant, otherwise it is an offspring and $M_i = 1$. Let $\mathbf{I}[t] = \max\{i : t_i < t, M_i = 0, z_i = j\}$, where $z_i \in \{1, \dots, M\}$ indicate the i th event type. The intensity for the j th component for the multivariate renewal Hawkes process, as defined in Stindl and Chen (2018) is

$$\lambda_j(t) := \eta_j \left(t - t_{\mathbf{I}_j[t]} \right) + \sum_{k=1}^{i-1} \vartheta_{j,z_k} h_{j,z_k}(t - t_k), \quad t \in \mathbb{R}, \quad (1.7)$$

where $\eta_j \left(t - t_{\mathbf{I}_j[t]} \right)$ is the immigrant intensity and $h_{j,z_k}(t - t_k)$ is the offspring density, and combined with the branching ratio is the excitation function. The immigration arrivals form a renewal process, whereas the immigrants in the Hawkes process in (1.5) arrive according to a Poisson process. There is reasonable evidence of the benefits of a non-constant immigration intensity function for financial applications, for example addressing the non-stationary behaviour of limit order book event arrivals over longer intra-day time scales. However, a marked Hawkes process with a general renewal process is yet to be developed and is beyond the scope of this thesis.

For the Hawkes process the intensity is increased proportionally by the branching coefficient ϑ given an event has occurred. When new orders (immigrants) arrive in the limit order book we expect a clustering effect, which results in an increase in secondary events called descendants. How fast this effect decays in time is governed by the decay function. The decay function in a Hawkes process is often referred to interchangeably as a kernel or decay function. The kernel function does not have to be monotonically decreasing, however for financial application it is natural for the event to decay over

time, i.e. the process of order arrivals progressively forgets the state of the arrivals in the increasingly distant past, at some increasing rate. Within this research we consider only decreasing families of functions and generally refer to the kernel as a decay function in a parametric setting. However, in the non-parametric setting we revert to the more appropriate terminology of kernel.

The decay function specification can be parametric, such as: an exponential decay, which is used in the majority of financial applications; power-law decay; or sums of exponential kernels. The sum of exponentials is an estimate of a power-law decay function whilst maintaining the Markov property of the exponential decay function. This is important for computational time when fitting the Hawkes process. Recent literature presented conflicting results about the appropriateness of a particular decay function, with some presenting support for the exponential decay function (Filimonov and Sornette, 2012, 2015). Others have advocated the power-law decay function (Hardiman et al., 2013; Hardiman and Bouchaud, 2014). More recently, non-parametric methods of calibrating the kernel function from empirical data have been proposed, see a review by Bacry et al. (2015). If the decay function is unknown then the non-parametric method may be preferred rather than the standard MLE method for estimating parameters of exponential and power-law decay functions (Bacry et al., 2012). Ultimately the choice of the decay function should reflect prior knowledge of the problem at hand. It is not a case of one-size-fits-all, as assumed in the majority of financial literature which construct a Hawkes process with an exponential decay function, and with no supporting evidence for their choice. For realistic practical development and assessment of such models these simplifying choices of an exponential decay function, though often convenient, may not always be appropriate. We will provide detailed evidence of this in the limit order book modelling context in future chapters.

The functional form of the intensity process can be extended to a non-linear intensity function (Zheng et al., 2014; Lu and Abergel, 2017; Rambaldi et al., 2018). As discussed in Hawkes (2018), this allows for greater flexibility in the shape of the kernels and reproduction of inhibition effects, where the intensity is decreasing instead of an exciting effect. The intensity can be transformed by a non-linear link function $h(\cdot)$ with support \mathbb{R}^+ Bremaud and Massoulié (1996). Typical choices for h include $h(x) = \mathbf{1}_{x \in \mathbb{R}^+}$ and $h(x) = e^x$. These forms are infrequently used and are not an extension considered in this thesis.

1.5.1 Event times

A range of financial data sets across many markets and asset types have been used in the fitting of Hawkes processes for a broad range of applications. We provide a brief overview of how various researchers have defined the ‘event time’ of a Hawkes process before delving into discussions of the applications and overview of the literature.

Trade data was used in the first application of Hawkes processes to financial data by Hewlett (2006) and Bowsher (2007). For example, an event time may be defined by a trade (market order) arrival. Bacry et al. (2012), Bacry et al. (2013a) and Bacry and Muzly

(2014) utilized trade arrivals along with event times defined by mid-price changes and price data, in the applications of market microstructure (Section 1.5.5) and price impact models (Section 1.5.4). Alfonsi and Blanc (2016) used event times defined by trade arrivals in the application of optimal execution strategies. Whilst studies by Toke and Pomponio (2012) and Stindl and Chen (2018) considered event times defined by the trade arrivals, they also used limit order book level 2 data to create a subset of the event times.

Event times can be defined by *price data*, for example when there is a change in the traded price which may not correspond to a trade arrival if multiple trades arrive at the same price. Event times defined by price data have been used by Hawkes process applied to macroeconomic applications (Section 1.5.2) by Ait-Sahalia et al. (2015), Calcagnile et al. (2018) and Yang et al. (2018). Another variation of price data used to define event times are studies of extremal prices (Section 1.6), which define the event time as the days where the closing price is greater than some threshold (Embrechts et al., 2011), or intra-day 15 minute time buckets of returns with event times based on the time buckets that exceed some threshold (Chavez-Demoulin et al., 2005).

Event times defined by *mid-price changes* are core to the studies on market reflexivity (Section 1.5.3), with some notable research by Filimonov and Sornette (2015) and Hardiman and Bouchaud (2014) and earlier papers by the same authors. *Quote data* refers to the best bid or ask prices. An example of an event time defined by quote data is a time related to a change in the best bid or ask price. In the application of market microstructure (Section 1.5.5), Zheng et al. (2014) defined the event process as a change up or down in the best bid or ask price.

Limit order book level 1 data has been used in a number of ways in current literature to define event arrival times. For example, an event arrival time such as a market order, limit order or cancellation can be determined by a change in volume and/or price on level 1 of the limit order book. Almost all literature utilizing limit order book level 1 data to define event times, with the exception of Large (2007), has almost exclusively appeared within the last 5 years. Of this research, the applications are extensions to existing work that uses a lower frequency data set. Some examples include: Kirchner (2017a,b), who extended on earlier work on market reflexivity (Section 1.5.3); Bacry et al. (2016), who extended prior studies on price impact models (Section 1.5.4); and Lu and Abergel (2017) and Achab et al. (2017), who extended early research on market microstructure (Section 1.5.5).

Finally, we consider *limit order book depth data*, for example the event time could be defined by the event arrivals, such as market orders, limit orders and cancellations in the depth of the limit order book. To-date, there are no studies that apply the Hawkes process to limit order book depth data. Such an application will be a focus of this thesis.

1.5.2 Hawkes processes for macroeconomic applications

The first collection of research we consider includes the application of a Hawkes process for the study of market contagion. Despite the disagreements on how to define and measure contagion there is a fairly broad agreement about the channels through which a shock can be transmitted. Caprio et al. (2013, Chapter 38) present two definitions of contagion.

The first states that in a broad sense, contagion occurs when a shock from one or a group of countries, markets, or institutions, spreads to other countries, markets, or institutions. This broad definition is precipitated when: there is a spread of a financial crisis from one country to another; devaluation of currency; or if a bank experiences a run on its deposits, which leads to other banks also experiencing a run. The second definition refers to narrow contagion, which occurs when a shock from one or a group of markets, countries, or institutions, spreads to other markets, countries, or institutions and when the spread of the shock or the co-movement, caused by the shock, is excessive relative to some ‘norm’. Under both definitions contagion is associated with co-movement. The norm refers to co-movement that is greater than usual, increases and declines rapidly, and is excessive relative to fundamentals.

In the application of portfolio credit risk modelling the introduction of Hawkes processes was due to the insufficient explanation that a default intensity depended linearly on macroeconomic variables, when in fact there is a clustering phenomena of defaults around economic recession. Errais et al. (2010) introduced the concept of credit contagion. Extending on this idea, and still within the realm of credit risk modelling, Dassios and Zhao (2011) introduced a dynamic contagion process by generalizing a Hawkes process (with exponential decay) and the Cox process. This extension allowed for both self-excited and externally excited jumps, which could be used to model the dynamic contagion impact from endogenous and exogenous factors of the underlying system (Dassios and Zhao, 2011).

Aït-Sahalia et al. (2015) (with preliminary research in Aït-Sahalia and Jacod (2009)) presented a Hawkes-jump-diffusion model with exponential decay, which is a generalization of Poisson jump-diffusion model. The model incorporated elements of drift and stochastic volatility and it was applied to the study of jump excitation around 5 world markets. The aim was to create a measure of market stress, with the model presenting high self- and cross-excitement across markets during the Global Financial Crisis, consistent with the empirical data.

The application of peaks-over-threshold is presented by Schneider et al. (2018) to identify abrupt liquidity dips in the limit order book data. Descriptions of liquidity dry-ups and synchronized asset moves can be found in Section 1.1 of this thesis. Using a multivariate Hawkes process for multiple assets, Schneider et al. (2018) presented two ideas: the first being *illiquidity spirals*, which are the self-excitation of extreme changes in liquidity; and *illiquidity spillovers*, which reflect the cross-excitation across assets and can be thought of as illiquidity contagion. This method was applied to the sovereign bond market. The identification of liquidity dry-ups and spillovers is a significant risk factor for both investors and regulators. In the same fashion that Calcagnile et al. (2018) discovered a synchronization of large price moves across assets since 2001, Schneider et al. (2018) discovered a synchronization of extreme illiquidity across assets for the same period, amplifying this risk further. Their method allowed for the identification of such events and gave a directional measure for spillovers and extreme events.

Calcagnile et al. (2018) studied the high frequency dynamics of a set of U.S. stocks

from 2001 to 2013, and measured the fraction of systematic instabilities contributed to the exogenous component, being macroeconomic news. Rather than directly modelling the asset they modelled how numerous assets moved together, i.e. grouping stocks by the number of co-jumps. This was done to reduce the dimensions (> 140) of the Hawkes process. They discovered a decrease in extreme events, however a stronger synchronization of large price movements across assets since 2001. This was attributed to the greater connectedness of markets in recent times and the growth of trading algorithms. The majority of the extreme movements were endogenous in nature (Calcagnile et al., 2018).

Unmarked Hawkes processes have been used for many other financial applications. Investor sentiment (positive and negative) and market returns (positive and negative) were modelled by Yang et al. (2018) as a multivariate Hawkes process. Yang et al. (2018) found that the self- and cross-excitation features were strong for all event types across 15 minute time scales and were not present for smaller time scales. They observed that the news sentiment events spline intensity followed different patterns than the generally observed U-shaped intra-day return pattern. They observed positive sentiment shocks generated negative price jumps, presenting a contrarian response.

1.5.3 Hawkes processes for market reflexivity applications

The efficient market hypothesis, at its essence, states that the price formation fully reflects all available information and as a result markets are assumed to be driven by external inputs of information. In contrast to this, the concept of social reflexivity was first qualitatively described by Soros (1987) in the context of economics and finance.

If investors believe that markets are efficient then that belief will change the way they invest, which in turn will change the nature of the markets in which they are participating (though not necessarily making them more efficient).
–Soros (1987)

Despite a substantial part of market activity argued to be endogenous or reflexive in its origins, with exception to the ensuing literature, there has been no statistical tool to enable one to define and measure the degree of reflexivity in financial markets (Hardiman et al., 2013).

The role market reflexivity plays in price formation was explored by Filimonov and Sornette (2012). They suggested that price dynamics are mostly endogenous and driven by positive feedback mechanisms involving investors anticipation, that lead to self-fulfilling prophecies. They described the endogenous mechanism as resulting from herding and strategic orders splitting, which led to long-range correlation in the series of trade initiations (buy side or sell side). Filimonov and Sornette (2012) introduced the first quantitative measure of market reflexivity (endogeneity in financial markets), being the branching ratio ϑ of a univariate Hawkes process in (1.6). The event times occur when there is a mid-price change in the futures over the S&P 500 equity index in the U.S., with these changes representing ‘market activity’. In a financial context this may be describe as the prevalence of endogenously driven market activity,

measured as a proportion ϑ versus exogenous (external or news related) market activity. The immigration intensity η in (1.6) represents the exogenous events. A single exogenous event may trigger additional endogenous events. As noted by Filimonov and Sornette (2012), the key property of the Hawkes model is that the branching ratio represents the degree of self-excitation of the system, not the average rate of events. The three regimes that exist are, sub-critical $\vartheta < 1$, critical $\vartheta = 1$ and supercritical $\vartheta > 1$.

Studies by Filimonov and Sornette (2012, 2015), Hardiman et al. (2013) and Hardiman and Bouchaud (2014) presented a robust debate on the range of values the branching ratio has taken in financial markets since 2008 and the appropriateness of using an exponential versus a power law decay kernel within the Hawkes process. Both studies defined the event times as changes in the mid-prices for the E-mini S&P 500 futures data, from the period 5-Jan-1998 to 29-Aug-2010.

Filimonov and Sornette (2012) presented findings that the level of endogeneity as measured by the branching ratio, has increased significantly from $\vartheta = 0.3$ in 1998 to $\vartheta = 0.7$ in 2007. The underlying Hawkes process had an exponential kernel with no marks. They argued that the increase in the branching ratio is a result of increased self-excitation of the system, not trading activity, as transactions or volume do not enter the model directly. However, despite the fact that the model doesn't account for these features it does not mean that the true process hasn't been influenced by an increase or a decrease in these components. This cannot be attributed to a causal reason, it simply means the model does not address this issue.

Research presented by Hardiman et al. (2013) extends the work of Filimonov and Sornette (2012), considering the idea of a market which exhibits critical reflexivity. In context of Hawkes process, criticality relates to both long-range memory and the prevalence of endogeneity over exogeneity (Hardiman et al., 2013). A Hawkes process that is stationary with a finite mean and variance, but with long-range dependence of the event rates, must be critical ($\vartheta = 1$) (Hardiman et al., 2013). The mathematical existence of a stationary critical Hawkes process was proven by Bremaud and Massoulié (1996). Contradicting previous findings, they showed that the markets in recent times have not increased in reflexivity (endogeneity) due to automated trading, rather they have been extremely stable over time and equal to the critical value ($\vartheta = 1$). They argued that the appropriate Hawkes kernel has a power-law decay, which is approximated by a sum of exponential functions with power-law weights. They stated that the results from Filimonov and Sornette (2012), were due to the use of an exponential kernel and this led to conflicting conclusions. It should be noted that, Bacry et al. (2012), Bacry and Muzy (2014), Bacry et al. (2016), and Hardiman et al. (2013) have shown that the kernels are slowly decreasing and are well described by a power-law behaviour, with exponents close to 1. Thus, the two key findings from the research by Hardiman et al. (2013) are: the Hawkes model is very close to the stability threshold of $\vartheta = 1$, being the critical case, showing finite mean and variance with long range dependence; and the empirical kernel is well described by the power-law function.

In response to Hardiman et al. (2013), Filimonov and Sornette (2015) presented a care-

ful study of the challenges of using a Hawkes process to model high frequency financial data. They argued that factors, such as regime shifts on the parameters, due to day-to-day non-stationarity, led to an upward bias in the measure of the criticality index. Whilst the aim is to mount evidence against the claim made by Hardiman et al. (2013), that the markets have been functioning close to criticality, this research also provides a detailed description of the optimization challenges for Hawkes processes, such as edge effects. Edge effects impact simulation of long memory processes, especially Hawkes processes with power-law decay functions. This is due to events occurring before the time window not being account for. The claimed appropriateness of a power-law decay function was rebutted by Filimonov and Sornette (2015), who demonstrated that the challenges of data integrity leads to biases in microstructure research of high frequency financial data. These challenges are covered further in Chapter 3.

Finally, Hardiman and Bouchaud (2014) presented a method for simplifying the estimation of the branching ratio via the first two empirical central moments of the event count data. This method was intended to eliminate the biases raised in Filimonov and Sornette (2015) about the choice of kernel in the maximum likelihood estimation approach. Using this new method they reconfirmed their findings in Hardiman et al. (2013), that the markets are in fact critical and exhibit long memory.

Filimonov et al. (2015) extended the methods of calculating market reflexivity by providing a direct comparison of the Hawkes and ACD models, which was introduced in Section 1.4. Based on numerical simulations they showed that an introduced composite parameter into the ACD model, which serves as an effective measure of endogeneity, can be mapped onto the branching ratio of the Hawkes process. Despite the different classes of models, both exhibit similar mathematical properties, for example, the introduced composite parameter and the Hawkes branching ratio both characterize stationarity properties of the models. They were able to show the monotonous relationship between the introduced composite parameter of effective degree of endogeneity in the ACD model and the branching ratio corresponding to the Hawkes model. They have expanded the quantification of endogeneity to the class of ACD models. In a novel method of interpretation, they showed that there is strong support for the hypothesis of a dominant endogenous or reflexive component to dynamic markets.

Research by Rambaldi et al. (2018) aimed to enrich former multivariate Hawkes models with level 1 data (Bacry and Muzy, 2016) by accounting for the identity of market participants. Rambaldi et al. (2018) used data obtained on the CAC40 index future from Euronext, with the key advantage of two numeric IDs indicating who submitted the order and whether the order was sent through one or more connections. They constructed a multivariate Hawkes process with non-parametric kernels. The components reflected the number of event types (mid-price jumps, market, limit, cancel orders at the bid an ask) by the number of agents. Their research gave insight into the agent's behaviours that affect volatility, and the proportion of activity that is of reflexive nature (endogenous) or driven by some external information flow (exogenous) for different agents. High frequency traders are more endogenously driven compared to other agents. This study is not transferable

to most exchanges as market participant IDs are typically not disclosed.

1.5.4 Hawkes processes for optimal execution, price impact and order-flow applications

The link between order flows and price formation is non-trivial and it is of primary interest for practitioners when developing price impact models to measure transaction costs. From an academic stand-point, this relates to the relationship between price formation and order flow, that is, the endogenous nature of price fluctuations (Bouchaud et al., 2004). Price impact models are also important in the design of optimal execution strategies. For example, an implementation shortfall strategy attempts to address the classic execution problem of minimizing market impact cost, which is the investor's trade impact on the price of the asset, with timing risk or urgency of execution. Extensions to this are presented in Cartea et al. (2015), via three execution models where the investor can: continue execution conditional on the asset not breaching some critical boundary; incorporating order flow to take advantage of trends in mid-price, which may be due to pressures on the buy or sell side of the market; and trades in both lit and dark pool venues. For a thorough coverage of optimal strategies from a modern view of dynamic stochastic optimization, see Cartea et al. (2015).

An early application of the Hawkes process was in the area of price impact and the development of an optimal liquidity strategy (Hewlett, 2006). In a similar vein to the work by Bowsher (2007), they constructed a bivariate Hawkes process of buy and sell trades for the FX market. This model was used to predict the future imbalance of the buy and sell trades, conditional on the history of recent trade arrivals. They constructed a linear price impact function that has a closed form solution when the self-excitation function is exponential. This was used to construct a mean-variance optimization strategy for the liquidation of an asset.

To model the dependence on the bid-ask spread and generate accurate predictions of orderflow, Vinkovskaya (2014) introduce a regime-switching variant of the Hawkes process. This extension accounts for the fact that certain events sharply increase in intensity when the bid-ask spreads widen. Vinkovskaya (2014) show that the one-dimensional, multivariate and regime switching multivariate Hawkes models perform well and achieves a higher degree of accuracy in short term predictions of order flow, than the Poisson model and time series models of the AR, MA and ARIMA form.

Bacry and Muzy (2014) present the idea of a meta-order, which is the sum of smaller orders that a trader submits over a period of time to reduce the potential market impact of the trade. As Bacry and Muzy (2014) suggested, this is at the centre of market micro-structure regulation.

Bacry and Muzy (2014) defined a continuous time version of a market impact price model, which incorporates stylized facts, such as high frequency mean reversion quantified by the Hawkes process in Bacry et al. (2013a). They proposed 'superior' non-parametric techniques for kernel construction for big data sets arising in limit order book modelling. They made comparisons with earlier non-parametric methods, demonstrating the ability

to jointly capture the mid-price movements (upward/downward jumps) and market order dynamics (buying/selling market order) via a four dimensional Hawkes process. However, they did not account for the volume of the orders nor price jump size Bacry et al. (2015). They were able to reproduce microstructure noise as strong microscopic mean reversion or de-correlation of the increments. Further extensions on the impact models were made by Bacry et al. (2016), utilizing level 1 limit order book data to construct an eight dimensional counting process (upward/downward mid-price moves, market, limit and cancel orders that don't move the bid or ask), to measure exogeneity and causality between these events.

The framework by Bacry and Muzy (2014) was extended by Jaisson (2015), with the development of a diffusive framework to build a two dimensional model with a focus on meta- market orders. They assumed the impact of meta-orders is linear, the price is a martingale, and the market orders can be approximated as an unstable Hawkes process with long memory. They recovered stylized facts with a power-law impact function, whose exponent was linked to the long memory of the sign of market orders.

The standard approach of using a Hawkes process is to model the increase in jump intensity towards the close. Khashanah et al. (2018) addressed the challenge of capturing the decline in jump intensity observed at the start of a daily trading regime. They began with a univariate Hawkes process, then introduced a classic birth-death-immigration process to model a jump process with activity near the beginning of the day. They finally showed that this can be extended to the bivariate case, resulting in a self-depressing process of birth-death-immigration Hawkes process. The authors noted that the applications of this approach extended to financial risk propagation, portfolio management and algorithmic trading strategies.

Using the model presented in Aït-Sahalia et al. (2015), a natural extension to the application of optimal portfolio selection was proposed in Aït-Sahalia and Hurd (2016). They showed that due to the asset prices being subject to mutually exciting jump processes, the clustering of jumps led to time-varying optimal asset allocations. They provided solutions to the optimal dynamic portfolio problem.

Recently, Achab et al. (2017) designed a new non-parametric estimation technique for the kernels of a multivariate Hawkes process and applied the method to limit order book data, using the same data set and model structure as Bacry et al. (2016) for comparison. Their model was able to produce stylized features of the limit order book, such as the symmetry of the bid and ask, and prices being cross-excited. Also, the market, limit and cancel orders were shown to be strongly self-excited, which demonstrated persistence of order flows and splitting of meta-orders into sequences of smaller orders.

Extending their earlier work in Achab et al. (2017), Achab et al. (2018) introduced a new non-parametric method that allowed for a direct, fast and efficient estimation of the branching matrix, which can be considered a multivariate extension of the method proposed in Hardiman and Bouchaud (2014). The model structure is very similar to the 8 dimensional process presented in Bacry and Muzy (2014), however rather than just considering mid-price moves, they also made use of level 1 limit order book data. They applied this method to the EUREX exchange for a single asset, level one data for a

12 dimensional process (upward/downward mid-price moves due to market, limit and cancellation orders and bid/ask market, limit cancellation orders that don't move the mid-price) and then analysed two assets simultaneously in a 16 dimensional framework. They also considered a two asset, 16 dimensional framework to study the joint dynamics of the assets that share exposure to similar risk factors, but have different characteristics. Similarly to work by Bacry and Muzy (2014), aspects such as volume of the orders and price jump sizes were ignored.

A trade-through is defined as an order that is carried out at a suboptimal price, even though a better price was available on the same exchange or another exchange. This is often not allowed or highly regulated. The United States Security and Exchange Commission (SEC) regulation, Reg NMS, ensures investors receive the best price for their executions, thus protecting against trade-throughs. In the context of research by Toke and Pomponio (2012), Pomponio and Abergel (2013) and Stindl and Chen (2018), a trade-through is defined as a market order that is executed against more than one price level of the limit order book. It is advised in future research that a new term is coined for this type of market order to avoid confusion with the more formally defined meaning of a trade-through. Henceforth, we refer to this type of market order as a *VWAP-MO*, given the market order execution price will be the volume weighted average price (VWAP) in (1.1), due to the multiple price levels associated with the trade execution.

The concept of limit order book resiliency was proposed by Large (2007) and refers to the ability of the limit order book to replenish after a large trade has occurred. The replenishment of the limit order book after a large trade has a direct impact on the price formation process. Resiliency is captured by how large trades alter future intensities of fresh limit order submission. Large (2007) defined a 10 dimensional, multivariate Hawkes process, whereby the dimensions reflect aggressive traders who move the bid/ask and non-aggressive traders for market order, limit order and both bid/ask side. Cancellations were included for the bid and ask side. They found that aggressive limit orders were submitted to replenish liquidity taken from aggressive market order and to replace 'stale' limit orders. Replenishment of the limit order book after a large trade, only occurred 40% of the time and within 20 seconds. Consistent with this, is the study previously mentioned by Achab et al. (2018), which established that when a market order consumes the liquidity available at the best bid/ask and moves the mid-price, it has an inhibitory effect at short time scales. It is unlikely that the price will be moved in the same direction by other market orders as the price becomes more unfavourable.

The VWAP-MO is essentially the inverse of the previously described concept of a meta-order, where the meta-order is the splitting of limit orders to minimize market impact. The VWAP-MO is a submission of a large single order that consumes volume on the best bid/ask and deeper levels of the limit order book, with an investor placing a higher value on timing risk (i.e. urgency to complete the transaction), rather than market impact. An example of this, is an index fund manager rebalancing a portfolio on the quarterly rebalance dates. They have a benchmark price of the closing price of a stock on the rebalance date. Typically large volumes are traded at the end of the day to minimize the

deviation of the overall traded price and the benchmark price.

Modelling VWAP-MO has relevance to the price formation process. The resulting market impact is important for a practitioner's decision of the potential gains by electing to submit a large market order. Modelling VWAP-MO by Toke and Pomponio (2012) was motivated by the empirical evidence of clustering of event times presented in Pomponio and Abergel (2013). They used a similar framework to Bowsher (2007), by introducing an exogenous immigration intensity in a bivariate Hawkes process, with an exponential decay function of the bid and ask intensities, finding strong self-excitation, but only weak cross-excitation.

An interesting observation by Toke and Pomponio (2012) is that the aggressive markets orders (Large, 2007), defined as moving the best bid/ask, were not in perfect intersect with the VWAP-MO, that is, not all VWAP-MO move the best bid/ask price. It is not clear how this could be the case, given the VWAP-MO consumes more than one price level, thus creating a new best bid/ask. In addition, the TRTH data which is used within the research by Toke and Pomponio (2012), and discussed in Chapter 3 and in Definition 5, consolidates the limit order book at the price level, such that for side s and limit order book level l , the limit order book price is $P_{t_i}^{(s=bid,l=1)} < P_{t_i}^{(s=bid,l=2)}$ and $P_{t_i}^{(s=ask,l=1)} > P_{t_i}^{(s=ask,l=2)}$. If these two order types are not recorded in the limit order book as a single market order, respectively, then it may be the case that the VWAP-MO can walk up the book until filled. Whilst this is a very interesting research, these aspects need further clarification, with a clear distinction between how one defines an aggressive market order and a VWAP-MO before further use of the Hawkes process within this application.

A more recent study by Stindl and Chen (2018) extended the framework of the Hawkes process, to allow the immigration intensity to follow a general renewal process in (1.7), rather than a homogeneous Poisson process. They make a comparison with the model proposed by Toke and Pomponio (2012), applying their model to VWAP-MO of the same asset and similar bivariate structure representing the bid and ask. Rather than a two hour window, as per Toke and Pomponio (2012), they applied the model to the entire day. They showed that the immigration process for both bid and ask exhibits heavy clustering and over-dispersion relative to a Poisson process. They concluded that the Weibull renewal multivariate Hawkes process provided a better fit than the classic Hawkes process.

Dark pools were introduced in Section 1.2.1 as private exchanges for trading securities that are not accessible by the general public. The liquidity available on these exchanges is called dark pools of liquidity. The lack of transparency of a dark pool serves the purpose of giving an institutional investor a platform to execute very large orders, with potentially lower market impact and remaining anonymous to the market. Refer to recent work by Comerton-Forde and Putnins (2015) on dark pools and price discovery.

Gao et al. (2017) explored the application of Hawkes process to dark pools. They modelled the order execution data from a dark pool with an exogenous immigration intensity, which may represent the intra-day pattern of dark pool liquidity. Gao et al. (2017) analysed various performance metrics such as, time-to-first-fill, time-to-complete-fill and expected fill rate.

1.5.5 Hawkes processes for market microstructure applications

The price formation mechanism has been studied extensively using Hawkes processes, with the majority of applications to financial data to improve the understanding of market microstructure by reproducing stylized facts of financial markets. Market microstructure is defined by O’Hara (1997) as the study of the process and outcomes of exchanging assets under explicit trading rules. While much of economics abstracts from the mechanics of trading, microstructure literature analyses how specific trading mechanisms affect the price formation process (O’Hara, 1997).

The first application of the Hawkes process to the study of market microstructure was made by Bowsher (2007). This research presented the advantages of using a Hawkes process over autoregressive conditional duration and intensity models, and has been the cornerstone of the application of Hawkes process to financial data, with discussions of the advantages in Section 1.4. Bowsher (2007) introduced both a univariate and a bivariate Hawkes process with a vector valued exogenous immigration intensity. Bivariate in this context means that they consider two event types, the two-way interaction between timing of trades and mid-price changes, and thus have an intensity function with $j = 2$ components. The intensity provides an approximation to the instantaneous price volatility of an asset.

The data used by Bowsher (2007) was one second time granularity, meaning that trades and mid-price events could be reported on the same time stamp, breaching a core assumption of the Hawkes process of non-simultaneous event times. To address this, they adjusted the occurrence of the times of the mid-price events by a uniform random component, reasoning that the actual occurrence times (in continuous time) will rarely be the same for trades and mid-quote (Bowsher, 2007). The adjustment takes t as the original occurrence time (in seconds), then the time becomes $t - 0.5 + U$, where U is the uniform random variable on $(0, 1)$. This work is over a decade old. Nowadays, trade and limit order book datasets are reported at a minimum of a millisecond and often at the nanosecond for hedge funds that have direct access to exchanges. Despite the finer granularity of reporting, the growth in the activity of financial markets still results in a sizeable number of simultaneously reported events, even with considerably lower frequency trade data. The acknowledgement of events reported on the same event time is still a prevalent issue for data processing and the choice of Hawkes process, however the proposed adjustments may not be sufficient with the substantial growth in financial data, despite the small time granularity with which it is reported.

The next body of research we consider is the application of the Hawkes process to reproduce stylized facts that are observed in very high frequency transaction data, such as the microstructure bias introduced in the evaluation of realized volatility as the time interval becomes small. The realized volatility can be measured by the long term second moments of the number of jumps in a Hawkes process framework, which is discussed below. The bias is due to the mean reverting behaviour at short time intervals, and as presented in Bacry et al. (2013a), the realized volatility becomes unstable at very small time intervals due to this microstructure noise. A diagnostic tool for this is the volatility

signature plot, which helps reveal the severity of the microstructure bias as the sampling frequency increases (Andersen et al., 2000). As presented by Bacry et al. (2013a), if $X(t)$ is the price of some asset at time t , the signature plot can be estimated from the quadratic variation of $X(t)$ over time period $[0, T]$ at the sampling frequency (scale) $\tau > 0$, where T/τ is integer and the so-called realized volatility estimator is

$$\hat{C}(\tau) = \frac{1}{T} \sum_{n=1}^{T/\tau} |X((n+1)\tau) - X(n\tau)|^2. \quad (1.8)$$

The microstructure noise effect manifests itself through an increase in the realized variance when the time scales go from large to small, $\tau \rightarrow 0$ (Bacry et al., 2013b).

The *Epps effect* (Epps, 1979) quantifies the dependence of asset return cross-correlations on the sampling frequency. For high frequency resolution data the cross-correlations are significantly smaller than their asymptotic value as observed on daily data (Tóth and Kértész, 2009). There have been many mechanisms proposed that contribute to the Epps effect. Münnix et al. (2011) demonstrated two major causes. The first being the asynchrony of time series, which refers to a time series that features an arbitrary lag for a given point in time, but the average lag is zero. This is simply due to non-synchronous pricing of stocks, for example stocks trading at several stock exchanges simultaneously. The second cause is the tick-size, which is the lowest possible price change an asset can make. Prices tend to cluster at certain multiples of the tick-size, resulting in an effective tick size. This so-called ‘imposed discretization’ of prices leads to information loss, which grows for smaller return intervals (Münnix et al., 2011). Similarly to the use of a signature plot to visualize the change in realized volatility across different times scales, the Epps effect can be visualized in a analogous fashion. As presented by Bacry et al. (2013a), if $X_1(t)$ and $X_2(t)$ are prices of two assets, a correlation coefficient over a time period $[0, T]$ can be estimated from high-frequency price increments as

$$\hat{\rho}(\tau) = \frac{\hat{C}_{12}(\tau)}{\sqrt{\hat{C}_1(\tau)\hat{C}_2(\tau)}}, \quad (1.9)$$

where

$$\hat{C}_{12}(\tau) = \frac{1}{T} \sum_{n=1}^{T/\tau} [X_1((n+1)\tau) - X_1(n\tau)] \times [X_2((c+1)\tau) - X_2(n\tau)].$$

The first in the series of papers on Hawkes processes in the application of finance by Bacry and contributors, is Bacry et al. (2013a) and this is the only research contribution by the authors (to date) that considers a parametric Hawkes processes. Later research by this group resides within the non-parametric framework.

Bacry et al. (2013a) created a model based signature plot that reflected the statistical structure of the empirical data constructed signature plot. They used a multivariate Hawkes process with $j = 2$ components, an exponential decay function and with event times which were the arrival of upward and downward mid-price changes to reproduce the behaviour of the signature plot. To study the Epps effect they extended this model to a

multivariate Hawkes process with $j = 4$ component, which described the joint dependence of the mid-price dynamics for two futures, Euro-BOBL and Euro-Bund. They showed that closed-form expressions can be obtained for the models second-order properties at all time scales. This allowed the recovery of the major high-frequency stylized facts, such as the microstructure bias in the calculation of realized volatility as the time interval becomes smaller, and the Epps effect.

Following on from research by Bacry et al. (2013a) and their first item within the non-parametric framework, was a paper by Bacry et al. (2012). This introduced a non-parametric estimation method, calibrating the shape of the kernel function with the empirical auto-covariance of the counting process for a Hawkes process. When this methodology is applied to two futures assets the Hawkes process kernels were shown to have a power-law decay.

Two further advancements in the study of market microstructure apply the Hawkes process to higher frequency data. Zheng et al. (2014) introduced a generalization of the model by Bacry et al. (2013a), with an exponential decay function and investigated the interplay between the spread and the mid-price. This was achieved by modelling the best bid and ask prices based on events of upward and downward ticks, leading to a four dimensional non-linear Hawkes process, with additional constraints on the intensity function. They showed that the model is able to recover signature plots, with shapes that are similar to empirical shapes.

Lu and Abergel (2017) constructed the counting process from level 1 limit order book data, where an event can be a limit order, market order, or a cancellation and can affect the best bid or best ask. The events were additionally tagged according to whether they moved the mid-price or not. In total they formed $j = 12$ components for a multivariate Hawkes process.

They extended upon this model formulation by constructing a non-linear Hawkes process to address shortcomings of the linear Hawkes model to reproduce stylized characteristics, represented by the signature plots. The said shortcomings relate to inhibition effects (decreasing intensity, rather than an exciting effect) not being allowed in the linear setting, and this was overcome by introducing negative kernels. For the non-linear Hawkes process, the intensity is written as a non-linear function $h(\cdot)$ with support \mathbb{R}^+ . In the case of research by Lu and Abergel (2017), the $h(\cdot)$ denotes the element-wise positive part function. Lu and Abergel (2017) acknowledged that no two events can arrive at the same time, which would void the construction of the model they propose, stating that the high time resolution assures this. The bid and ask side events were modelled with a multivariate Hawkes process. However, as we will discover in Chapter 3, it is in fact highly likely that events on the bid and ask will occur simultaneously. This may not be the case for their specific data set of stocks trading on the Frankfurt Stock Exchange, however investigation of simultaneity of events would need to be conducted before applying this proposed model.

1.5.6 Summary

The extensive applications of the unmarked Hawkes process to financial markets provides many insights, which will be valuable when extending the Hawkes process to the marked case. These insights and research opportunities are summarized below.

- Within the construction of the Hawkes process the immigration intensity has been extended to a deterministic function of time and a renewal process in (1.7) across applications, such as the study of market microstructure, VWAP-MO and dark pool liquidity. There is reasonable evidence of the benefits of a non-constant immigration function for financial application, for example, in the application of longer intra-day time scales where event arrivals in the limit order book may be non-stationary. However, the development of the marked Hawkes process with a renewal process is beyond the scope of this thesis.
- The decay function has been specified in many different ways, with the classical approaches being an exponential or power-law decay and sum of exponentials. Newer, non-parametric methods have seen many iterations of development. The exponential decay function remains the dominate choice, however the majority of the literature often does not justify this choice, rather just stating the use. Of the few papers comparing the various options of a decay function, there is no consensus even with the same application in mind. When considering an appropriate decay function, current literature provides numerous options but very little guidance on the choice for ultra high frequency data.
- The specification of the counting process is broad and varies with the particular application to which it is applied. The majority of the literature defines the event using market order arrivals, price changes and mid-price changes. More recent literature extends existing applications to incorporate quote and level 1 limit order book data. Whilst some studies by Toke and Pomponio (2012) and Stindl and Chen (2018) use level 2 limit order book to determine a subset of market orders to model, this is an entirely different proposition to modelling limit order book arrivals. Of the literature reviewed, no applications have utilized the limit order book data beyond level 1, despite the empirical evidence to suggest that there is much to gain from extending data used in research on the price discovery mechanism, to include limit order book depth.
- Numerous variants of the functional form of the Hawkes intensity process have been proposed, for example a non-linear intensity, however there is loss of mathematical tractability for most properties of these processes. The development of the Hawkes-Cox process and the Hawkes-jump-diffusion model was motivated by specific applications, for example both jump-diffusion and Hawkes-Cox process have been used for market and credit risk contagion, respectively. These extensions are yet to be demonstrated across a range of applications, and the event processes include returns and mid-price changes, not limit order book depth/event arrivals. Whereas the linear

multivariate Hawkes process has been extensively studied and applied to financial applications, including level 1 order arrivals. On face value, the multivariate Hawkes process would appear to be a natural starting point for modelling the limit order book, with components representing levels of the limit order book. However careful consideration needs to be given to the property of a simple point process, that no two events can occur at the same time. This presents substantial challenges in data preparation to construct a suitable observable data set and the appropriate model choice, for example univariate versus multivariate Hawkes process. These questions have not been addressed by current literature, including in the setting of unmarked Hawkes processes.

- The Hawkes process provides a powerful tool for many financial applications. These extend from: market microstructure; reproducing the signature plot and Epps effect; optimal execution strategies; market impact models; detecting meta-orders; VWAP-MO and association with aggressive market orders; market reflexivity (endogeneity) and critically of markets; limit order book resiliency; credit risk contagion; measures of market stress; endogeneity of extreme moves; identification of liquidity dry-ups and spillovers; investor sentiment; and liquidity in dark pools. Of this impressive list of applications all utilize an ‘unmarked’ Hawkes process, despite many authors proposing future extensions to incorporate the substantial information available in the limit order book by the inclusion of marks. The application of market reflexivity is one area where an unmarked and a marked Hawkes process has been applied. A key finding by Kirchner (2017b), is that the exclusion of explanatory variables leads to a large upward bias in branching ratios and very heavy-tailed decay functions, when in fact they are more appropriately represented by an exponential decay (Kirchner, 2017b). This gives some preliminary insight into the necessity for the inclusion of marks to improve the specification of the model for financial application.

1.6 Marked Hawkes processes

The marked Hawkes process was originally introduced by Ogata (1988) as a model of earthquake occurrences, with the inclusion of the magnitude of the earthquake as a mark. The generalization of the Hawkes process to include marks forms the foundation of this research. The studies to-date that include marks in a financial application are Kirchner (2017b), Alfonsi and Blanc (2016), Rambaldi et al. (2017), Fauth and Tudor (2012), Chavez-Demoulin and McGill (2012), Embrechts et al. (2011) and Liniger (2009). What is distinctive between this review and that of the unmarked Hawkes process in Section 1.5, is the limited number of papers that study Hawkes processes with marks.

The definitions and properties of the univariate Hawkes process with multivariate marks can be found in Chapter 4. For the purposes of this discussion we present the multivariate and univariate marked Hawkes intensity processes, which closely follows the work of Embrechts et al. (2011) and Liniger (2009).

The multivariate Hawkes intensity process with vector valued marks has observed

points $\{(t_i, m_i, \mathbf{X}_i), i = 1, \dots, n\}$ $N_t \in \mathbb{N}$ with event times $0 = t_0 < t_1 < t_2 < \dots < t_n \leq T$ a component index $m_i \in \{1, \dots, m\}$ and marks $\mathbf{X}_i \in \mathbb{R}^d$. The j th component $j \in \{1, \dots, m\}$ intensity is defined as

$$\lambda_j(t) := \eta_j + \sum_{k=1}^m \vartheta_{j,k} \int_{[0,t) \times \mathbb{R}^d} w_j(t-s; \alpha) g_k(\mathbf{x}; \phi, \psi) N_k(ds \times d\mathbf{x}), \quad t \in \mathbb{R}, \quad (1.10)$$

where the mark impacts the intensity through the boost function $g_k : \mathbb{R}^d \rightarrow \mathbb{R}_+$ for $k \in \{1, \dots, m\}$, which is parametrized with vectors ψ_k and ϕ_k and which is required to specify the distribution of the marks.

The univariate marked Hawkes intensity process with vector valued marks has observed points $\{(t_i, \mathbf{X}_i), i = 1, \dots, n\}$ $N_t \in \mathbb{N}$ and a vector of d marks $\mathbf{X}_i \in \mathbb{R}^d$. This is defined as

$$\lambda(t) := \eta + \vartheta \int_{[0,t) \times \mathbb{R}^d} w(t-s; \alpha) g(\mathbf{x}; \phi, \psi) N(ds \times d\mathbf{x}), \quad t \in \mathbb{R}, \quad (1.11)$$

where the $d \times 1$ marks \mathbf{X} impact the intensity through the boost function $g : \mathbb{R}^d \rightarrow \mathbb{R}_+$, which is parametrized with a vector ψ and ϕ .

The boost function controls how strong the effect of the mark of some event is on the intensity and is influenced by both the time lag and the mark value after an event. The marks are assumed to be independent of the past intensities of the Hawkes process. The marks allow a complete specification of the dynamics of the Hawkes process (Liniger, 2009). The boost function can take many functional forms, such as polynomials as in Embrechts et al. (2011) and Kirchner (2017b), exponential, or power law as in Fauth and Tudor (2012). Boost functions can be combined multiplicatively or additively (refer to Section 4.2.1 for further discussions on boost functions)

As we will present in Section 1.7, there are many well established stylized features of the limit order book. It is a purpose of a model of the limit order book to best replicate these stylized features to describe the underlying dynamics and price discovery mechanisms. Within the financial data sets there is a wealth of information that can contribute to the specification of the models. A very simple example being, if a trader submits a market order to buy a stock, will the intensity of the limit order book change if the volume is very large versus the volume being very small? This information and so much more, can be captured by marks.

Excluding marks from the modelling fails to accurately capture the dynamics of the limit order book in all applications, whether it be the link between order flows and price formation, the endogenous nature of price fluctuations, predicting buying and selling intensities for the purpose of constructing an optimal liquidation strategy, measure market reflexivity, market stress, capturing volatility clustering and more.

Why is there so little research on the Hawkes process with multivariate marks? It is not difficult to define marks that could be included in a Hawkes process. For example, one can capture the volume of a market order, limit order and cancellation, or easily define the spread. The challenge lies within the generalization of the Hawkes process to capture the distributional properties of the marks. Hautsch (2012) aptly presents some of the

challenges that a researcher faces when incorporating statistical features, being the serial correlation, long memory properties and the specification of an appropriate parametric distribution, which may be heavy tailed, continuous or discrete.

Secondly, there is a significant computational challenges related to full joint likelihood parameter estimation of the Hawkes process. This has proven challenging even in the simplest case of a univariate Hawkes process, largely due to the vast volumes of data and the appropriate specification of the decay function. A number of researchers have attempted to address this issue by proposing non-parametric kernels (see review by Bacry et al. (2015)) and the integer valued autoregressive (INAR) approximations by Kirchner (2016). The number of marks that could be defined for a single asset limit order book are upwards of hundreds, if not thousands. Even if it were possible to accurately specify the statistical properties of the marks included in a Hawkes process, assessing their statistical significance would be impossible with the available methods.

Within the limited literature on the application of marked Hawkes process to financial data, there have been two approaches of including marks in the model for the intensity process. The first approach by Embrechts et al. (2011), Chavez-Demoulin and McGill (2012), Fauth and Tudor (2012) and Kirchner (2017b) is via a boost function, which is the approach we propose in this research. The second method used the marks to create a multivariate Hawkes process by Rambaldi et al. (2017). We will proceed by reviewing the applications of each in turn.

Hawkes processes are used in the modelling of *extremal returns*, due to the ability to capture volatility clustering. Embrechts et al. (2011) and Chavez-Demoulin and McGill (2012) both applied a Hawkes process with multivariate marks in the modelling of extremal behaviour of financial time series. Embrechts et al. (2011) presented two illustrative examples of the application of the Hawkes process with multivariate marks to capture the extremal clustering of daily and intra-day returns of the Dow Jones Industrial Average over a period of six years, marked with the absolute values of excess. The significant contribution in this paper was the presentation of simulation and estimation of univariate and multivariate marked Hawkes processes, with an exponential distribution and a linear boost function. Embrechts et al. (2011) incorporated a Gaussian copula to capture the cross dependence of the marks. In addition, they presented a univariate Hawkes process with marks, modelling the excess portfolio returns of three indexes. The marks were the absolute value of excess returns for the three indexes, which were modelled with a gamma distribution.

Studies by Chavez-Demoulin and McGill (2012) built upon earlier work of Chavez-Demoulin et al. (2005) and proposed new risk measures for the application of high-frequency trading. They presented a family of Hawkes processes that appropriately capture the volatility clustering behaviour of the intra-day extremal returns. The marked Hawkes process for the excesses over high threshold, had a Pareto distributed mark of excess sizes to capture the high kurtosis and volatility clustering of log returns, with a linear boost function. The Hawkes process provided a suitable estimation of high-quantile based risk measures, such as Value-at Risk (VaR) and Expected Shortfall (ES).

For completeness, we review a large body of work by Embrechts and Kirchner (2018), Kirchner (2016), Kirchner and Bercher (2018) and Kirchner (2017b). Their final application presented in (Kirchner, 2017b) contributes to work in *market reflexivity*. As described by Embrechts and Kirchner (2018), the Hawkes graph summarizes the immigration and branching structure of a multivariate Hawkes process as a directed graph with weighted vertices and edges. The vertices represent the possible event types and the vertex weights correspond to the immigration intensities. The Hawkes skeleton is the Hawkes graph, disregarding the weights, with the main purpose to lay the groundwork for the graph estimation, and which serves to reduce the complexity of the a priori fully connected network. The Hawkes graph estimation quantifies the estimated excitements. The method is a useful tool for preliminary analysis when examining large, multi-type event-stream data in the Hawkes framework.

Kirchner (2017b) used a non-parametric estimation method to estimate the Hawkes Skeleton and Graph from the data. The aim was to use this method in the preliminary analysis for the specification and fitting of the parametric Hawkes model. Kirchner (2016) presented the interrelationship between the Hawkes process and the INAR process, with the INAR process being interpreted as a discrete-time version of the Hawkes process. They developed an estimation method in Kirchner (2017a), which also leveraged the research in Embrechts and Kirchner (2018) and compared this with the performance of the MLE in Kirchner and Bercher (2018). They concluded that whilst the INAR has a larger standard deviation than MLE, the computational efficiency is a compelling reason to choose the INAR representation.

Kirchner (2017b) extended the work of Filimonov and Sornette (2012, 2015), Hardiman et al. (2013) and Hardiman and Bouchaud (2014) in the financial application of market reflexivity by considering higher frequency data and marked Hawkes processes. Order arrivals in the limit order book were modelled as a multivariate Hawkes process, with marks of order size and the state of the limit order book characterized by the limit order book imbalance. Two stocks from the NASDAQ exchange were considered. They used the non-parametric method of Kirchner (2016) for the decay function and utilized research by Bacry and Muzy (2014) to extend their method to the marked case. They chose a power-law decay and a linear boost function for their model. Kirchner (2017b) concatenated the baseline intensities, branching coefficients and parameters for decay functions, boost functions and mark densities and created an approximate log-likelihood. A specification and estimation of the mark distribution was not considered. Their results for the branching ratio differed greatly from the critical case reported in Hardiman et al. (2013) and Bacry and Muzy (2014). They argued that the exclusion of explanatory variables led to large branching ratios and very heavy-tailed decay kernels. They did not support the findings of markets being close to criticality levels, rather, their findings were in line with Filimonov and Sornette (2012, 2015). This is an important finding and provides further support to the necessity to consider marks in the Hawkes process. However, in contrast, Filimonov and Sornette (2012) considered including functions of volumes or increments of price movements, but based on the excellent fit of the unmarked process, they argued

that fitting a model with intensity boosted by marks was not justified, but without further detail provided.

Fauth and Tudor (2012) modelled the limit order book of the FOREX market currencies EUR/USD and EUR/GBP with a four component multivariate Hawkes point process. They studied the bid and ask quote, when a currency had an up or down move as a result of an event type, limit order or market order. Fauth and Tudor (2012) incorporated a mark, being volume associated with the upward or downward move of the currency via a normalised power boost function (as proposed by Liniger (2009) and Embrechts et al. (2011)). They found that small and large volumes impacted the arrival intensities in different ways, thus supporting the need for a marked point process.

Alfonsi and Blanc (2016) utilize a Hawkes process within the framework of an *optimal liquidation strategy* to model the flow of market orders. They considered a bivariate marked Hawkes process for event times being the buy and sell market order arrivals, with an exponential kernel. This work follows on from Bacry et al. (2013b) and Bacry and Muzy (2014), who studied a price impact model of market buy and sell, with up and down movements. However, this model is extended to include marks representing the volumes of the incoming market orders. They solved the optimal execution problem, which represents the strategy with the minimum expected cost, explicitly in the mixed-market-impact Hawkes price model. They showed that Poisson rate arrivals of orders led to price manipulation strategies, which are liquidation strategies with negative expected cost. Using the proposed mixed-market-impact-Hawkes price model, the price manipulation strategies disappeared.

The second approach, and also with the application of market microstructure, creates a multivariate Hawkes process by using information associated with the events. For example, Rambaldi et al. (2017) binned the data according to volume size and used this to create the components of a non-parametric Hawkes process. By studying two futures contracts, German Bund Future and the DAX Future, level one limit order book data, for 16 months, Rambaldi et al. (2017) demonstrated that the non-parametric kernels do not have the same shape, so the volumes on the limit order book are not independent of the point process. In a similar vein and conclusion to Fauth and Tudor (2012), but using this different structure, the research by Rambaldi et al. (2017) demonstrated that big volumes have an impact on small, but not vice versa. They also showed that trades, particularly large ones, have a significant influence on the arrival of limit orders on the opposite of the limit order book. In addition, the use of non-parametric kernels is argued to be appropriate when modelling large amounts of data and when the kernel is not localized. It is important to emphasize that within this setting, they are not modelling the mark vector as a boost to the intensity function via some boost function, rather they stratify the data by the mark and using this to create the components of the Hawkes process. This method is limited in its ability to include many marks simultaneously, by way of construction of the multivariate Hawkes process.

Summary

The main points from the literature on the application of a marked Hawkes process to financial data are:

- The literature does not provide guidance of potential marks for a model of limit order book dynamics. Firstly, and to-date, the selection of marks in the literature is limited to the absolute value of excess returns and limit order and market order event volumes. The research does not present alternative marks and there is limited motivation presented for the selection utilized. Secondly, the models only consider returns series of best bid and ask of the limit order book, not the limit order book depth which is the focus of this thesis;
- The literature does not provide guidance on the appropriate boost functions, with only linear and power functions presented. The detailed study of the distributional and dependence properties will be required to guide the appropriate selection of boost functions for the inclusion of marks into the Hawkes process;
- The research does not present formal studies on properties of the marks, rather they have been assumed. The assessment of the distributional form of the marks and studies such as serial dependence in the marks were not presented, despite these being well documented features in empirical studies (Hautsch, 2012);
- Guidance of potential marks will therefore come from literature presented in Section 1.7, concerning empirical features of the limit order book.

1.7 Foundation for marks identification

Extensive empirical studies on financial data provides insight on the stylized features of the limit order book, which we may wish to capture with a model of the dynamics of the limit order book. These studies are a stepping stone to the identification of appropriate marks. There have been extensive studies of the statistical properties of high frequency data, however the majority of the literature primarily focuses on returns series of market orders, or more commonly referred to as trade returns. Whilst the two are not mutually exclusive, it is important to distinguish between trade return-based studies and the studies relating to the statistical properties of the order flows on the limit order book, which is a focus of research in this thesis. The brief summary below will provide an important back-drop to mark identification and associated statistical properties. We refer the reader to the surveys by Chakraborti et al. (2011), Gould et al. (2013) and Abergel et al. (2016) for a more comprehensive review on empirical studies of limit order books.

Early studies of the statistical properties of the limit order book by Biais et al. (1995), Gopikrishnan et al. (2000), Challet and Stinchcombe (2001), Maslov and Mills (2001), Bouchaud et al. (2002), Zovko and Farmer (2002), Potters and Bouchaud (2003) and Smith et al. (2003) studied themes such as: order flow rates conditional on the state

of the limit order book; price impact and market order aspects; distribution of order size; depth profiles; cancellation rates; and more.

The majority of the empirical studies relating to the shape of the limit order book considered the limit order book in its entirety, rather than looking at the volume on the individual levels. Using equity limit order book data from Paris Bourse,¹⁸ two studies by Biais et al. (1995) and Bouchaud et al. (2002) originally proposed that the shape of the limit order book was monotonically increasing away from the best bid/ask with the highest level of flow occurring at the best bid/ask. Consistent with these findings, a study by Challet and Stinchcombe (2001) considered the limit order book for the Island ECN¹⁹ and described the shape of the limit order book as convex and peaking at the best bid/ask. However, more recent studies contradict these findings. An investigation by Potters and Bouchaud (2003) on the statistical properties of the NASDAQ used a zero-intelligence model to reproduce the empirical results and described the shape of the order book as being humped (ie the maximum being away from the best bid/ask), due to the non-uniform, power-law distributed flow of incoming orders. This study raised the important consideration of non-uniform cancellation rates, due to the higher frequency of amendments of orders at, or around the best bid/ask. These findings have been further supported in studies by Gu et al. (2008) on 23 listed stocks on the Shenzhen Stock Exchange and Chakraborti et al. (2011) on the Paris Bourse, both observing the maximum away from the best bid/ask.

When considering the distributional properties of the volume on the limit order book, studies were categorized into those considering the volume at each price point (tick) away from the best bid/ask and those considering volume at each consolidated price level. Bouchaud et al. (2002) described a Gamma distribution to be the best fit for volumes at each price point. They found that the Gamma distribution (for three stocks) had a scale parameter of between (0.7000, 0.8000) and a shape parameter of approximately 2,700. Gu et al. (2008) also considered volume at a price point and found that the volumes on the best bid/ask were best represented by a log-normal distribution. For order sizes that are smaller than average, they found that the distribution deviates from a log-normal distribution and exhibits power-law behavior with exponents for different levels: 4.1900 ± 0.0900 for level 1, 2.6100 ± 0.0300 for level 2, 2.6700 ± 0.0500 for level 3. Using NASDAQ level II data, Maslov and Mills (2001) conducted an empirical study of statistical properties of the limit order book, considering volume at each consolidated price level. They initially investigated a power law distribution for limit order sizes and found that it was consistent with an exponent of 1.0000 ± 0.3000 . They also implemented a better fitting log-normal distribution to the data, which has an effective power law exponent equal to 2 in the middle of the observed range.

The U-shaped curve (volume smile) which is described by Biais et al. (1995), is well known for trade data and it has also been shown to be applicable to limit order data. Both Challet and Stinchcombe (2001) and Chakraborti et al. (2011) demonstrated that the U-

¹⁸Paris Bourse was the historical Paris Stock Exchange and it is now known as Euronext Paris.

¹⁹Island ECN was an alternative trading system in the U.S., where prices were aligned with prices on NASDAQ. Island ECN merged with Instinet in 2002 and was subsequently acquired by NASDAQ in 2005.

shaped curve does exist for limit orders. Challet and Stinchcombe (2001) also showed that orders have a tendency to cluster, both in size and position, they tend to have a size of a multiple of 10, 100 or 1000, and to be placed at round prices, or at halves. Other research on the limit order book volumes has not indicated this feature and it may be specific to the ECN data used, rather than a general feature of a limit order book.

In a number of settings, it has been hypothesized that larger volumes lead to more market order submissions. This can be described as traders becoming more aggressive, leading to more trades executed. Consistent findings in support of this hypothesis have been presented by Engle and Lunde (2003), Ranaldo (2004), Lo and Sapp (2010), Fauth and Tudor (2012) and Rambaldi et al. (2017). Engle and Lunde (2003) presented a model that revealed larger volumes lead to informational traders, those which are more aggressive and likely to submit a higher number of market orders. Ranaldo (2004) studied seven hypotheses in relation to a traders aggressiveness, with limit order submission as passive and market order submission as aggressive. They found that patient traders became more aggressive under a number of limit order book conditions, for example, when the opposite side of the book is thicker (when aggregate volumes are larger). Research conducted by Lo and Sapp (2010) considered the trade off between order aggressiveness and quantity when an order is submitted. They found that more order depth (on the best bid/ask only) encouraged market orders initiated from the opposite side of the market, but discouraged limit orders.

Another important feature demonstrated by Gu et al. (2008) was the assessment of temporal dependency. They characterized the temporal dependency by the Hurst index and compared the first 3 levels of the limit order book. The results were consistent with long memory, with Hurst indexes significantly larger than 0.5 for both bid and ask side. Bartolozzi et al. (2007) investigated the temporal evolution of the *local* Hurst exponent to explore the dynamical properties of the correlations in different futures contracts. Abergel and Jedidi (2013) began with a simple agent-based market model and developed an order book model as a multidimensional, continuous time Markov chain. This was used to show how elementary changes in the price and spread are linked to the shape of the order book and order flow.

Challet and Stinchcombe (2001) studied the rate of cancellation of orders and found that the average lifetime of a limit orders fits a power law distribution. Chakraborti et al. (2011) also calculated the cancellations, although their methods were somewhat crude given the use of Level II data and the approximations applied. They observed a power law decay for both the limit orders that were subsequently cancelled and the limit orders that were executed, resulting in a market order.

Relative price is defined as the price away from the best bid (or ask) that a limit order or cancellation is submitted. Arrival rates of orders have been shown to be related to the relative price, rather than the actual price (Gould et al., 2013; Biais et al., 1995; Bouchaud et al., 2002; Potters and Bouchaud, 2003; Zovko and Farmer, 2002) and do not depend on features such as spread or mid-price. These studies also reported that relative price exhibits a power law distribution with varying values of power-law exponents.

Numerous studies (Biais et al., 1995; Lo and Sapp, 2010; Toke and Pomponio, 2012; Engle and Russell, 1998; Rinaldo, 2004; Hall and Hautsch, 2006; Cao et al., 2008) have shown that arrival rates are dependent on spread, with an increase in the spread associated with an increase in the arrival rate of orders and/or increase in limit orders arriving inside the spread. In the context of ACD modelling and for transaction data only (market order), Engle and Russell (1998) interpreted an increase in transactions rates (increased market order) and price movements if the spread is relatively wide, due to informed trades being active. In the narrow spread case, they suggested that liquidity traders are inferred to be dominant. An extension of the ACD models to include trades and quotes was conducted by Engle and Lunde (2003) with consistent findings. They found that the most conspicuous of the explanatory variables was the change in spread, whereby a rise in spread leads to a rise in the trading (market order) intensity (Engle and Lunde, 2003). Consistent to the above studies, Ellul et al. (2003) conducted a study on limit order book data from the NYSE and found that narrower spreads decrease the probability of market orders.

A measure of higher volatility should suggest that the asset price is expected to undergo larger price changes than an asset with lower volatility (Gould et al., 2013). However, methods of estimating volatility of the limit order book are subject to micro-structure noise, rather than meaningful price changes. As discussed in Gould et al. (2013), submission of a limit order, followed by an immediate cancellation within the spread, causes microstructure noise, which is not a meaningful change in price. The choice of how one measures volatility becomes crucial.

Rinaldo (2004) hypothesized that changes in the limit order book, such as spread and volatility, affect the limit and market order trading in opposite ways. He showed that the the marginal probability of a buy LO will be positive if there is more volume on the bid side, less volume on the ask side, decreased spread, decreased volatility and decreased speed of order submission (vice versa for sellers and inverse for market order). In-line with these findings, Ahn et al. (2001) demonstrated that when there is an increase in volatility on the bid (ask) side, investors will submit more limit order on the ask (bid) side than market order on the ask (bid) side. In the context of ACD modelling, Engle and Russell (1998) showed that volatility was a key factor, with increased volatility associated with periods of higher transactions rates (market order arrivals). This was consistent with findings from Rinaldo (2004) and Ahn et al. (2001).

1.7.1 Summary

Some key findings that will provide the foundation of the construction of marks for a Hawkes process with multivariate marks are:

- The shape of the limit order book has been described as humped, with the maximum volume away from the best bid or ask. Volumes are an important consideration when studying the dynamics of a limit order book. Studies that only consider level 1 or trading data, as is the case for the financial applications of Hawkes processes, are potentially precluding important information from the process;

- The distributional features of volumes of price points include Gamma and log-normal. Literature on volumes consolidated at the price level suggest a log-normal distribution. The research on distributional properties of volumes is largely inconclusive and given the important role that volumes play in limit order book dynamics, the statistical features of volume profiles should be studied up-front and in detail;
- The relationship between volume and market order flow has been well studied. In addition, limit order book depth has been linked to the choice between submission of a limit order versus a market order. This highlights the importance of both the role of volumes in limit order book models and the importance of not precluding deeper levels of the limit order book in studies of market microstructure;
- Features such as changes in price and spread have been linked to order flow. Spread has been associated with an increase in arrivals of limit orders;
- Temporal dependence has been studied in the first 3 levels of the limit order book, with results being consistent with long memory. When investigating the statistical features of proposed marks, serial dependence will need to be assessed;
- Arrival rates of orders have been shown to be related to relative price, i.e. price away from the best bid or ask for a limit order or cancellation;
- Volatility has been associated with an increase in market order arrivals.

Given the literature on marked Hawkes processes is insufficient to inform the selection of marks for a model of the limit order book, we rely on the research on the stylized features of the limit order book. This research initiates the idea generation of possible marks that can be derived from trade and limit order book data for the inclusion into a Hawkes process. The scope of the research in this thesis considers endogenous marks only, however it should be noted that future studies should also consider exogenous based marks. Despite the empirical research on limit order books initiating research on mark identification, many challenges still lie ahead for the final construction of a Hawkes process with multivariate marks. For example: how to construct the marks from the trade and limit order book data; identification of other endogenous aspects of the limit order book that might be informative for a limit order book model; defining the marks mathematically; studying the statistical properties and dependence features of the defined marks; investigating the appropriate formulation of the boost functions and marks distribution; given the vast number of potential marks, investigation into a rigorous selection methods of marks; and the parameter estimation via the full joint likelihood with the added complexity of possibly many marks that may not be i.i.d. and may not have easily identified distributional properties.

1.8 Theoretical properties of the Hawkes process

From the research that has been reviewed, it is apparent that the marked Hawkes process for financial applications has many benefits and to-date, has been underutilized. There are many marks that can be derived from the trade and limit order book data, however fitting challenges are likely to exist due to the statistical properties that have been noted in the empirical studies of limit order books. This will be further amplified in the attempt to fit a Hawkes process with many potential marks and very large limit order book data sets. There is no selection criteria for marks that does not involve the full joint likelihood method for parameter estimation. To apply a marked Hawkes process using limit order book data in the study of price formation, it is paramount that a screening tool is developed for mark selection. To that end, we present a brief review of relevant theoretical literature that will support the derivation and construction of such a screening tool.

The majority of the literature on the development of theoretical aspects of the Hawkes process relates to the linear case, with Daley and Vere-Jones (2007) and Liniger (2009) providing excellent surveys. The log-likelihood and associated asymptotic statistical properties for the marked Hawkes process has been extensively studied by Ozaki (1979), Ogata (1978) Anderson et al. (1996) and Clinet and Yoshida (2017). The Ergodic properties of an unmarked Hawkes process were established by Bremaud and Massoulié (1996).

Ozaki (1979) presented the maximum likelihood estimation procedure for the unmarked Hawkes process with exponential decay and further verified this with simulations. Ozaki (1979) provided an explicit expression of the log-likelihood of the model, its gradient and Hessian. Ozaki (1979) showed that the asymptotic properties of the likelihood estimates, based on the stationary and non-stationary version, are equivalent for the exponential decay Hawkes process.

Ogata (1978) provided proofs and developed the asymptotic properties of the maximum likelihood estimator for both the stationary and non-stationary versions of the Hawkes process, proving that the estimator is consistent, asymptotically normal and efficient. He considered theoretically, a conditional log-likelihood under the information from the infinite past.

Clinet and Yoshida (2017) derived the asymptotic properties of the Quasi Maximum likelihood estimator and the Quasi Bayesian estimator, where the quasi likelihood was constructed, using observations available over $[0, T]$. For the unmarked and exponential decay Hawkes process Clinet and Yoshida (2017) showed, using results of Bremaud and Massoulié (1996), that a suitable probability space exists on which a stationary version can be defined, and for which the non-stationary version of the intensity process converges suitably to the stationary version.

The log-likelihood for the stationary marked case with infinite past history was formally defined in Liniger (2009) and Embrechts et al. (2011). The impact of the marks enters through a boost function, which multiplicatively modifies the decay function specifying the intensity process. Liniger (2009) presented the calculation of moment measures and the existence and uniqueness of stationary solutions. For computations, (Liniger, 2009) noted that a truncated version of the log-likelihood must be used.

1.9 Contributions

- *Limit order book volume profiles.* The heavy tailed features of limit order book volumes on levels 1 to 5, aggregated to short, evenly spaced time intervals are investigated and found to require a variety of heavy tailed distributional models to adequately capture their statistical features.
- *Accurately describing the limit order book and identifying marks.* A detailed analysis of the process by which the physically operating order book is transformed into data suitable for analysis is presented. The description of this process is novel and important, because it establishes clearly the limitations of available limit order book data for point process modelling. The limit order book is found to consist of events that frequently occur at the same time, breaching a core assumption of any regular point process, particularly the multivariate Hawkes self-exciting point process often used in this field. To overcome the challenges arising from simultaneity of events in the observed limit order book, events that occur at the same time are treated as a single event and the additional information of the level and type of orders occurring is incorporated in marks attached to these events. This leads to consideration of marked Hawkes self-exciting processes to describe clustering of events in the limit order book. Empirical evidence of the appropriate decay function is presented. Substantial empirical research, guided by available literature, is presented to identify and describe a large number of potential marks, which could likely impact the intensity of the point process being modelled.
- *A score test for the detection of marks.* Joint estimation via maximum likelihood of the Hawkes process parameters and those needed to describe the marks distribution is challenging. In view of the number and complexity of the marks identified as being relevant for modelling the limit order book, a method for screening marks that is computationally straightforward to implement is developed. This new approach is based on the score test, leading to a test statistic requiring the unboosted Hawkes process to be fit once, to the sequence of observed event times. The estimates for the test statistic can be obtained parametrically or non-parametrically from the moments of suitable functions of the marks. The test has an asymptotic chi-squared distribution under the null hypothesis that the marks do not impact intensity. Extensive simulations are presented to confirm the utility of this for realistic models and sample sizes. Additionally, extensive simulation studies of the power of the new test statistic are presented. Extensions of this method to serially dependent marks are developed, something which proves to be essential in detecting appropriate marks in the limit order book.
- *Hawkes process with multivariate marks for the limit order book.* A detailed description of the univariate Hawkes self-exciting point process with multivariate marks is presented. Dependence between marks is modelled using various copulas with heavy tailed marginal distributions. Methods for simulating and estimating, via maximum

likelihood estimation of these marked processes is summarized. A description of the substantial MATLAB implementation developed for this research is given. Application of the score test is made to futures data. The marks identified as having significant impact on the intensity are shown to be serially dependent, leading to the need for a new, ‘decoupled’ approximate method of likelihood estimation. Simulation experiments show that this method reduces the modelling assumptions on the statistical properties of the marks and leads to estimation of the Hawkes process parameters and the boost function, which has good performance.

1.10 Organization of the thesis

Chapter 1

This chapter introduces the role that limit order books for derivatives play in financial markets and financial motivation for modelling the order book. Based on a literature review of its empirical features, motivation is given for developing marked Hawkes point process models for the limit order book. A detailed review of Hawkes processes in financial application is presented and research areas not addressed by the current literature are identified.

Chapter 2

Chapter 2 considers the heavy tailed features of the limit order book volumes at level 1 to 5 and aggregated to short, evenly spaced time intervals. Heavy tailed statistical distributions are found to be required to adequately capture the statistical features of volumes. A variety of heavy tailed distributional models are assessed. The methods summarized and developed in this chapter, establish the necessary foundation required to model other heavy tailed features which occur in many marks that may be incorporated in the Hawkes process.

Chapter 3

Chapter 3 extends the studies to irregularly spaced time intervals. We present the complicated structure of limit order book data and trade data, and the challenges associated with matching and modelling this data. The development of a comprehensive description of the limit order book data and the development of advanced algorithms to match and aggregate the data, such that maximal information is retained for construction of marks is presented. The identification of the limit order book events that regularly occur simultaneously, thus breaching a core assumption of the Hawkes process, highlights the unsuitability of the multivariate Hawkes process. The univariate Hawkes self-exciting point process, which can be enhanced via the inclusion of marks, is proposed as a way to model event clustering on the limit order book.

Chapter 4

Chapter 4 defines the Hawkes process and we present algorithms used for the full joint likelihood parameter estimation, with extensions to incorporate copula models for joint dependence between the marks. Extensive simulation studies explore the challenges of fitting a complex model structure, which prevents the application of marked Hawkes process in most literature. Recommendations are made to address these challenges.

Chapter 5

Guided by literature on the stylized features of the limit order book, Chapter 5 builds on the work from Chapter 3 and proposes a selection of marks derived from the limit order book. The inclusion of marks, which provide a summary of information about the nature, level and associated information of the amalgamated events to be modelled, can enrich the Hawkes process. However, the number and complexity of potential marks highlights the necessity for new methods for selecting which marks have significant impact on the intensity process. Furthermore, evidence is provided necessitating a method for detecting marks that is robust to assumptions, such as distributional shape of marks and serial dependence.

Chapter 6

Chapter 6 introduces a detection method based on the likelihood score statistic, which allows for large scale screening of marks without full model fitting. The score statistic is asymptotic chi-squared with r degrees of freedom. Under various combinations of sample sizes, boost function and marks density, we show that the chi-squared distribution can be used for obtaining sufficiently accurate quantiles for application. We present simulation studies that conclude that the score test used for marks with joint dependence, serial dependence and heavy tailed marginals, has excellent power properties. Coupled with the ease of implementation, this makes the score test a powerful and flexible tool to use when identify appropriate marks for the Hawkes process.

Chapter 7

Chapter 7 brings together the various strands of research in the thesis to apply the score test to the screening of the marks identified in Chapter 5. We fit the Hawkes process with multivariate marks to futures order book data. Fitting the Hawkes process to limit order book data via a likelihood based method, presents substantial practical challenges, which are investigated via simulation of a variety of model formulations. A decoupled approximate likelihood method is proposed to address the challenges present in the identification of suitable parametric distributions for the marks.

Chapter 2

Statistical properties of the limit order book volume process

In Section 1.1 we introduced the critical role that market liquidity plays in the price discovery mechanism of well-functioning financial markets. With the majority of exchanges now electronic, and with the evolution of the limit order book (LOB), there has been a shift in the traditional price discovery mechanism relying on trades to a far more complex process of trade and LOB data. The complexity of the interconnected financial markets, and the increasingly elaborate trading instruments and strategies executing on electronic exchanges, have added extensive complications in understanding liquidity. However, undisputed is the role liquidity plays for both market participants transacting and regulators designing policies for financial markets.

Market regulation aims to improve the quality of price determination in markets. Regulations, which were discussed in Section 1.1, for example Reg NMS in the U.S. and MiFID II in Europe, require a thorough understanding of the part liquidity plays in price discovery, to ensure the design of policies and procedures improve the price formation mechanism and limit degradation in market quality and available liquidity during periods of crisis or market malpractice. For example, MiFID II Liquidity Provision Agreement¹ provides incentive schemes, such as reimbursement of trading fees for transactions executed during stressed markets. This is conditional on firms meeting certain requirements, such as providing liquidity in times of increased volatility. Understanding the dynamics of price formation, and in particular, the characteristics of liquidity, provide the foundation for developing models for the limit order book process for simulation purposes, which can be beneficial to test regulation and to test price formation under different conditions and scenarios. Further discussion on the importance of this research in context of electronic exchange regulation can be found in the published work of this research, Richards et al. (2015).

As we discovered in Section 1.7 there is limited literature on the statistical properties of volumes in the LOB. Of these limited number of studies, the volumes considered are aggregated at a fixed grid of price points (ticks) away from the best

¹https://www.eurexchange.com/blob/3086116/2853e6f23dc29c94ba4b309ff32c76e8/data/20170629_mifid-ii-market-making-and-liquidity-provisioning.pdf

bid and ask, and typically don't assess intra- and inter-day volume features. The work undertaken in this thesis on this aspect of LOB modelling extends from papers such as, Biais et al. (1995), Challet and Stinchcombe (2001), Maslov and Mills (2001), Bouchaud et al. (2002), Gu et al. (2008), Chakraborti et al. (2011) and Gould et al. (2013). Previous studies have primarily focused on the two parameter (shape, scale) light tailed gamma distribution family (Bouchaud et al., 2004), which we demonstrate is limited in its ability to accurately capture the skewness and kurtosis features of intra-day volume processes for a range of assets. Considerations of tail properties of the volume process have not been explored in previous studies on LOB volume data. In addition, previous studies only consider assets from a single exchange, whereas this study considers futures comprised of different asset classes across five different exchanges. The time frame of estimation of both the inter- and intra-day level is across one year, which represents a more comprehensive and extended study compared to those carried out in the literature on this aspect of LOB liquidity modelling.

The aim of this chapter is to create the building blocks for a flexible dynamic model of the LOB, which we've identified in Chapter 1 to be the Hawkes process. It is natural to consider volume up front as it plays a key role in models of order flow in the LOB, which is well established in empirical literature (Section 1.7). LOB volumes will form the core component for the construction of many marks, that may be considered in a marked Hawkes process. For example, depth of the LOB, depth at the price level, LOB imbalance, volumes of different order types, to name a few (Chapter 5). In addition, volume has been utilized as a mark in financial literature applying marked Hawkes process.

Of the many ways that the volume process can be defined, we consider the instantaneous volume on the LOB at each price level on regularly spaced time intervals. Classical estimation and empirical analysis is more amenable to equally spaced time intervals. Within the classical framework, the study of the statistical and distributional properties will provide valuable insights into the challenges that may need to be considered in the development of the observable data sets on irregularly spaced time. This definition will be advanced to irregularly spaced time intervals in Chapter 3.

Definition 1 (Volume process). *The volume process $V_{t_i,d}^{l,a} \in \mathbb{R}_+$, is the instantaneous volume (snapshot) at time t_i , which is equally spaced, taking spacing of $i \in \{1, 2, 5, 10\}$ seconds. The LOB price level l contains levels for both the bid and ask side, where the bid side is $l \in \{-10, \dots, -1\}$ and the ask side is $l \in \{1, \dots, 10\}$. Further to this, $l = -1$ is the price level of the best bid and $l = 1$ is the price level of the best ask. The volume process is defined for asset, a and trading day, d .*

Within this chapter we make inference on common features of several LOB multivariate stochastic structures, with the focus on sub-exponential behaviour in the tails of the marginal distributions of the volume processes, for each level of depth on the bid and ask. We study long memory of the volume processes via the Hurst exponent and autocorrelation for lower sampling rates. We proceed by fitting a range of flexible statistical models to the LOB volume processes, to investigate these features on an intra- and inter-day time scale, which will help inform potential distributional specifications for the marks of a

Hawkes process in Chapter 5. These models include the following families, generalized extreme value distributions, the generalized Pareto distributions and the univariate α -stable distributions. These parametric models, developed for heavy tails in the continuous case, have a well understood statistical interpretation and this will directly inform the statistical attributes of these stochastic volume processes on the LOB. To ensure the accuracy of the statistical and financial conclusions drawn from the analysis, we consider several parameter estimation approaches for each model which include, generalized method of moment based approaches, empirical percentile based approaches, mixed maximum-likelihood and moment based methods and L-moment based estimators. We study the impact of the trade-off between the variance in the parameter estimates, due to reduced sample sizes and the bias introduced at the higher sampling rates. In the discussion of each method, we comment on the suitability for practical estimation of such models using ultra-high frequency LOB data sets.

2.1 Data specifics

For this research, we study volumes of futures, which is a financial derivative instrument. Section 1.2.2 provides an introduction to futures. The futures we study are listed in Table 2.1. Recall, futures contracts are a binding agreement to take, or make delivery of a per-specified quantity and quality of an asset, on a predetermined date and location. The value of the futures contract is derived from the eventual use of the underlying asset. The predetermined date is known as the expiry/delivery date, at which time the buyer takes delivery of the underlying asset, unless having exited the position prior, and the seller has to provide the underlying asset at the expiration date. At any one point in time there are many futures contracts with the same underlying asset, but different expiry dates. The contract closest to the expiry date is the *front* contract and has the highest liquidity. Traders roll over futures contracts close to expiry from the soon-to-expire front contract, to the *back* contract and as a result, liquidity increases in the back contract in the days surrounding the roll over, known as the roll date (different to the expiry date). The roll date is not a scheduled date, it represents the time the process of roll over commences. In this research, we study the front futures contract. The historical data used within this research contains a continuous futures history, or the *on-the-run* front contract, so joining contract data is not necessary.

2.1.1 Liquid market hours

We consider limit orders only during *liquid market hours* (Table 2.1), as defined in Richards et al. (2015), excluding auction periods and excluding days when the exchange is open on a public holiday. Price discovery, which is the purpose of the continuous double-sided auction seen in the marketplace, becomes less efficient when liquidity is low (Chordia et al., 2008). This is driven by reported prices in the marketplace (at the best bid and ask) having insufficient volume during low liquidity times to support meaningful transaction sizes. Low liquidity indicates that few participants are currently active in the

marketplace, hence the rate of information arrival is slow and the reported price becomes inefficient. Inefficiency in the price discovery mechanism leads to asset prices that deviate from their underlying equilibrium levels. Therefore, liquid market hours are times of the trading day when the transaction rate in the marketplace is sufficiently high to guarantee that the price discovery process is operating efficiently, and that the prices reported can be transacted in reasonable size. With sufficient volume per unit of time trades can be executed while still controlling for the cost of execution. As a general rule, the consumption of liquidity (ie, the portion of the market volume flow that is consumed in filling trades) should never exceed more than about 10%-15% of the volume available per unit time. Technically, liquid hours can be disjoint sets or intervals, however we take a contiguous interval where illiquid periods within the specified liquid market hours are included. The use of liquid market hours ultimately focuses the modelling of the volume dynamics on the stochastic volume process attributes, rather than the short periods related to exchange specific pre-market open auction mechanisms.

Table 2.1: Asset description used in the analysis and modeling. Market hours refer to the liquid market hours in local trading time of the exchange.

Asset Name	Acronym	Liquid Market Hours (local time)	Exchange
Interest rate futures			
1. 5 Year T-Note	<i>5YTN</i>	7:30:00 to 14:00:00	CBOT
2. Euro-BOBL	<i>BOBL</i>	8:00:00 to 19:00:00	EUREX
Equity index futures			
3. SIMEX Nikkei 225	<i>NIKKEI</i>	8:00:00 to 14:00:00	SGX
4. E-mini S&P 500	<i>SP500</i>	08:30:00 to 15:00:00	CME
Precious metals futures			
5. Gold	<i>GOLD</i>	06:30:00 to 13:30:00	COMEX
6. Silver	<i>SILVER</i>	08:30:00 to 13:00:00	COMEX

The data used within this study is the Market Depth (Level II) LOB volumes. For a detailed discussion of the components and construction of the LOB from the exchange to the data vendor, we refer the reader to Chapter 3 of this thesis. A distinguishing feature of this research compared with other empirical literature, is the vast quantity of data utilized. To put this into context, for a single random day in 2010, GOLD has 814,580 event times recorded, which are associated with one or more event types, such as a market order, limit order or cancellation, across 10 levels in the LOB data, for the on-the-run front contract. 5YTN has 790,006 event times and SP500 has 3,717,465 event times. This study considers 250 trading days in the year of 2010.

We analyse volumes as defined in Definition 1, for a range of high frequency sampling rates, with sub-sampling frequencies of 10 seconds, 5 seconds, 2 seconds and 1 second intraday for each trading day of 2010, however for brevity, results are only presented for the 10 second sampling rate. We have several objectives for considering these sampling rates, with the first to disambiguate the real stochastic behaviour of the heavy tailed features of the LOB volume process structures from the high-frequency micro-structure noise (Dayri, 2012; Bacry et al., 2013a). The second is to understand, for a given market and asset class, whether basic statistical features such as heavy tailed attributes are persistent in

the stochastic processes and possibly arising at different sampling rates. Finally, small sample rates provide guidance on what statistical properties and appropriate distributional models can be expected for the specification of marks on irregularly spaced time intervals for a Hawkes process.

2.2 Empirical analysis of limit order book volumes

Table 2.2 presents some descriptive statistics of the volumes at level one of the LOB for 2010 for each asset. We observe that the mean volume, the variance and the total daily spread of volumes on level 1 of the LOB, for 5YTN, BOBL, SP500, NIKKEI, GOLD and SILVER can be large. All assets show high levels of positive skewness, with GOLD demonstrating the highest level (mean \pm standard error) 7.61 ± 4.88 on the bid side and 8.74 ± 7.04 on the ask side. The mean level of kurtosis is high for all assets, however the standard deviation of kurtosis may indicate that high kurtosis is not always present in the data. For example, NIKKEI kurtosis is 6.55 ± 3.90 on the bid side and similarly for the ask side, 6.54 ± 3.40 .

Table 2.2: Descriptive statistics for volumes on level 1 of the LOB across all trading days, for 2010, using sub-sample data of 10 seconds.

Asset	Side	Max	Min	Median	Mean	Std	Kurtosis	Skew
5YTN	Bid	2553.05	1.05	390.88	420.19	324.95	8.73	1.30
	Ask	2514.93	1.04	399.45	424.80	322.76	7.58	1.20
BOBL	Bid	2633.98	1.04	430.67	465.35	298.44	14.42	1.59
	Ask	2606.95	1.04	429.43	462.18	293.88	14.42	1.55
SP500	Bid	4040.79	5.01	540.34	635.55	436.54	13.30	2.00
	Ask	4448.10	4.69	541.34	650.44	474.15	16.13	2.29
NIKKEI	Bid	531.62	1.17	103.65	115.43	75.35	6.55	1.33
	Ask	536.11	1.20	105.14	116.20	75.83	6.54	1.32
GOLD	Bid	172.96	1.00	4.62	6.78	8.66	139.65	7.61
	Ask	206.85	1.00	4.62	6.80	9.64	184.71	8.74
SILVER	Bid	79.17	1.00	6.56	7.83	6.39	46.77	3.63
	Ask	76.02	1.00	6.56	7.81	6.30	38.63	3.38

The empirical studies that follow assess the shape of the volume process for the first 5 levels of the bid and ask of the LOB. The empirical literature on the LOB presents a humped shape description of the LOB and this motivates greater consideration of levels beyond the best bid and ask. The assessment will determine whether these features persist in the asset class and high frequency data, which we study. Despite being recognized in empirical literature, dependence structures of high frequency data are often ignored in models of the LOB. We study the long range dependence of each asset using two methods, the Hurst exponent analysis and the Extremogram. Finally, we will present empirical evidence for heavy tails on LOB volumes by using two non-parametric techniques, exponential quantile plot and the mean excess plot, to assess whether the proposed gamma distribution (Bouchaud et al., 2004; Gu et al., 2008) is sufficient to capture the features of volume processes. This empirical study will guide whether alternative distributions need to be assessed, contributing important ground work for the application of the marked

Hawkes process to LOB data in later chapters, where distributional specification for the marks are required.

2.2.1 Shape of limit order book volume profiles

To assess the shape of the LOB volume process, we develop a graphical representation of the volume on the LOB, which we denote as the volume profile for each asset obtained. The volume profile is evaluated by taking the median of the 10 second volumes for each hourly time increment throughout each trading day of the year 2010. In addition, the median volume per level on the LOB, levels 1 to 5, per day across the year of trading is considered. With this information we develop an understanding of the general volume features of the LOB, inclusive of depth considerations.

The originally proposed idea of a LOB shape suggested a monotonically decreasing function away from the best bid and ask (Biais et al., 1995; Bouchaud et al., 2002; Challet and Stinchcombe, 2001). More recent findings from Potters and Bouchaud (2003), Gu et al. (2008) and Chakraborti et al. (2011) suggested a humped shaped LOB. Consistent with the more recent findings, Figures 2.1, 2.2 display a heat map of the intra-day volume processes for the year and the common feature that appears between the 5YTN and BOBL, respectively, is the humped shaped LOB. This is also consistent for NIKKEI and SILVER. SP500 and GOLD appears to have monotonically increasing volumes in the first 5 levels of the LOB. As shown in Figure 2.2, the heat chart for BOBL volumes are significantly higher at the start of the year and drop off towards the end of 2010. This feature is also present in NIKKEI and to a lesser extent, the 5YTN (Figure 2.1) and SILVER. SP500 and GOLD volumes tend to be relatively consistent throughout the year, contrary to the clear change in the volume profile dynamic throughout the year of 2010 for several of the other key futures assets.

The common feature that appears in all assets, is the inherent symmetry in the median volumes on the bid and ask at each level of the LOB. Despite the precious metals GOLD and SILVER having significantly lower volumes placed on the LOB compared with other assets, they still demonstrate the inherent symmetry observed in the other assets.

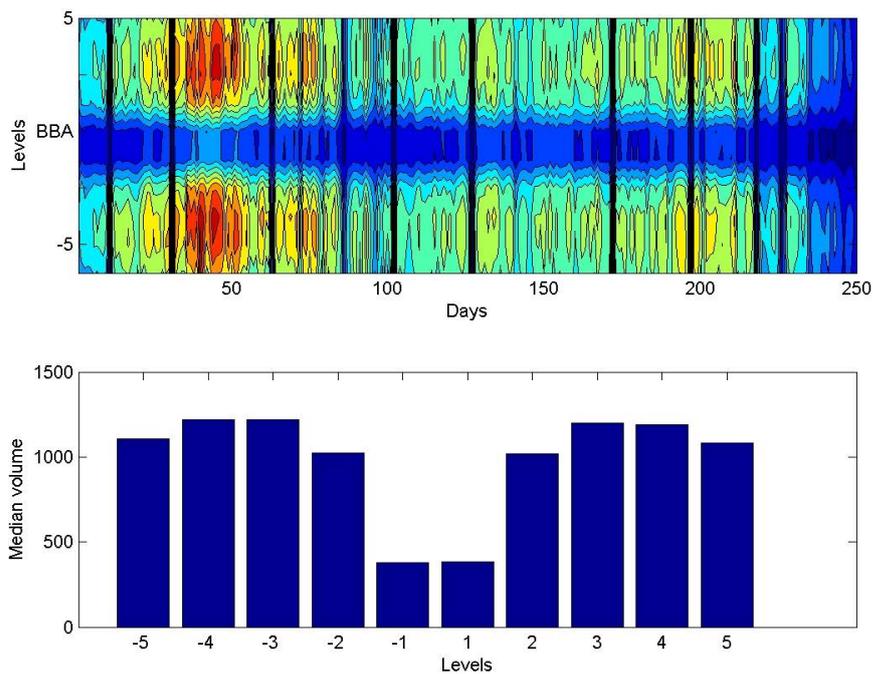


Figure 2.1: 5YTN: Heat maps of the volume for the first five levels of the LOB and the median volume on each level of the LOB on the bid and ask for 2010. The black lines represent the 9 U.S. public holidays that the CME observes.

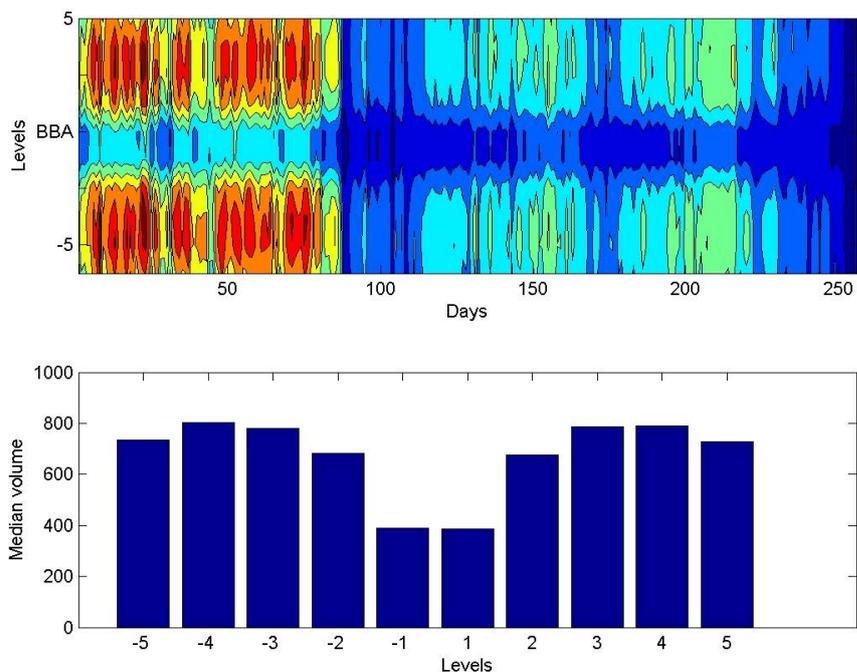


Figure 2.2: BOBL: Heat maps of the volume for the first five levels of the LOB and the median volume on each level of the LOB on the bid and ask for 2010.

2.2.2 Assessing long range dependence

Liquidity is the ease at which assets can be sold without loss of value and as we presented in Section 1.1, this plays an important role in price discovery. The LOB provides liquidity, which is essential for a well functioning trading mechanism. The characterization of liquidity has three dimensions, spread, depth and resiliency (Harris, 2003). Resiliency has been classified as a return to some former price level, or alternatively, as the replenishment of volume in the LOB after some shock (Panayi, 2015). Whilst the aim is not to study the resiliency of the LOB or the effect of shocks to the LOB, general persistence or long range dependence in volumes plays an important role in the LOB being able to provide consistent liquidity for transactions. This persistence does not necessarily correspond to persistence of the mid-price.

Available liquidity on the LOB is an important consideration of the evaluation of price impact models, which are a core component of any optimal execution framework. It is also an important aspect of supply and turnover for trading activities. If long range dependence exists, but the models do not account for this dependence, the predicted future liquidity available on the LOB will be biased and may have a material impact on the liquidity assessment. From a regulatory perspective, market regulators policies are aimed at promoting consistent liquidity to ensure well functioning markets. For recent work on market liquidity and a large-scale study of commonality in liquidity and resilience, we refer the reader to Panayi (2015); Panayi et al. (2015).

Two methods are considered when studying the long range dependence of an asset with the first being the well known Hurst exponent. We then confirm our findings from the Hurst exponent analysis by also studying a more recent technique called the Extremogram (Davis and Mikosch, 2009).

Hurst exponent (long memory) for the limit order book volume process

To study empirically the possibility of long memory in the LOB volume processes at each of the 5 levels of volume on the bid and the ask, we considered the autocorrelation function, which preceded this study, and with results suggesting the presence of long memory for all assets. Gu et al. (2008) utilized the Hurst exponent to test for long memory in the volume of the LOB by implementing a de-trended fluctuation analysis to estimate the Hurst exponent on the 1 minute averaged volumes at the first 3 tick levels of the LOB. De-trended fluctuation analysis is a well-established scaling method for the detection of long-range correlations in time series Kantelhardt et al. (2001); Hu et al. (2001). This provides an *index of long-range dependence*, giving a quantitative measure of the relative tendency of a time series to regress strongly to the mean or cluster. As a guide, values for the index in the range $0.5 < H < 1$ indicate a time series with long-term positive autocorrelation Simonsen and Hansen (1998). This indicates momentum in the intra-day volume process, whereby high volume in the series is likely to be succeeded by another high volume period. Values of the Hurst exponent between $0 < H < 0.5$ indicates a time series with long-term switching between high and low volumes in adjacent 10 second time increments.

With our aim to provide a richer data analysis, we implement the Hurst exponent estimation across a wider range of assets for every trading day of 2010, considering all 5 levels of the bid and ask. Figure 2.3 shows the Hurst exponents for each intra-day volume process at level 1 to level 5 of the bid and ask, in a sequence of box plots comprised of estimates for each trading day of 2010. In the analysis performed for each of the assets, there is a range of results between $(0.65, 1.00)$ across all 5 levels on the bid and ask side volumes for all assets. This is strongly consistent with what would be expected from data exhibiting long memory. Such features in general financial time series have previously been noted by Lobato and Velasco (2000), where they show that trading volume and volatility show the same type of long memory behaviour. We have extended the range of assets and depth in the order book to show this long memory feature in the volume exists in a range of different asset classes in the futures market. This forms a key consideration when modelling the dependence structure within models for the LOB.

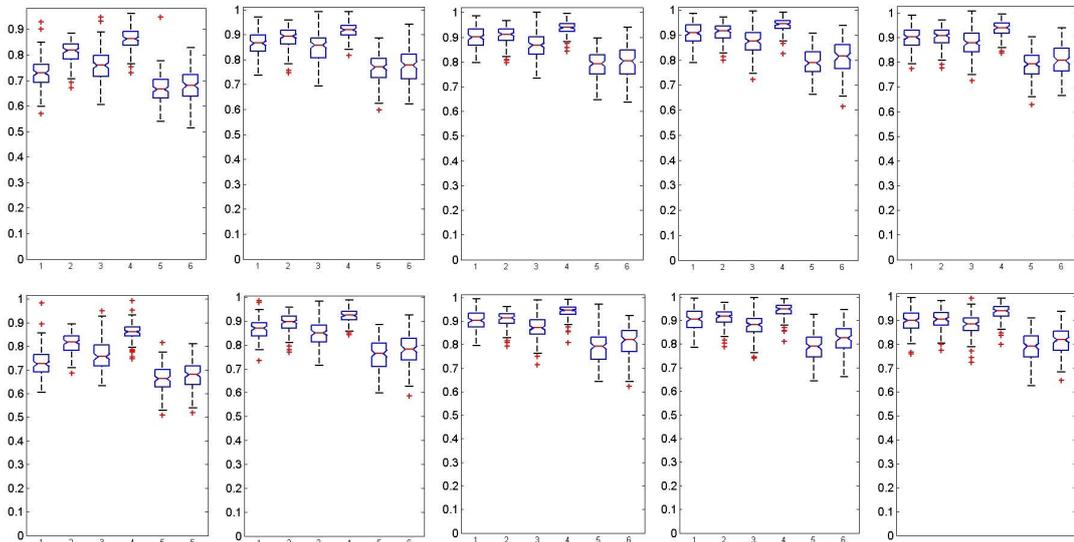


Figure 2.3: Boxplots of daily Hurst exponent for 2010. *Top Row - Left to Right*, level 1 to level 5 bid volumes; *Bottom Row - Left to Right*, level 1 to level 5 ask volumes; *Each Subplot*, assets left to right (1 to 6) are, 5YTN, BOBL, SP500, NIKKEI, GOLD and SILVER.

For each asset and each trading day, we also tested the volumes on each level for stationarity, and in all cases the data was stationary at level 1 of the LOB for time increments of 10 seconds. We also considered deeper levels of the LOB. For level 2, for assets SP500 and NIKKEI, we observed a stochastic trend, which was removed by first differencing in approximately 1% of trading days. Moving to level 3 of the LOB, the following assets exhibited non-stationarity, with the indicated portion of days with this feature reported in brackets, 5YTN (bid 5.70%, ask 6.10%), BOBL (bid 0.39%, ask 0.00%), SP500 (bid 4.30%, ask 2.80%) and NIKKEI (bid 6.90%, ask 6.50%). Level 4 demonstrates increasing days with a stochastic trend, 5YTN (bid 15.90%, ask 14.70%), BOBL (bid 0.39%, ask 0.00%), SP500 (bid 5.12%, ask 5.98%) and NIKKEI (bid 8.57%, ask 8.88%).

Level 5 of the LOB shows, 5YTN (bid 10.49%, ask 15.51%), SP500 (bid 7.17%, ask 2.08%) and NIKKEI (bid 10.61%, ask 8.16%). As we investigate deeper levels of the LOB, a stochastic trend is present and the long memory appears to increase. However, the results for which a subset of the series required differencing, due to stochastic trend removal, did not vary the long memory findings significantly from the original Hurst exponent results presented in Figure 2.3.

To further validate the results of this test, a randomized experiment for the Hurst exponent was implemented. For all assets across the 250 trading days considered, we found the Hurst exponent estimates to be robust to heavy tailed data.

Finally, we consider the impact of varying the sub-sampled time increments for level 1 of the LOB. We consider a range of time intervals of one second to one minute. Table 2.3 shows that in all cases we can see an increase in long memory as the time increment becomes finer. NIKKEI demonstrates the highest degree of long memory for the one second time increment, with a Hurst Exponent of 0.9020. What is clear from this study is that for any realistic trading time interval, long range dependence is a persistent feature of volume processes in the LOB and cannot be removed by considering lower frequency data.

Table 2.3: Mean Hurst Exponent across all trading days, using varying sub-sampled data of volumes on *level 1* of the LOB.

Asset	Side	1 second	2 seconds	5 seconds	10 seconds	1 minute
5YTN	Bid	0.7948	0.7765	0.7553	0.7265	0.6286
	Ask	0.7918	0.7728	0.7559	0.7276	0.6234
BOBL	Bid	0.8585	0.8460	0.8271	0.8088	0.7463
	Ask	0.8606	0.8483	0.8286	0.8105	0.7447
SP500	Bid	0.8206	0.8038	0.7811	0.7545	0.6524
	Ask	0.8236	0.8070	0.7857	0.7587	0.6489
NIKKEI	Bid	0.9020	0.8922	0.8764	0.8605	0.8046
	Ask	0.9004	0.8918	0.8753	0.8589	0.8042
GOLD	Bid	0.7483	0.7295	0.6968	0.6666	0.5833
	Ask	0.7452	0.7252	0.6915	0.6596	0.5730
SILVER	Bid	0.7651	0.7448	0.7070	0.6763	0.5802
	Ask	0.7633	0.7430	0.7048	0.6763	0.5830

Extremogram (serial dependence in the upper tail)

In this section we apply a more recent technique called the extremogram, which was developed by Davis and Mikosch (2009) and provides a quantitative measure of dependence of extreme events in a stationary time series. Stationarity of the series was assessed in the previous section, with all LOB volumes exhibiting stationarity on level 1 of the LOB. To simplify notation for the discussion in this section, we set $V_t = V_{t,d}^{l,a}$ in Definition 1, where $d \in \{1, \dots, 250\}$ is each trading day that we summarize across, $l = 1$ is level 1 on the bid and ask, a is the stated asset in each example, and t is the stated sub-sampling frequency in each examples.

For a strictly stationary \mathbb{R}^d valued time series (V_t) , the extremogram is defined by

$$\rho_{A,B}(h) = \lim_{v \rightarrow \infty} \mathbb{P}(v^{-1}V_h \in B | v^{-1}V_0 \in A), \quad h = 0, 1, 2, \dots \quad (2.1)$$

provided the limit exists. Because the volumes are positive, the *extremogram* has been applied in this study by choosing $A = B = [1, \infty)$. This reduces the extremogram to the upper tail dependence, which is often used in extreme value theory and quantitative risk management (McNeil et al., 2005; Davis et al., 2012).

To estimate the extremogram, the limit on v in (2.1) is replaced by a high quantile $(1 - 1/a_m)$ of the process (Davis et al., 2012). We select a_m as the 20th percentile in order to be consistent with the peaks over threshold approach, used when fitting the generalized Pareto distribution in later sections of the analysis. The sample extremogram, which is based on the observations V_1, \dots, V_n , is given by

$$\hat{\rho}_{A,B}(h) = \frac{\sum_{t=1}^{n-h} I(V_{t+h} \geq a_m, V_t \geq a_m)}{\sum_{t=1}^n I(V_t \geq a_m)}. \quad (2.2)$$

The extremogram, being the conditional measure of extremal serial dependence, is suitable for studying the persistence of a shock in the volume at a future time instant. The persistence of increased liquidity, for example, will allow a market participant to increase their position in order to take advantage of the extra liquidity, without incurring additional transaction costs associated with liquidity. BOBL (Figure 2.4) and SP500 show persistent extremograms across all lags for the 250 trading days. SP500 and 5YTN demonstrate lower levels of serial dependence compared with SP500. GOLD and SILVER show almost no serial dependence in the upper tails.

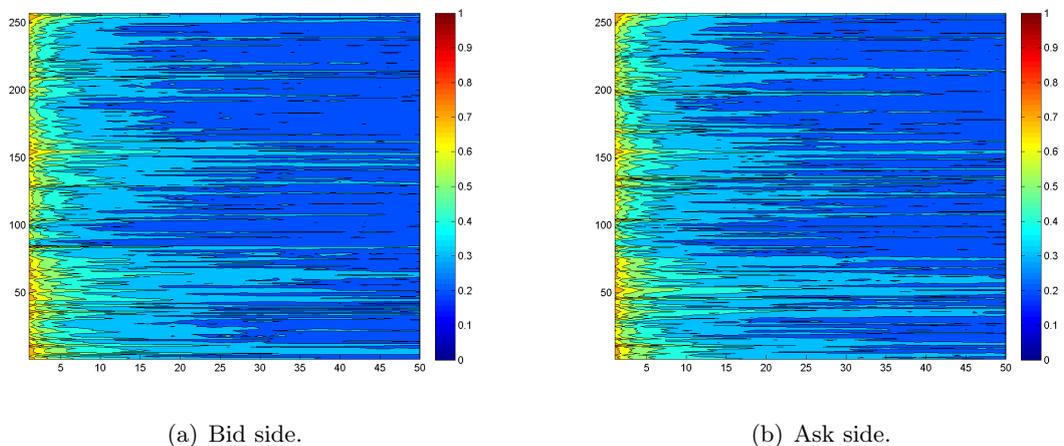


Figure 2.4: BOBL: extremogram heat map for 250 trading days, using *10 second* sub-sampled data, $a_m = 80$ th percentile and geometric distribution p-value= $1/200$.

The bootstrapped sample extremogram

We consider the bootstrapped sample extremogram to measure the significance of the extremogram estimates. The method implemented is detailed by Davis et al. (2012), who recommended the block re-sampling scheme, which was introduced by Politis and Romano (1994).

This method defines blocks by first choosing a starting point at random from a range $(1, n)$. The length of the block is chosen geometrically with a probability of $1/200$, as recommended by Davis et al. (2012). The second and subsequent blocks are generated until the total length of the concatenated blocks is equal to, or greater than our original sample size. The 95% confidence intervals are produced by using 10,000 stationary bootstrapped replicates and evaluating the indicator functions defined in (2.2). The bounds are found using the 0.025 and 0.975 quantiles from the empirical distribution of the bootstrapped replicates. This provides consistent estimators of the variability of the extremogram.

When generating the confidence intervals for each of the assets across the 250 trading days, we consider a random selection of days for the bid and ask side. The solid horizontal line of height 0.20, represents the extremogram under an independence assumption. It is worth noting that Davis et al. (2012) use 0.04, however they had a very long time series of tens of thousands of observations, markedly larger than the few thousand data points of the time series used in this study. In addition, the 0.20 remains consistent with the peaks over threshold approach used when fitting the generalized Pareto distribution in later sections. If the solid horizontal line is well outside these confidence bands, this will confirm the serial extremal dependence in the upper tail.

As demonstrated in Figure 2.5, BOBL has serial dependence in the upper tail up to the 25th lag for both bid and ask side at the 2.5% level of significance. For the 5YTN and SP500, serial dependence in the upper tail up to the 9th lag is observed. NIKKEI confirms serial dependence up to the 18th lag. Consistency of serial dependence across both bid and ask side exists for all assets. GOLD and SILVER shows serial upper tail dependence up to lag 3 only. It is worth highlighting, that in all cases, the median of the simulation always exceeds the independent, and smoothly decays over time. This provides a strong indication of persistent long range dependence. The results for the Extremogram for each asset are largely consistent with what we have observed in the Hurst exponent, that is, GOLD and SILVER for the Hurst and Extremogram show the lowest levels of long memory and serial dependence in the upper tails, respectively. And conversely NIKKEI and BOBL demonstrated the highest levels of long memory and serial dependence in the upper tails.

Long range dependence has been noted by Gu et al. (2008). By implementing two procedures, the Hurst exponent and the more recent technique being the extremogram, confirms the finding of long range dependence across all assets and with increasing dependence for shorter time increments.

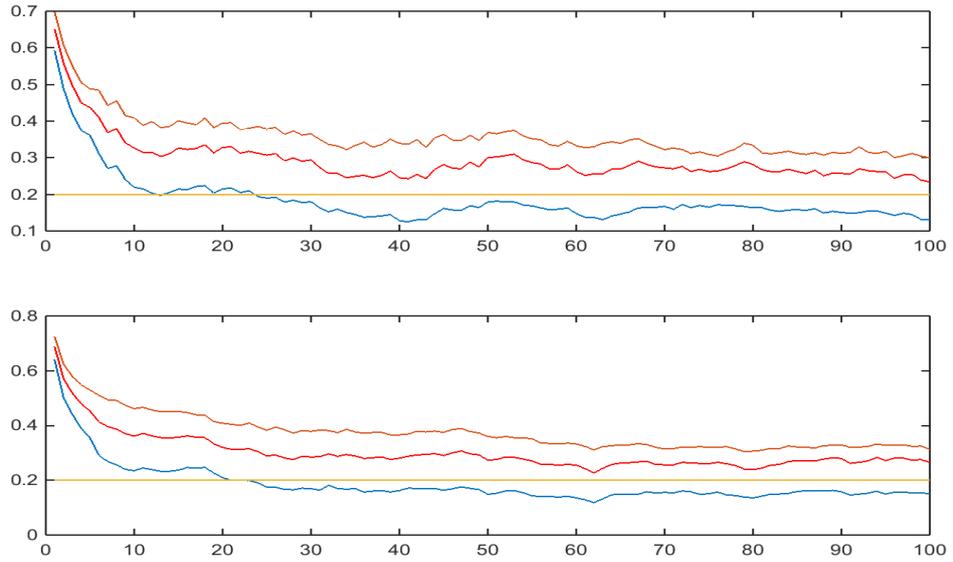


Figure 2.5: BOBL: extremogram for volume level 1 of the LOB, using 10 second sub-sampled data, 10,000 bootstrapped samples and for 100 lags. The upper chart shows the bid-side and the lower chart shows the ask-side. The solid horizontal line of height 0.20, represents the extremogram under independence. The bounds are found using the 0.025 and 0.975 quantiles from the empirical distribution of the bootstrapped replicates.

Extreme value theory and dependence

This section details the context in which one may apply elements of extreme value theory, typically applied to i.i.d. data sets, to a time series structure, such as the volume process on the bid and ask at each level of the LOB.

Volume spikes followed by additional volume spikes, being the presence of persistent heavy tailedness, is of particular importance to all market participants. Whether it be the agency broker seeking best execution, the market-makers providing liquidity, or the traders and arbitrageurs systematically increasing their positions to take advantage of the increased liquidity, they can all benefit from a better understanding of the heavy tailed features of the volumes. When studying the heavy tailed features of the volume process, we adopt techniques from econometrics, statistics and probability, which includes working with appropriate aspects of extreme value theory Beirlant et al. (2004). Even with the time series structure observed in the volume process, it is appropriate and meaningful to consider extreme value theory. Considering both the marginal and the joint distributions of the process, as highlighted in the statistical finance paper of Cont (2007), is crucial in developing a holistic understanding of the volume process.

For a generic volume process (V_t) observed over time, with the assumption that it is a strictly stationary time series, each individual observation of the volume process V_t will have the same distribution function, denoted by F . If that data didn't have any time series structure we would be safe to assume that we could study, under the extreme value theory framework, the maximum of the series of i.i.d. random variables of length

n , denoted by $M_n = \max\{Z_1, \dots, Z_n\}$. One could then trivially show in this case that $\mathbb{P}(M_n \leq y) = \{\mathbb{P}(V_j \leq y)\}^n = F^n(y)$, where the independence of V_t is used.

For dependent data, such as data obtained from a time series structure, this relationship does not hold exactly and the distribution of the maximum M_n is not determined solely by the marginal F alone, but rather from the complete distribution of the time series, i.e also the transition or conditional probabilities. However, it is also known in the extreme value theory literature that often a comparable, approximate extreme value theory relationship will be applicable, allowing for the application of extreme value theory in this context to time series data. In many settings, the following approximate relationship will be applicable to aid in this context, $\mathbb{P}(M_n \leq y) \approx F^{n\eta}(y) \geq F^n(y)$, which will apply for large time series samples. This will be the case when considering ultra-high frequency LOB data sets intra-daily, for example, large n with $\eta \in [0, 1]$ denoting what will be termed the extremal index or extreme value index. The extremal index is a critical parameter in extreme value theory related to the heavy tailed nature of the data, see discussions in Embrechts et al. (1997).

More precisely, in the independent case one can say that for $\tau \in [0, \infty]$ and every sequence of real numbers $(u_n)_{n \geq 1}$, then it holds that as $n \rightarrow \infty$ then $n\bar{F}(u_n) \rightarrow \tau$ iff $\mathbb{P}(M_n \leq u_n) \rightarrow e^{-\tau}$. From this statement one can say that the distribution F of the volume process belongs to the domain of attraction of a generalized extreme value distribution.

In the context of a time series, which does not display independence, one has approximately the following extension of this result. We consider the extreme value index $\eta \in [0, 1]$, for the volume process time series (V_n) when, for certain τ and a sequence of real numbers u_n , one can show that $n\bar{F}(u_n) \rightarrow \tau$ and $\mathbb{P}(M_n \leq u_n) \rightarrow e^{-\eta\tau}$. Furthermore, if an η exists, then the value does not depend on the specific choice of τ, u_n . Therefore, using this approximate relationship between the distribution of the maximum and the exceedence probability, one obtains directly, as u_n grows large and $\bar{F}(u_n) \approx 0$, the following approximate relationship, $\mathbb{P}(M_n \leq u_n) \approx e^{-\eta\tau} \approx e^{-\eta n\bar{F}(u_n)} \approx (1 - \bar{F}(u_n))^{n\delta} = F^{n\delta}(u_n)$. In this context, one may adopt aspects of extreme value theory and apply it meaningfully to time series data.

The data in this study demonstrates dependence, but this does not remove the necessity for one to consider the heavy tailed features present in the data, nor diminish the importance of a study on heavy tailed marginal distributions. The study of heavy tails in a marginal context has important effects on the accuracy of the estimators (Francq and Zakoïan, 2013). Studies that consider heavy tailed features in financial market data include and are not limited to, Cont (2001) and Francq and Zakoïan (2013). These features are a decisive contribution to a LOB model such as a Hawkes process, which requires a marginal distribution specification for the marks. In this study, we will use methods such as MLE, based on the assumption that observations are independent, hence the methods will not be efficient. However, we have large sample sizes on which to fit the extreme value distributions and it will be sufficient to obtain consistent estimators of the marginal distribution and the heavy tailed features.

2.2.3 Empirical evidence for heavy tails in limit order book volumes

The literature on LOB modelling to date has considered only simple classes of two parameter shape-scale models for the statistical modelling of LOB data for price or volume. Papers by Bouchaud et al. (2002) and Gu et al. (2008) consider the distributional features of the LOB volumes and conclude that a two parameter shape-scale model, given by a Gamma distribution is most suitable. However, from Table 2.2 we observed high levels of positive skewness and kurtosis for all assets. These findings are indicative of heavy tails and since the volumes are strictly positive, we would expect heavy right tails.

In this section, we demonstrate the need for more sophisticated, flexible parametric models. We began by fitting the gamma distribution, as suggested in previous literature (Bouchaud et al., 2002; Gu et al., 2008). To obtain the estimators of the gamma distribution, we equated the population moments with the sample moments (moment matching). To assess the stability of the parameters and to assess how well the gamma distribution represents the skewness and kurtosis in the data, via moment matching, we estimated the parameters for all assets and for every time segment across an entire year of trading. From these parameters we estimated the mean, variance (which should be consistent with the sample estimates), skewness and kurtosis. For all assets, the gamma distribution provided a poor estimate of skewness and kurtosis.

The observations made are likely to be due to the right tail of the volume distribution, being heavier than can be obtained from the gamma distribution. We further investigate this particular aspect of the distribution by using two non-parametric techniques, with the first being the exponential quantile plot. This acts as a visual comparison between a *medium sized tail* and allows us to identify a relatively *fat-tailed* distribution. If the empirical CDF lie on the *dashed* straight line, then the volume process is consistent intra-daily with an exponential distribution. However, the presence of a concave relationship, whereby the plot bends upwards away from a linear fit, indicates a fat-tailed distribution in the sub-exponential class.

Figure 2.6 shows the QQ-plot for the LOB data for the 5YTN, relative to a generalized Pareto distribution with a tail index of $\zeta = 0$, making this a comparison between the right tail of the empirical CDF and the right tailed exponential distribution. The results presented are estimated intra-daily for every 25th trading day of the year and for each of the 5 levels of the LOB on the bid and ask sides. The choice of the 25th trading day provides an illustration of the general results we observed consistently for each trading day of the year, without overwhelming the visual representation and excluding days, such as a futures expiry.

The findings from the QQ-plots, comparing the empirical CDF to the exponential model, demonstrate several interesting features. Firstly, starting with the 5YTN, at all levels of the volume process there is a convex relationship, relative to the exponential quantiles, indicating light tails for these profiles, see Figure 2.6. However, on some days there is a clear evidence for a concave relationship between the empirical CDF and the quantiles of the exponential distribution, indicating the existence of a power law relationship. To help distinguish these days, we have emphasized examples of these particular trading days

with a thicker solid line on the QQ plots. It is also clear from this analysis that there is a stronger tendency for power law relationships for the right tail of the volume process on the ask side, relative to the bid side for the 5YTN. BOBL also has occasional trading days indicative of the presence of heavy right tails for the intra-day volume process, indicating a power law relationship. In this case, it is clear that there is a stronger tendency for such heavy tailed features to occur on all 5 levels of the bid side throughout 2010, as opposed to the ask side, which indicates far fewer examples of power law tails. NIKKEI demonstrates occasional concave relationships for the right tail of the volume process, for example on level 1 of the ask and level 3 of the bid and ask.

For the SP500 it is apparent that the power law relationship in the tails is prominent more often in the volume process at all levels 1 to 5 on both the bid and ask sides. GOLD (Figure 2.7) indicates strongly the presence of heavy tailed relationships on every trading day analysed on all levels of the LOB for the bid and ask. SILVER, similarly has consistent evidence of heavy tailed relationships in the right tail of the intra-day volume process.

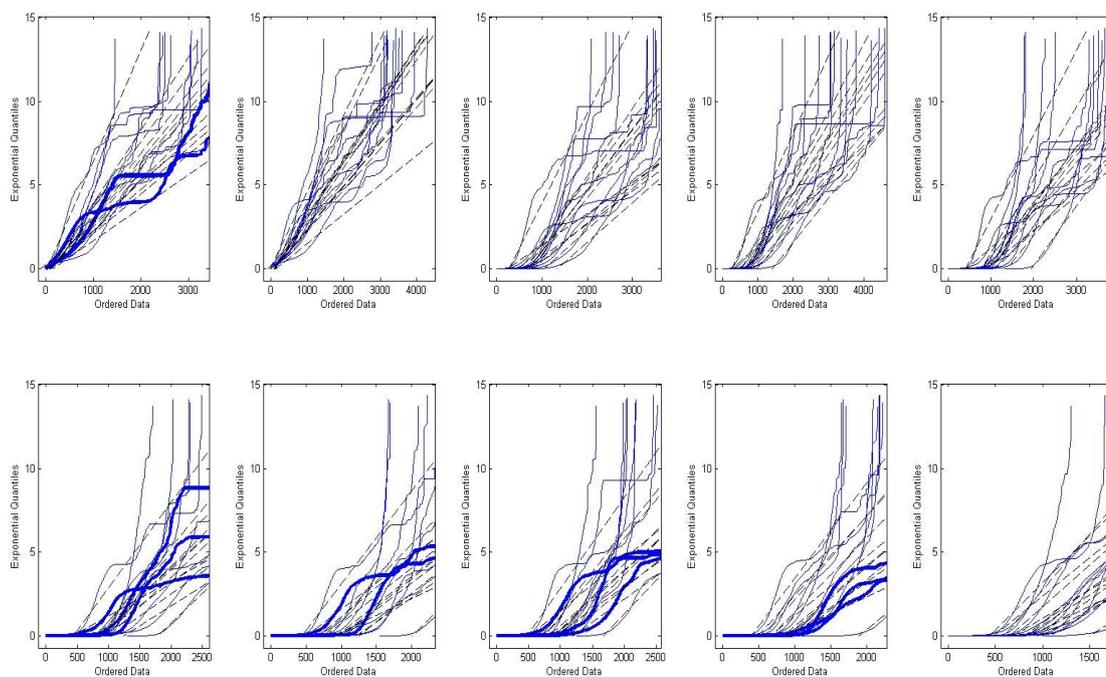


Figure 2.6: 5YTN: Quantiles for exponential distribution model versus sample order statistics, for intra-daily volume data, every 25th trading day of 2010. *Top Row Bid* from left to right is level 1 to level 5 of LOB, and *Bottom Row Ask* from left to right is level 1 to level 5 of LOB.

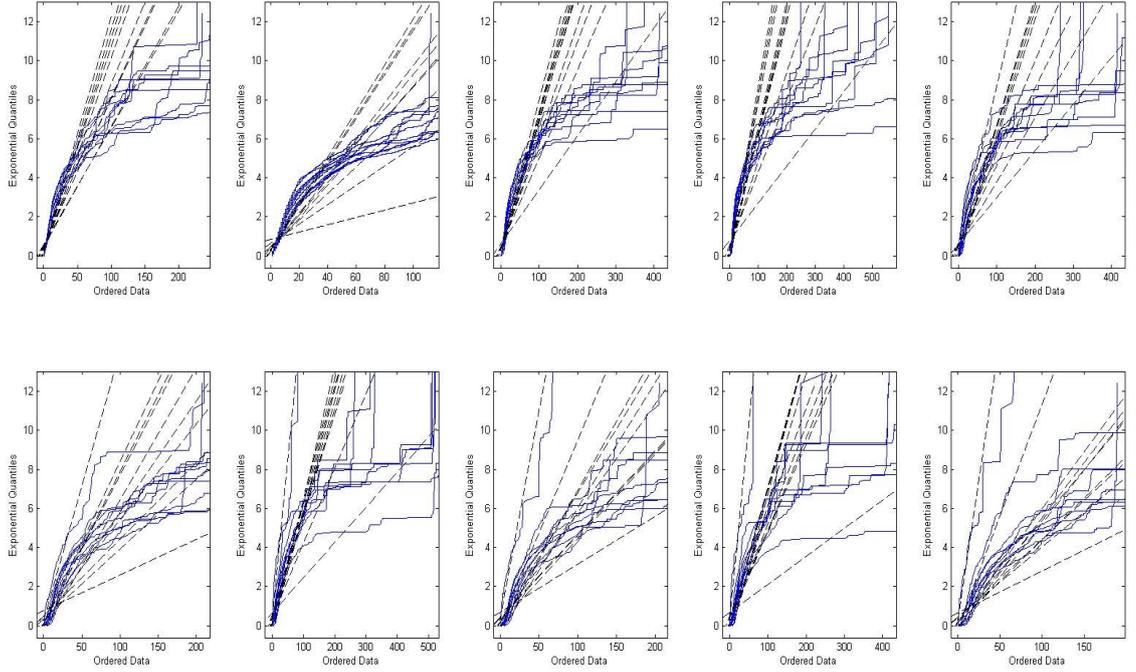


Figure 2.7: GOLD: Quantiles for exponential distribution model versus sample order statistics, for intra-daily volume data, every 25th trading day of 2010. *Top Row Bid* from left to right is level 1 to level 5 of LOB, and *Bottom Row Ask* from left to right is level 1 to level 5 of LOB.

The second technique we consider, is the mean excess plot to aid the assessment of heavy tailed behaviour (Kratz and Resnick, 1996). We present the mean excess plot intra-daily for every 25th trading day of the year, and for each of the 5 levels of the LOB on the bid and ask sides at ten second sub-sampling frequency for each asset. The sample mean excess function defined by (2.3), represents the sum of the excesses over a threshold u , divided by the number of data points which exceed the threshold u . It approximates the mean excess function describing the expected exceedence amount for a particular threshold u , given an exceedence in the volume process has occurred. If the empirical mean exceedence function estimate has a positive slope for large thresholds u , then this indicates that the observed volume process data is consistent with a generalized Pareto distribution with a positive tail index parameter (Beirlant et al., 2004, Chapter 1). Worth noting is that the mean excess function is sensitive to the larger thresholds when the corresponding $e_n(u)$ defined in (2.3), contains fewer observations. Embrechts et al. (1997) suggests in that case to ignore the larger thresholds. However, this is not a necessary consideration for the large data sets of the LOB.

The sample mean excess is given by

$$e_n(u) = \frac{\sum_{i=1}^n (V_i - u) \mathbb{I}_{\{V_i > u\}}}{\sum_{i=1}^n \mathbb{I}_{\{V_i > u\}}}, \quad (2.3)$$

which estimates the conditional expectation $e(u) = \mathbb{E}[(V - u) | V > u]$.

Figure 2.8 displays the mean excess plot versus the threshold u , for the 5YTN. It

indicates a clear upward trend as the threshold (x-axis) exceeds 500 for all of the trading days explored on the bid at level one, consistently indicating the presence of heavy tailed power law relationships in the volume process. At level 1 of the ask, there is a mix between evidence for some days having heavy tailed attributes in the right tail of the volume process, and other trading days with lighter tails in the higher threshold region. This is also present throughout the other levels of the LOB on the bid and ask. The results for BOBL and SP500 demonstrate a strong indication of power law relationships in the right tail of the volume process on several of the trading days. SP500 is very pronounced at level 1 of both the bid and ask. NIKKEI also indicates the occasional presence of power law right tail. As expected from the QQ plots, GOLD (Figure 2.9) and SILVER indicate strong power law relationships consistently in the intra-day volume processes on the majority of trading days presented.

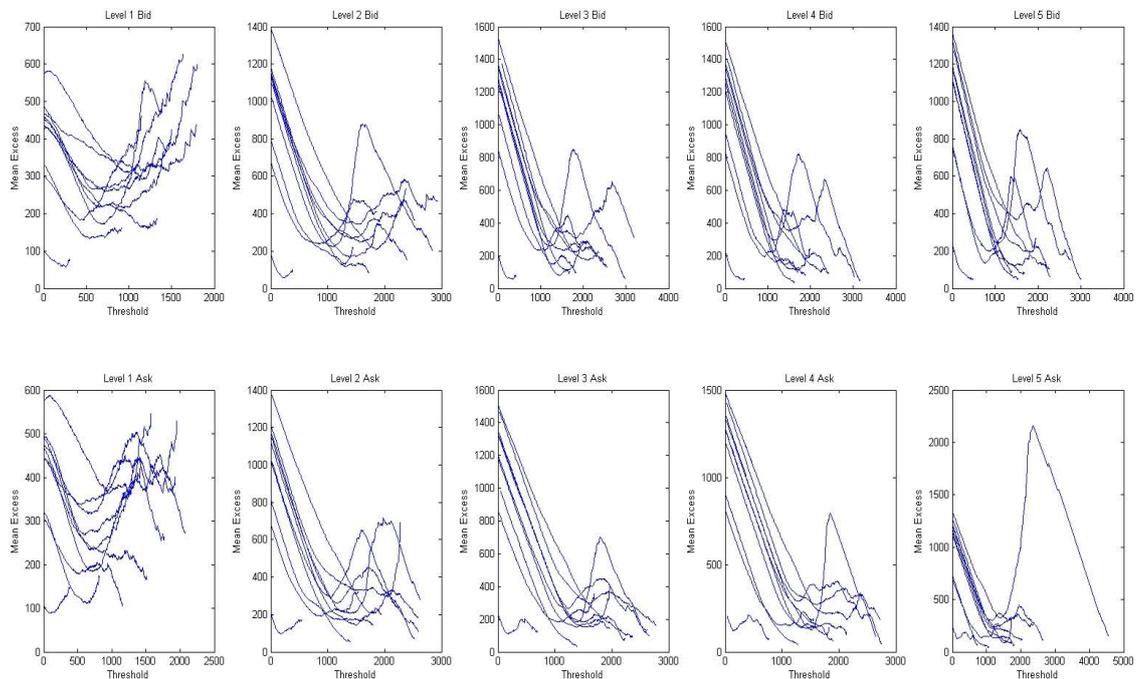


Figure 2.8: 5YTN: Mean Excess plot versus the threshold u , for intra-daily volume data, every 25th trading day of 2010. *Top Row Bid* from left to right is level 1 to level 5 of LOB, and *Bottom Row Ask* from left to right is level 1 to level 5 of LOB.

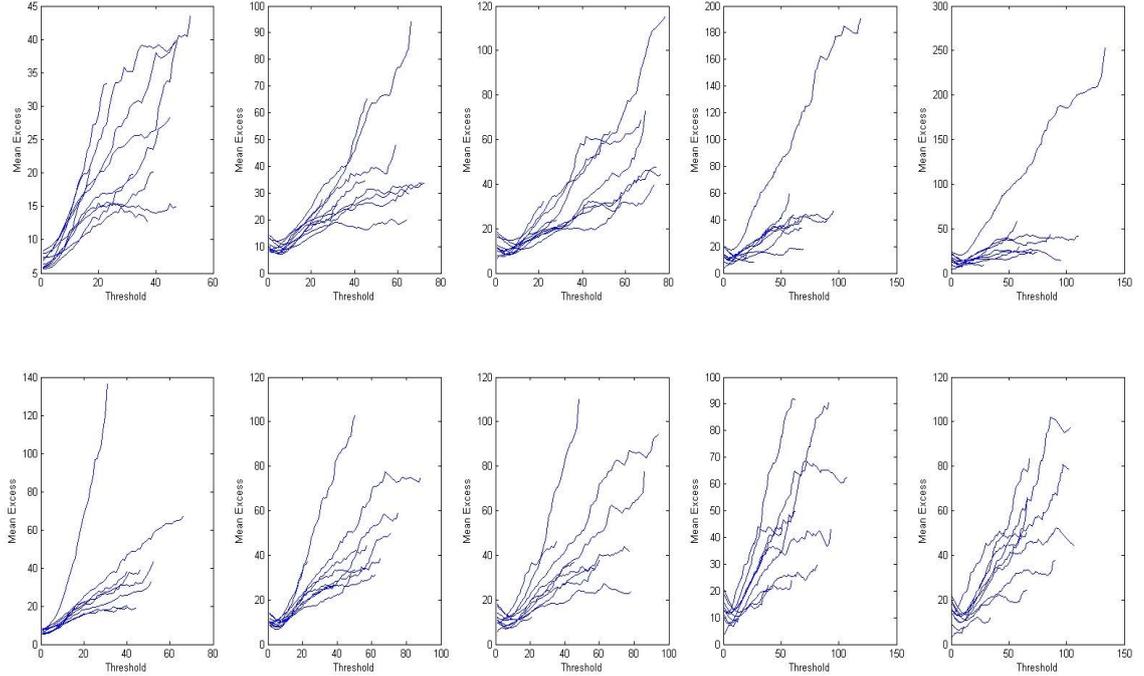


Figure 2.9: GOLD: Mean Excess plot versus the threshold u , for intra-daily volume data, every 25th trading day of 2010. *Top Row Bid* from left to right is level 1 to level 5 of LOB, and *Bottom Row Ask* from left to right is level 1 to level 5 of LOB.

In conclusion, we have convincingly demonstrated that the previously suggested gamma distribution is not capable of capturing the heavy tailed nature of the distribution of volumes in the LOB at all levels. Tails are heavier than exponential and more consistent with the power law. In the next section, we assess alternative distributions that are better able to capture this heavy tailed behaviour.

2.3 Statistical estimation methods for models of the limit order book volume process

The following section presents the parametric families of models that we consider for modelling levels 1 to 5 of the LOB volume processes. Stable distributions have been proposed as a description for the large data sets of the LOB volume data, due to their flexible heavy tail behaviour and asymmetry relationships. We also consider a second sub group of the sub-exponential family of models, being the extreme value distributions, which have a well established statistical theory and give a greater flexibility in capturing heavy tailed features. Both models have asymptotic power law tails.

There are numerous approaches that can be applied when fitting heavy tailed and flexible families of models such as the generalized Pareto distribution, generalized extreme value and α -stable. Each approach will have different merits related to statistical efficiency, bias and variance trade-offs, and importantly for the setting of analysis of LOB data (massive data sets), the methods being computationally robust and efficient. Computational robustness refers to computer science based robustness, whereby the algorithm

that implements the statistical techniques will need to continue to operate despite abnormalities in data inputs. Computationally efficient refers to the amount of resources used by the methods and the time it takes to obtain results.

In the cases where we discuss the Maximum Likelihood Estimation (MLE), we base this on the assumption of independence and as noted in the Section 2.2.2, MLE won't be statistically efficient, but it will be consistent. Efficiency considerations are secondary to our main objective of determining consistent estimators of distribution parameters.

In Table 2.4 we provide a summary of the methods utilized for each of the distributions. This table shows the distributions and methods used for volume data on level 1 of the LOB for each of the six asset, 5YTN, BOBL, SP500, NIKKEI, GOLD and SILVER. For each asset we consider varying time resolutions of 1 second, 2 seconds, 5 seconds and 10 seconds across each trading day in 2010. The discussion that follows, will provide insight into the dynamics of the parameters over time, varying time resolutions, different estimation methodologies and associated implementation issues. In particular, we provide details of several less well known, but efficient and statistically robust approaches.

Table 2.4: Distributions and methods fit to volume data for six assets, 5YTN, BOBL, SP500, NIKKEI, GOLD and SILVER. The methods include, MLE, method of moments, McCullochs quantile based estimation (McCulloch, 1986), Pickands' estimator (Pickands III, 1975) and empirical percentile method.

Method	α -stable	General. extreme value	Generalized Pareto distribution
MLE		✓	✓
McCullochs	✓		
Mixed L-moments			
Pickands		✓	✓
Emp. percent. method			✓

2.3.1 α -stable distribution parameter estimation

Models constructed with α -stable distributions possess several useful properties, including infinite variance, skewness and heavy tails (Zolotarev, 1986; Alder et al., 1998; Samorodnitsky and Taqqu, 1994; Nolan, 2007). Considered as generalizations of the Gaussian distribution, they are defined as the class of location-scale distributions, which are closed under convolutions. α -stable distributions have found application in many areas of statistics, finance and signal processing engineering, as models for impulsive and heavy tailed noise processes (Mandelbrot, 1960; Fama, 1965; Fama and Roll, 1968; Nikias and Shao, 1995; Godsill, 2000; Melchiori, 2006; Peters et al., 2010, 2011, 2012; Gu et al., 2012). We consider using this class of models for the modelling of the volume processes in a LOB stochastic process.

The univariate α -stable distribution is typically specified by four parameters (Lèvy, 1924). The tail index $\alpha \in (0, 2]$, determines the rate of tail decay, $\beta \in [-1, 1]$ determines the degree and sign of asymmetry (skewness), $\delta > 0$ is the scale (under some parametrizations) and $\mu \in \mathbb{R}$ is the location. The parameter α is typically termed the characteristic exponent, with small and large α implying heavy and light tails respectively, and $\alpha < 2$ having infinite variance. Gaussian ($\alpha = 2, \beta = 0$) and Cauchy ($\alpha = 1, \beta = 0$) distributions

provide the only simple analytically tractable sub members of this family. In general, as α -stable models admit no closed form expression for the density, which can be evaluated point-wise, inference typically proceeds via the characteristic function.

For modelling the LOB data, we use the following parametrization in which a random variable V , is said to have a stable distribution if its characteristic function has the following form

$$\mathbb{E}[\exp(i\theta V)] = \begin{cases} \exp\{-\delta^\alpha |\theta|^\alpha (1 + i\beta(\text{sign}(\theta)) \tan(\frac{\pi\alpha}{2})(|\delta\theta|^{1-\alpha} - 1)) + i\mu\theta\} & , \text{ if } \alpha \neq 1, \\ \exp\{-\delta|\theta|(1 + i\beta(\frac{2}{\pi})(\text{sign}(\theta))\ln(\delta|\theta|)) + i\mu\theta\} & , \text{ if } \alpha = 1. \end{cases} \quad (2.4)$$

The series expansions for the corresponding density and CDF are provided in Zolotarev (1983).

The method used to estimate the parameters for the α -stable is based on the quantile approach of McCulloch (1986) and McCulloch (1997). This approach was selected because it is known to be computationally robust and efficient. Estimates of the model parameters are based on sample quantiles, while correcting for estimator bias due to the evaluation of the sample quantiles. We let $v_{(p)}$ be the p -th population quantile and $\widehat{v}_{(p)}$ to be the sample quantile from the order statistics of the sample. The four parameters in the α -stable model, under the parametrization presented, are determined from a set of five predetermined quantiles for the parameter ranges $\alpha \in [0, 2.0]$, $\beta \in [-1, 1]$, $\delta \in [0, \infty)$ and $\mu \in \mathbb{R}$ (McCulloch, 1986). The stages of this estimation involve the following details.

Step 1 Obtain a finite sample consistent estimator of quantiles, with the v_i arranged in ascending order. The skewness correction is made by matching the sample order statistics with $\widehat{v}_{s(i)}$, where $s(i) = \frac{2i-1}{2n}$. Then a linear interpolation to p , from the two adjacent $s(i)$ values, is used to establish $\widehat{v}_{(p)}$ as a consistent estimator of the true quantiles. This corrects for spurious skewness present in finite samples and $\widehat{v}_{(p)}$ is a consistent estimator of $v_{(p)}$.

Step 2 Obtain estimates of the tail index α and skewness parameter β . McCulloch (1986) provide non-linear functions of ν_α of α and ν_β of β in terms of the quantiles, as detailed in (2.5).

$$\nu_\alpha = \frac{v_{(0.95)} - v_{(0.05)}}{v_{(0.75)} - v_{(0.25)}}, \quad \nu_\beta = \frac{v_{(0.95)} + v_{(0.05)} - 2v_{(0.5)}}{v_{(0.95)} - v_{(0.05)}}. \quad (2.5)$$

We can therefore estimate quantities $\widehat{\nu}_\alpha$ and $\widehat{\nu}_\beta$ using the sample estimates of the quantiles $\widehat{v}_{(p)}$. To obtain the actual parameter estimates $\widehat{\alpha}$ and $\widehat{\beta}$, we numerically invert the non-linear functions ν_α and ν_β . This can be done efficiently through a look up table provided for numerous combinations of α and β , and provided in tabulated form in McCulloch (1986).

Step 3 Obtain estimates of δ , given estimates of α and β using quantile matching. McCulloch (1986) provide a third non-linear function, which is explicit in δ ,

and implicit in α and β , in terms of the quantiles, as detailed in (2.6).

$$\nu_\delta(\alpha, \beta) = \frac{v_{(0.75)} - v_{(0.25)}}{\delta}. \quad (2.6)$$

An estimate then follows given $\hat{\alpha}$, $\hat{\beta}$ and consistent sample quantiles $\hat{v}_{(0.75)}$, $\hat{v}_{(0.25)}$.

2.3.2 Generalized extreme value parameter estimation

We present the characterization of the extreme value theory families considered for modelling of the right tail of the volume distribution at each level of the LOB profile. We focus on the generalized extreme value and the generalized Pareto distribution families of distributions, which will be presented next.

Definition 2 (Generalized extreme value distribution). *The generalized extreme value distribution is defined by*

$$\mathbb{P}(V < v; \mu, \delta, \zeta) = \exp \left\{ - \left[1 + \zeta \left(\frac{v - \mu}{\delta} \right) \right]^{-1/\zeta} \right\}, \quad (2.7)$$

for $1 + \zeta(v - \mu)/\delta > 0$, where $\mu \in \mathbb{R}$ is the location parameter, $\delta > 0$ the scale parameter and $\zeta \in \mathbb{R}$ the shape parameter. Furthermore, the density function is given by

$$f(v; \mu, \delta, \zeta) = \frac{1}{\delta} \left[1 + \zeta \left(\frac{v - \mu}{\delta} \right) \right]^{(-1/\zeta)-1} \exp \left\{ - \left[1 + \zeta \left(\frac{v - \mu}{\delta} \right) \right]^{-1/\zeta} \right\}. \quad (2.8)$$

In addition, the support of a random variable $V \sim H_\zeta \left(\frac{v - \mu}{\delta} \right)$, is given by

$$S_X = \begin{cases} \left[\mu - \frac{\delta}{\zeta}, \infty \right], & \zeta > 0, \\ [-\infty, \infty], & \zeta = 0, \\ \left[-\infty, \mu - \frac{\delta}{\zeta} \right], & \zeta < 0. \end{cases} \quad (2.9)$$

The estimation of the generalized extreme value model parameters involves a block maximum based analysis with its associated estimation procedures and properties (Beirlant et al., 2004; Bensalah, 2000; Embrechts et al., 1999). In short, the block maximum approach is a data preparation procedure that involves taking the maximum volume recorded for that LOB level, within the sub-sample time increment used. The preparation of the volume process data, for fitting the generalized extreme value model, involves taking intra-daily data for each asset over the period 2010 and splitting the data into blocks. The maximum volume submitted in the sub-sample time block is recorded, producing a set of K ordered realized observations $\{v_{(1,K)}, \dots, v_{(K,K)}\}$. Comparing this to the α -stable preparation of the data, with K samples, the key difference is the use of the maximum volume, rather than the last volume recorded in the specified time increment, when constructing the observed time series data.

A number of methods have been proposed in the literature for estimating parameters in the generalized extreme value family. Here we focus on two such methods, MLE for cases

where $\zeta < 0.5$ (Prescott and Walden, 1980; Smith, 1985) and the method of L-moments (Hosking, 1990). We develop a mixed approach combining MLE and L-moments. A detailed discussion on L-moments can be found in Hosking (1990), however it is worth noting that in the context of the large data sets utilized for this study, a choice of a mixed method overcomes the instability found in the MLE parameter estimators (Hosking, 1990). In addition, the sample L-moments are numerically stable and robust when L-skewness and L-kurtosis estimators are used directly in L-moments estimation. It has also been observed by several authors, Royston (1992) and Vogel and Fennessey (1993), that the L-moments are less sensitive to outlying data values. The recent developments of mixed methods for inference provide, greater computational efficiency, statistical accuracy and robustness in the estimation.

A detailed discussion of the mixed MLE and L-moments based estimation, including asymptotic properties can be found in Morrison and Smith (2002). To present the mixed MLE and L-moments based approach, we first define the L-moments, given by Hosking (1990), for the real valued random variable V with distribution $F(v)$ and quantile function $Q(p)$, according to Definition 3.

Definition 3 (L-Moments). *The Population L-moments of a real valued random variable $V \sim F(v)$, for which there is a K sample realization with order statistics given by $V_{(1,K)} \leq V_{(2,K)} \leq \dots \leq V_{(n,K)} \leq \dots \leq V_{(K,K)}$, is defined as*

$$\lambda_r = r^{-1} \sum_{k=0}^{r-1} (-1)^k \binom{r-1}{k} \mathbb{E}[V_{(r-k,r)}], \quad \forall r = 1, 2, \dots \quad (2.10)$$

In practice, for such a mixed approach, the parameter range of the extreme value index is restricted to $\zeta \in [-0.5, 0.5]$, ensuring the moments are finite, and appropriate regularity conditions for the MLE are satisfied. Discussion on such items in the generalized Pareto distribution setting and the generalized extreme value setting, can be found in Beirlant et al. (2004, Chapter 5). We focus on the simplest mixed approach based on the MLE for ζ and L-moments for μ, δ , detailed for a sample size of K , according to the following stages.

Step 1 Re-parametrize the generalized extreme value model likelihood in terms of the extreme value index ζ . Express the parameters μ and δ as functions of ζ , via constraints on the population L-moments, given by

$$\begin{aligned} \lambda_1 &= \mathbb{E}[V_{(1,1)}] = \mu - \frac{\delta}{\zeta} (1 - \Gamma(1 - \zeta)), \\ \lambda_2 &= \frac{1}{2} \mathbb{E}[V_{(2,2)} - V_{(1,2)}] = -2 \frac{\delta}{\zeta} (1 - 2^\zeta) \Gamma(1 - \zeta). \end{aligned} \quad (2.11)$$

Step 2 **Estimate the population L-moments empirically via the sample L-moments.** This utilizes the ordered data realizations $\{v_{n,K}\}_{n \in \{1,2,\dots,K\}}$ given by

$$\hat{\lambda}_1 = \frac{1}{K} \sum_{n=1}^K v_{(n,K)}, \text{ and } \hat{\lambda}_2 = \frac{1}{K(K-1)} \sum_{n=1}^K \left\{ \binom{n-1}{1} - \binom{K-n}{1} \right\} v_{(n,K)}. \quad (2.12)$$

Then utilize these estimates to obtain the estimators for μ and δ with respect to ζ , according to the expressions incorporating these L-moment estimates given by

$$\begin{aligned} \hat{\delta} &= -\frac{1}{2} \zeta \hat{\lambda}_2 \left[(1 - 2^\zeta) \Gamma(1 - \zeta) \right]^{-1}, \\ \hat{\mu} &= \hat{\lambda}_1 + \frac{1}{\zeta} (1 - \Gamma(1 - \zeta)) \left\{ -\frac{1}{2} \zeta \hat{\lambda}_2 \left[(1 - 2^\zeta) \Gamma(1 - \zeta) \right]^{-1} \right\}. \end{aligned} \quad (2.13)$$

Step 3 **Perform maximum likelihood estimation for the extreme value index parameter ζ , subject to the constraints on L-moments imposed by the estimates in Stage 2.** The maximization of the re-parametrized likelihood for $\zeta \neq 0$, is given by

$$\begin{aligned} \ln l(\mathbf{v}_{(1:K,K)}; \hat{\mu}, \hat{\delta}, \zeta) &\approx -K \ln \left(-\frac{1}{2} \zeta \hat{\lambda}_2 \left[(1 - 2^\zeta) \Gamma(1 - \zeta) \right]^{-1} \right) \\ &- (1 + 1/\zeta) \sum_{n=1}^K \ln [1 + \zeta \mathcal{S}(v_{(n,K)}, \zeta)] - \sum_{n=1}^K [1 + \zeta \mathcal{S}(v_{(n,K)}, \zeta)]^{-1/\zeta}, \end{aligned} \quad (2.14)$$

where the function $\mathcal{S}(v_{(n,K)}, \zeta)$ is defined according to

$$\mathcal{S}(v_{(n,K)}, \zeta) = \left(\frac{v_{(n,K)} - \left(\hat{\lambda}_1 + \frac{1}{\zeta} (1 - \Gamma(1 - \zeta)) \left\{ -\frac{1}{2} \zeta \hat{\lambda}_2 \left[(1 - 2^\zeta) \Gamma(1 - \zeta) \right]^{-1} \right\} \right)}{-\frac{1}{2} \zeta \hat{\lambda}_2 \left[(1 - 2^\zeta) \Gamma(1 - \zeta) \right]^{-1}} \right). \quad (2.15)$$

If ζ is in the neighbourhood of the origin ($\zeta \in \text{n.e.}(0)$), the likelihood is given according the Gumbel limit of the generalized extreme value distribution, specified as

$$\begin{aligned} \ln l(\mathbf{v}_{(1:K,K)}; \mu, \delta, \zeta) &\approx -K \ln \left(-\frac{1}{2} \zeta \hat{\lambda}_2 \left[(1 - 2^\zeta) \Gamma(1 - \zeta) \right]^{-1} \right) \\ &- \sum_{n=1}^K \mathcal{S}(v_{(n,K)}, \zeta) - \sum_{n=1}^K \exp [-\mathcal{S}(v_{(n,K)}, \zeta)]. \end{aligned} \quad (2.16)$$

2.3.3 Generalized Pareto distribution parameter estimation

An alternative specification of such extreme value theory models, is the peaks over threshold based formulation, which is used for the generalized Pareto distribution model. For details of alternative data preparation and parameter estimation procedures, see Beirlant et al. (2004). In the case of peaks over threshold, the last observation in the time increment is recorded. Furthermore, we then retain the data that sits above a pre-

specified threshold. If the threshold was 120, for example, we would not include this observation in the model.

Definition 4 (Generalized Pareto Distribution). *A random variable $V \sim GP(\zeta, \delta)$, which has a distribution and density conditional upon translation to the origin, where the location parameter $\mu = 0$, given by*

$$F_V(v; \zeta, \delta) = \mathbb{P}r(V < v | V \geq \mu) = \begin{cases} 1 - \left(1 + \frac{\zeta v}{\delta}\right)^{-\frac{1}{\zeta}}, & \zeta \neq 0, \\ 1 - \exp\left(-\frac{v}{\delta}\right), & \zeta = 0, \end{cases} \quad (2.17)$$

$$f_V(v; \zeta, \delta) = \frac{\delta^{\frac{1}{\zeta}}}{(\delta + \zeta v)^{\frac{1}{\zeta} + 1}}, \quad (2.18)$$

with shape parameter $\zeta \in \mathbb{R}$ and scale parameter $\delta \in (0, \infty)$. In addition, the support of the density is given by

$$S_V = \begin{cases} [\mu, \infty), & \zeta \geq 0, \\ \left[\mu, \mu - \frac{\delta}{\zeta}\right], & \zeta < 0. \end{cases} \quad (2.19)$$

We consider three approaches to estimate the parameters of the generalized Pareto distribution, re-parametrized MLE estimation, analytic Pickands' estimators and robust versions of the empirical percentile method. The re-parametrization of the MLE makes the procedure numerically more robust.

We consider intra-daily data for each asset during 2010, at time resolutions of 1, 2, 5 and 10 seconds, with trading days split into blocks and utilizing a peaks over threshold approach. Under the peaks over threshold approach, the last volume recorded in the specified time increment is retained only if it exceeds the threshold. Comparing this to the preparation of data in Sections 2.3.1 and 2.3.2, we can see that the α -stable estimation considers the full sub-sampled intra-day data sets, the generalized extreme value considers the maximum volume per sub-sample time increment and the generalized Pareto distribution considers only the largest percentage of volume defined by the specified threshold.

Under the assumption that the volumes collected from the exceedence data are i.i.d. in the peaks over threshold approach, the likelihood for the generalized Pareto distribution, as a function of the absolute exceedence data, is given for the case in which $1 + \frac{\zeta V_j}{\delta} > 0$, by

$$\ln l(\mathbf{V}; \zeta, \delta) = -J \log \delta - \left(\frac{1}{\zeta} + 1\right) \sum_{j=1}^J \log \left(1 + \frac{\zeta V_j}{\delta}\right), \quad (2.20)$$

where the condition $1 + \frac{\zeta V_j}{\delta} > 0$ ensures the log-likelihood is finite. If $\zeta = 0$, the likelihood is given according to the exponential based distribution,

$$\ln l(\mathbf{V}; 0, \delta) = -J \log \delta - \frac{1}{\delta} \sum_{j=1}^J V_j. \quad (2.21)$$

Given the likelihood and the moments of the generalized Pareto distribution distribu-

tion or the quantile function, there are numerous statistical approaches we could adopt to perform the parameter estimation. First we discuss how to perform maximum likelihood estimation for such models. Maximization of the generalized Pareto distribution likelihood provided in (2.20) and (2.21) with respect to the parameters ζ and δ , is subject to the constraints:

1. $\delta > 0$;
2. $1 + \zeta v_{(J)}/\delta > 0$, where $v_{(J)} = \max\{v_1, v_2, \dots, v_J\}$.

This second constraint is important if $\zeta < -1$ as $-\delta/\zeta \rightarrow y_{(J)}$, then the likelihood approaches infinity. Hence, to obtain maximum likelihood parameter estimates, one should maximize the likelihood subject to these constraints, and $\zeta \geq -1$. It is well known that re-parametrizing the generalized Pareto distribution likelihood aids in the numerical stability of the parameter estimation via an MLE approach. A re-parametrized MLE version is detailed in Section 2.3.3, along with the generalized Pareto distribution estimation of parameters via method of moments, which is analytic for the parametrization we consider.

Re-parametrization of the generalized Pareto distribution log-likelihood and maximization

In practice, it is beneficial to consider a re-parametrization of the generalized Pareto distribution log-likelihood function according to

$$(\zeta, \delta) \rightarrow (\zeta, \tau) = \left(\zeta, \frac{\zeta}{\delta} \right), \quad (2.22)$$

producing a re-parametrized log-likelihood model given by

$$\ln l(\mathbf{V}; \zeta, \tau) = -J \ln \zeta + J \ln \tau - \left(\frac{1}{\zeta} + 1 \right) \sum_{i=1}^J \ln(1 + \tau V_i). \quad (2.23)$$

This log-likelihood is then maximized subject to $\tau < 1/v_{(J)}$ and $\zeta \geq -1$. Under the first partial derivative this produces

$$\begin{aligned} \frac{\partial \ln l(\mathbf{V}; \zeta, \tau)}{\partial \zeta} &= 0; \\ \Rightarrow \zeta(\tau) &= \frac{1}{J} \sum_{j=1}^J \ln(1 - \tau v_j). \end{aligned} \quad (2.24)$$

Hence, the estimation is performed in two steps:

1. Estimate $\hat{\tau}^{MLE} = \arg \max \ln l(\zeta(\tau), \tau)$, subject to $\tau < 1/v_{(J)}$;
2. Estimate $\hat{\zeta}^{MLE} = \frac{1}{J} \sum_{j=1}^J \ln(1 - \hat{\tau}^{MLE} v_j)$. Then solve for the original parametrization via inversion $\hat{\delta}^{MLE} = \frac{-\hat{\zeta}^{MLE}}{\hat{\tau}^{MLE}}$.

Note that the log-likelihood $\ln l(\zeta(\tau), \tau)$ is continuous at $\tau = 0$. Hence, if the estimator $\hat{\tau}^{MLE} = 0$, then one should consider $\hat{\zeta}^{MLE} = 0$ and

$$\hat{\delta}^{MLE} = \frac{1}{J} \sum_{j=1}^J v_j. \quad (2.25)$$

In addition, in practice to ensure that $\zeta \geq 1$, the condition that $\tau < 1/v_{(J)}$ should be modified to $\tau < (1 - \epsilon)/v_{(J)}$, where ϵ is found from the condition that $\zeta(\tau) \geq -1$.

Remark 1. *It has been shown in Smith (1985) and Embrechts et al. (1997, Section 6.5.1), the case in which $\zeta > -1/2$, the MLE vector $(\hat{\zeta}^{MLE}, \hat{\delta}^{MLE})$ is asymptotically consistent. Further, the MLE vector is distributed according to a bivariate Gaussian distribution with asymptotic covariance, which is obtained using the MLE parameter estimates, given by the usual Fisher information matrix. When the data is not independent, the results will continue to be consistent asymptotically normal, but the asymptotic covariance matrix is substantially more complicated than the independent case.*

Empirical percentile method

The empirical percentile method approach to parameter estimation is based on the percentile based matching approach proposed in Castillo and Hadi (1997). We equate the model CDF evaluated at the observed order statistics, to their corresponding percentile values. This system of equations can then be solved for the models distributional parameters. In the case of the generalized Pareto distribution model for the volume processes on the bid and ask at level 1, there are two model parameters, so we require two distinct order statistics, as a minimum, to perform the estimation.

Consider a set of realized data obtained under a peaks over threshold approach, where the J volumes that have exceeded a per-specified threshold level u , are denoted by the data $\{v_i\}_{i=1:J}$ with order statistics denoted by $\{v_{(i,J)}\}_{i=1:J}$.

Given the CDF of the generalized Pareto distribution model in (2.26)

$$F(v; \zeta, \delta) = \begin{cases} 1 - \left(1 - \frac{\zeta v}{\delta}\right)^{\frac{1}{\zeta}}, & \zeta \neq 0, \delta > 0, \\ 1 - \exp\left(-\frac{v}{\delta}\right), & \zeta = 0, \delta > 0, \end{cases} \quad (2.26)$$

matching the CDF at two of the selected order statistics $i \neq j \in \{1, 2, \dots, J\}$ to the corresponding percentile values

$$F(v_{(i,J)}; \zeta, \delta) = p_{(i,J)} \quad \text{and} \quad F(v_{(j,J)}; \zeta, \delta) = p_{(j,J)}, \quad (2.27)$$

where the percentile is given for the generalized Pareto distribution model with J observations by

$$p_{(i,J)} = \frac{i - \eta}{J + \zeta}. \quad (2.28)$$

It is recommended in Castillo and Hadi (1997), that choices of $\eta = 0$ and $\zeta = 1$ provide reasonable results, so these settings were utilized in the studies performed. The solution to

this system of equations, in terms of the parameters is obtained by solving the equations for ζ and δ given by

$$1 - \left(1 - \frac{\widehat{\zeta}v_{(i,J)}}{\widehat{\delta}}\right)^{\frac{1}{\widehat{\zeta}}} = \frac{i}{J+1} \quad \text{and} \quad 1 - \left(1 - \frac{\widehat{\zeta}v_{(j,J)}}{\widehat{\delta}}\right)^{\frac{1}{\widehat{\zeta}}} = \frac{j}{J+1}. \quad (2.29)$$

Hence for any two pairs of order statistics i, j the solutions to these system of equations is

$$\widehat{\zeta}(i, j) = \frac{\ln\left(1 - \frac{v_{(i,J)}}{\widehat{\delta}(i, j)}\right)}{C_i} \quad \text{and} \quad \widehat{\delta}(i, j) = \widehat{\zeta}(i, j)\widehat{\delta}(i, j), \quad (2.30)$$

in terms of $C_i = \ln(1 - p_{(i)}(J)) < 0$ and $\widehat{\delta}(i, j)$. Here $\widehat{\delta}(i, j)$ is the solution to the equation,

$$C_i \ln\left(1 - \frac{v_{(j,J)}}{\widehat{\delta}}\right) = C_j \ln\left(1 - \frac{v_{(i,J)}}{\widehat{\delta}}\right), \quad (2.31)$$

which is obtained using a univariate root finding algorithm, such as bisection. Note that δ corresponds to a re-parametrization of the generalized Pareto distribution distribution when $\delta = \frac{\delta}{\zeta}$.

Remark 2 (Empirical percentile method and Pickands' analytic solution). *A special case of the empirical percentile method estimators is widely used in estimation of the generalized Pareto distribution model parameters, known as the Pickands' estimator. This correspond to the empirical percentile method setting, in which $i = \frac{J}{2}$ and $j = \frac{3J}{4}$. In these special cases, the bisection method is not required, as the system of equations can be solved analytically according to*

$$\widehat{\zeta} = \frac{1}{\ln 2} \ln\left(\frac{v_{(J/2,J)}}{v_{(3J/4,J)} - v_{(J/2,J)}}\right) \quad \text{and} \quad \widehat{\delta} = \widehat{\zeta} \left(\frac{v_{(J/2,J)}^2}{2v_{(J/2,J)} - v_{(3J/4,J)}}\right). \quad (2.32)$$

In general, we would not just pick two indexes i, j , instead we would combine the Algorithm's one and two discussed in Castillo and Hadi (1997), to produce an estimate of the generalized Pareto distribution parameters. Combining Algorithm 1 and Algorithm 2 of Castillo and Hadi (1997) we follow the stages outlined below to estimate the generalized Pareto distribution parameters via the empirical percentile method.

Step 1 Repeat the following steps for all order indexes $\{i, j : i < j, \text{ for } i, j \in \{1, 2, \dots, J\}\}$, such that $v_{(i,J)} < v_{(j,J)}$.

- (a) Compute the values $C_s = \ln\left(1 - \frac{s-\eta}{J+\zeta}\right)$ for $s \in \{i, j\}$.
- (b) Set $d = C_j v_{(i,J)} - C_i v_{(j,J)}$, if $d = 0$ let $\widehat{\delta}(i, j) = \pm\infty$ and set the EVI estimate $\widehat{\zeta}(i, j) = 0$, otherwise compute $\delta_0 = v_{(i,J)}v_{(j,J)}(C_j - C_i)/d$.
- (c) If $\delta_0 > 0$, then $\delta_0 > v_{(j,J)}$ and the bisection method can be used for the interval $[v_{(j,J)}, \delta_0]$, to obtain a solution $\widehat{\delta}(i, j)$. Otherwise the bisection method is applied to the interval $[\delta_0, 0]$.

(d) Use $\widehat{\delta}(i, j)$ to compute $\widehat{\zeta}(i, j)$ and $\widehat{\delta}(i, j)$ using

$$\widehat{\zeta}(i, j) = \frac{\ln\left(1 - \frac{v(i, J)}{\widehat{\delta}(i, j)}\right)}{C_i} \text{ and } \widehat{\delta}(i, j) = \widehat{\zeta}(i, j)\widehat{\delta}(i, j). \quad (2.33)$$

Step 2 Take the median of each of the sets of estimated parameters for the overall estimator to obtain

$$\begin{aligned} \widehat{\zeta}^{EPM} &= \text{median} \left\{ \widehat{\zeta}(1, 2), \widehat{\zeta}(1, 3), \dots, \widehat{\zeta}(J-1, J) \right\}, \\ \widehat{\delta}^{EPM} &= \text{median} \left\{ \widehat{\delta}(1, 2), \widehat{\delta}(1, 3), \dots, \widehat{\delta}(J-1, J) \right\}. \end{aligned} \quad (2.34)$$

2.3.4 Goodness-of-fit testing for heavy tailed models of the limit order book volume process

In this section we present a class of omnibus compound goodness of fit hypothesis testing procedures, specifically designed for heavy tailed models.

To assess the quality of the statistical model estimations, we considered a number of approaches. The first involves an exploration of the goodness of fit utilizing the Kolmogorov-Smirnov statistical test (classical Kolmogorov-Smirnov test), which aims to test the compatibility of the theoretical probability distribution with the empirical probability distribution. The classical versions of the omnibus goodness of fit test, based on the Kolmogorov-Smirnov supremum statistic, proved inappropriate for two reasons: the first is it inadequately assesses the key feature we consider, namely the heavy tails; secondly, the massive data sets tend to result in each hypothesis being tested under thousands of samples, resulting in criteria so strict that large rejections will arise unless the test is directly targeting the appropriate null. Since the test is not correctly attributing appropriate weights to the sub-exponential tails of the model, the test will incorrectly reject the null, as the sample increases, at a disproportionate rate to reality (Chicheportiche and Bouchaud, 2012), which is what we observed in the testing of all distributions. This is due primarily to the test expecting an exponential, rather than a power law decay in the tail probabilities, see detailed discussion in Koning and Peng (2008).

Due to the inappropriateness of the classical Kolmogorov-Smirnov test for heavy tailed, we implement the variance weighted modified version of the Kolmogorov-Smirnov test discussed in Chicheportiche and Bouchaud (2012). To implement the modified Kolmogorov-Smirnov test, consider i.i.d. random variables $\{V_i\}_{i=1}^N$ with distribution F . We let $Y_n(v) = \mathbb{I}_{\{V_n \leq v\}}$, of which the components are Bernoulli variables. The centred sample mean measures the difference between the empirical CDF and the theoretical CDF at point v . We define the centred sample mean, $Y(v)$ as

$$\bar{Y}(v) = 1/N \sum_{n=1}^N Y_n(v) - F(v), \quad (2.35)$$

where N is the sample size. Let $u = F(v)$ and thus

$$\bar{Y}(u) = 1/N \sum_{n=1}^N \sqrt{N} Y_n(F^{-1}(u)) - u. \quad (2.36)$$

The variance weighted Kolmogorov-Smirnov test then has equi-weighted quantiles, which is equally sensitive to all regions of the distribution, including the tails. The resulting weighting is defined as

$$\tilde{y} = y(u) \sqrt{\phi(u; a, b)}, \quad (2.37)$$

where

$$\phi(u; a, b) = \begin{cases} \frac{1}{u(1-u)}, & a \leq u \leq b, \\ 0, & \text{otherwise.} \end{cases}$$

The choices for a and b are $a = 1/N$ and $b = 1 - a$, corresponding to the minimum and maximum of the sample (Chicheportiche and Bouchaud, 2012). It follows that we evaluate the variance weighted supremum test statistic according to

$$K(a, b) = \sup_{u \in [a, b]} |\tilde{y}(u)| = \begin{cases} \frac{y(u)}{\sqrt{u(1-u)}}, & a \leq u \leq b, \\ 0, & \text{otherwise.} \end{cases} \quad (2.38)$$

Under the null, the critical threshold value k^* , corresponding to a 95% confidence level, is determined from the following equation

$$\frac{-\ln 0.95}{\ln N} \approx \sqrt{\frac{2}{\pi}} k^* \exp -\frac{k^{*2}}{2}. \quad (2.39)$$

We implement a root finding method to determine k^* , which is dependent on the sample size N .

2.4 Results and discussions on model estimation for the limit order book volume process

Previous studies have failed to capture the features presented in Section 2.2 when building statistical models for LOB volumes. This has direct ramifications when attempting to model the dynamics of the LOB via advanced models such as the Hawkes process, that require the distributional specification of the marks derived from LOB volumes.

The purpose of this section is to formally study the features of the volume processes of level 1 bid and level 1 ask in the LOB for futures markets, using more flexible families of parametric models. We firstly assess the appropriateness of statistical models and fit, at different sampling frequencies. We then consider the appropriateness of a heavy tailed assumption for the volume processes each day, as captured by the sub-exponential family models for volume process tails given by, α -stable, generalized Pareto distribution and generalized extreme value distributional families. In the process of estimating these models for the LOB volume process data, we also assess, study and recommend

sophisticated statistical estimation procedures and their performance for each model. The estimation procedures include, MLE, generalized method of moments and L-moment estimation, mixed methods of MLE and generalized moment matching, empirical percentile estimation and quantile based estimators.

2.4.1 α -stable model estimation results

The α -stable family of distributions were fitted to volume data on level one of the bid and ask, scaled by the inter quartile range. For each day of data, we utilized McCullochs method to estimate the parameters. We began with analysis of the dynamics of the parameters that define the α -stable distribution across the 6 assets analysed, 5YTN, BOBL, SP500, NIKKEI, GOLD and SILVER.

We first recall that for the α -stable distribution, the tail index $\alpha \in (0, 2]$, determines the rate of tail decay and is the key parameter of interest. The parameter α is typically termed the characteristic exponent, with small and large α implying heavy and light tails, respectively. Gaussian ($\alpha = 2, \beta = 0$) and Cauchy ($\alpha = 1, \beta = 0$) distributions provide the only simple analytically tractable sub members of this family.

For the majority of 2010, the 5YTN has tail index parameter estimates indicative of light tails for volumes on the bid and ask, with a mean of 1.9895. However, a particularly interesting feature involved a few pronounced periods in which a heavy tail model is clearly appropriate for the volume process. The days on which these heavy tailed volume processes occurred, did not correspond to the same days for the bid and ask volume processes, indicating an asymmetry in the volume process on the bid and ask over time, with respect to extreme volumes. Additionally, volumes for the 5YTN also appear to be heavily right skewed, with a $\beta \approx 1$ a large portion of the year.

In Figure 2.10 we contrast the daily estimated findings for the 5YTN to BOBL. We can see significantly greater variation in the tail index α and skewness parameter β , with the mean of the tail index parameter, 1.8195. This indicates that the daily volume process on the bid and ask is consistently more heavy tailed than the daily behaviour of the volume process for the 5YTN. In addition, the extreme volume process events observed occasionally in the 5YTN, are not present in the BOBL volume process until the end of 2010, where an event caused the volume process to demonstrate an infinite mean tail behavior for a few trading days in late November to early December. This is consistent with the observed extreme events estimated from the α -stable daily model fits during this period.

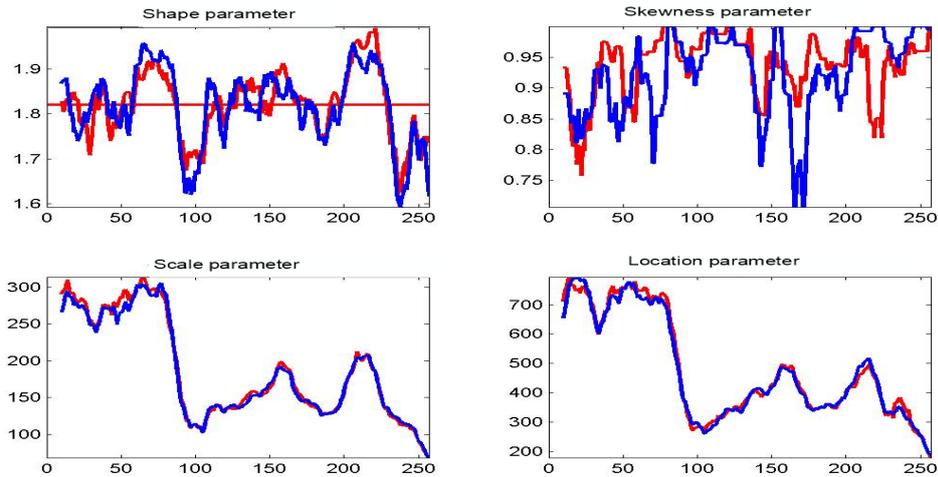


Figure 2.10: α -stable 10-day moving average of the daily parameter estimation, for the year 2010, using McCulloch's method for *BOBL*, at a time resolution of 10 seconds. *Top Left Plot*: tail index parameter α daily estimates. *Top Right Plot*: asymmetry parameter β daily estimates. *Bottom Left Plot*: scale parameter δ daily estimates. *Bottom Right Plot*: location parameter μ daily estimates. The blue line is the bid level 1 and the red line is ask level 1.

The daily volume processes for the LOB for NIKKEI are similar in nature to *BOBL*, with consistent heavy right tail attributes, strong skew and a mean tail index of 1.8022. Whilst there is symmetry between the bid and ask sides, there are some exceptions, with a marked relationship between asymmetry in the bid and ask volume processes, with the bid tending to produce a symmetric distributional fit when the ask is asymmetric, and vice versa.

Comparing the SP500 with the other assets considered, the estimated tail index parameter α is more pronounced, with a mean of 1.7474. In addition, there is a consistent daily tail profile, which has a tail index away from $\alpha < 2$. What is also of interest, is that when considering the SP500, which was globally the 2nd most actively traded equity future in 2010, there was actually a total volume decrease between 2009 and 2010 of -3.0% . However, this total change in volume did not affect the general attributes observed for the model estimation with regards to the heavy tailed behaviour, and the manner in which these contracts are traded on an intra-daily basis throughout the year.

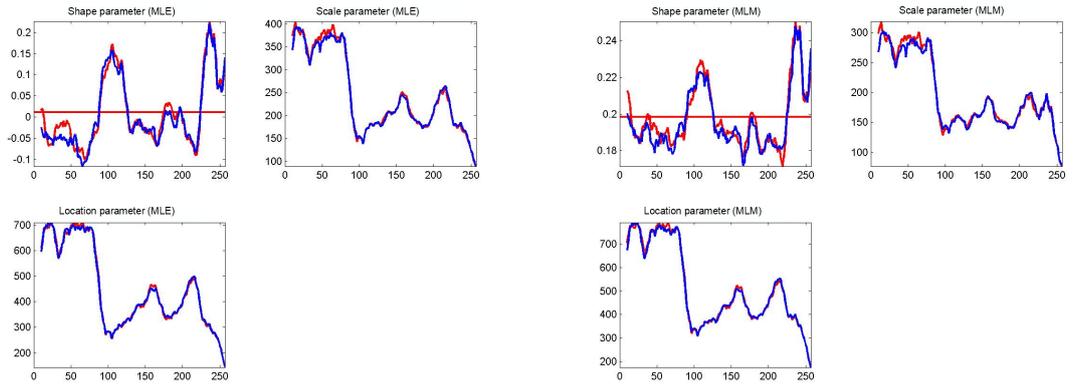
For the precious metals explored, the most prominent of the heavy-tailed volume processes is GOLD, which has a mean tail index parameter of 1.5076, and consistently heavy tailed behaviour intra-daily throughout all trading days in 2010. GOLD also demonstrates a strongly right skewed distribution for the volume process at level 1 of the bid and ask. The results for SILVER demonstrate a few marked periods in which the intra-daily volume process, on both the bid and ask, become exceptionally heavy tailed, most noticeably in the mid-year in which the ask side has tail index values around $\alpha \approx 1$.

2.4.2 Generalized extreme value model estimation results

Consistent with the findings from the α -stable model, the estimation results for the 5YTN show heavy-tailed behaviour is present in the volume processes between the 50th and 60th trading day, and the 140th and 150th trading days. Interestingly, the prominence of the heavy tailed features for BOBL is more pronounced under the generalized extreme value model fits, compared to the α -stable model. Additionally, the structural changes in the behaviour of the intra-daily volume process on the bid and ask are observed in the location and scale parameters for the BOBL around the 100th trading day, where there is a marked regime shift in the estimated model parameters. This is just as prominent in the generalized extreme value fit as it is in the α -stable model, indicating that it is entirely plausible that there was a dynamic change in the intra-day activity in this market mid trading year. A similar change is visible in the SP500 shape and scale parameters, but in this case the regime reverts gradually back to the behaviour present intra-day at the start of the year. This structural change is not as prominent in NIKKEI.

When considering the tail index parameter ζ , for the MLE method, we can see that the mean intra-daily estimated value averaged over all trading days in 2010 for the 5YTN, BOBL (Figure 2.11) and NIKKEI is close to zero, respectively, -0.0493, 0.0114, 0.0725. SP500 has a higher mean level for the shape parameter, being 0.1495. Again, consistent with the α -stable case, the generalized extreme value model estimations do indicate a reasonable variation in the tail index throughout 2010, indicating a number of days in which heavy tailed volume process attributes are appropriate. On a few days analysed, and for assets 5YTN, SP500 and NIKKEI, there are instances where the shape parameter spikes ($\zeta > 1$) indicating infinite mean-variance models are suitable. The occurrence of such events coincides with the trading days in which the α -stable model also indicated infinite mean heavy tailed behaviour as suitable. For all assets there is a correlation in the parameter estimations for the bid and ask side, and all assets show time variation across the year for scale and location.

The mixed L-moments approach is used to confirm our findings of the estimation of the tail index parameter, which is notoriously difficult to estimate. Comparing the subplots in Figure 2.11 for BOBL, we have a good representation of the features we found to be consistent across the assets 5YTN, SP500 and NIKKEI. The scale and location parameters are very similar for each method implemented. However, the shape parameter is systematically different when comparing the MLE and mixed L-moments methods, with the mean level for the shape parameter using the MLE method being between $(-0.0495, 0.1495)$ for all assets, whereas the mixed L-moments method produces a mean level of $(0.1951, 0.2321)$ for the shape parameter for all assets.



(a) MLE approach.

(b) Mixed L-moments approach.

Figure 2.11: Generalized extreme value intra-day 10-day moving average of the parameter estimation on each trading day of the year, for BOBL, bid and ask side, at a time resolution of 10 seconds. The blue line is the bid level 1 and the red line is ask level 1.

To further explore the discrepancies between the MLE and mixed L-moments, specifically the upward translation of the shape parameter by approximately 2, we performed a simulation study, which considered sample sizes ranging from 50 to 10,000. We randomly generated generalized extreme value distributed data series and replicate each sample 20 times. We then estimate the parameters of the generalized extreme value distribution for each replication and each sample size using MLE and mixed L-moments methods. Results show the same discrepancies between the different estimation methods for the shape parameter. As $\zeta \rightarrow 0$, there is an increased bias of ζ under the mixed L-moments method, but the trade-off is that the variance is reduced in the mixed L-moments method. Mixed L-moments becomes more reliable as the sample size decreases, thus prompting us to recommend the use of MLE for higher frequency LOB volume data.

In summary, we observed that the MLE method provided a more stable fit compared with the mixed L-moments method for these applications. Interestingly, we found from the analysis that as the data becomes significantly heavier tailed, as was the case for GOLD with a mean intra-day tail index parameter of 0.3564 for 2010, the results for the MLE estimation and the L-moments based solutions were in much closer in alignment. The observed bias present in the cases of light tailed volume processes on certain trading days in BOBL, was not present in the consistently heavy tailed GOLD.

2.4.3 Generalized Pareto distribution model estimation results

The generalized Pareto distribution family utilize a peaks over threshold preparation of the data, with a translation by a threshold corresponding to the 80th percentile of the data (the location (μ) parameter) (Embrechts et al., 1999). The features observed under the generalized Pareto distribution model are consistent with the findings discussed for the α -stable and generalized extreme value models. In particular, the prevalence for the heavy tailed attributes remain, as does the interesting features of the increased extreme

intra-day volume activities resulting in heavy tailed attributes appearing for BOBL, as observed in the generalized extreme value and α -stable fits.

The scale parameter δ , for the generalized Pareto distribution distribution using the MLE method for estimation, shows some structural shifts for 5YTN and SP500 in Figure 2.12, around the same time period as the structural shifts observed in the α -stable distribution.

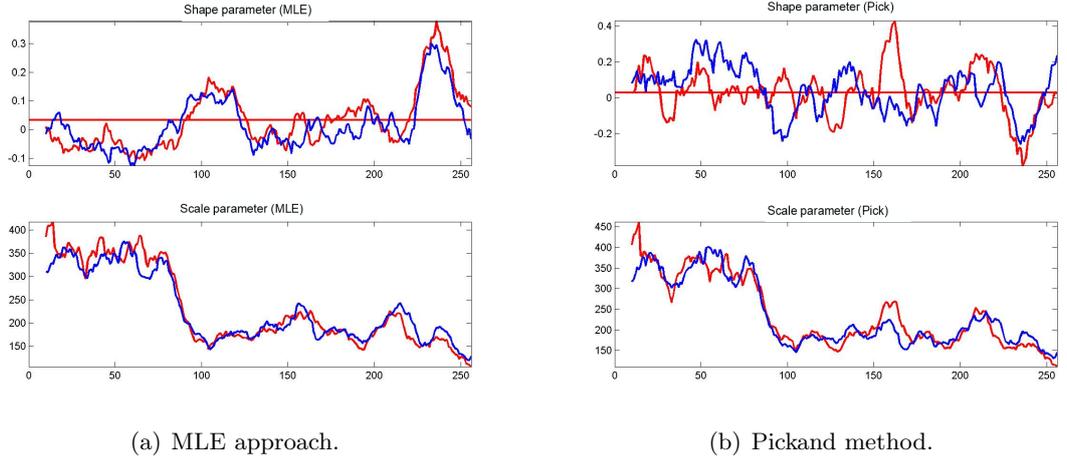


Figure 2.12: Generalized Pareto distribution intra-day 10-day moving average of the parameter estimation for each trading day of 2010, for BOBL, bid and ask side, at a time resolution of 10 seconds. The blue line is the bid level 1 and the red line is ask level 1.

In terms of the different estimation approaches, the scale parameter using the Pickands' estimator shows similar trending features as the MLE method. However, it also exhibits much higher variability in parameter estimation compared with the MLE method (Figure 2.12). The Pickands' estimator fails to capture the days where there is significant spikes in the shape and scale parameters.

The results observed for the empirical percentile method, demonstrate a substantial deviation from the MLE and the Pickand's estimator. The implementation of the empirical percentile method involved matching each pair of percentiles above a threshold of the median to obtain a solution for the model parameters, producing a very large number of solutions. Next we took the median of these solutions as a robust estimation of the model parameters under the empirical percentile method (Castillo and Hadi, 1997). Our findings indicated the sensitivity of this estimation approach to the inclusion of low percentile solutions into the estimation (median calculation) for the shape (tail index) parameter.

To further explore why the empirical percentile method gives substantially different results to the MLE and Pickands' methods, we considered a synthetic simulation study with 20 data sets, each with 500 randomly generated generalized Pareto data points used to estimate the MLE, Pickands' and empirical percentile method. We considered the impact of setting the shape parameter positive and negative. The results show that for a positive shape parameter, the three methods appear to be consistent. However, when the actual shape parameter is negative, there is a significant translation upwards, with

increased variability in the estimator.

We note that the empirical percentile method of estimation has significant issues when attempting to apply this method to real data. We found that the method produces a much more stable result when using higher starting percentiles for the grid search method for maximizing the log-likelihood function. However, for the simulated and real case there seems to be a systematic bias in the estimation under the empirical percentile method, but not the MLE when the data appears to have an *actual* negative shape parameter. It should also be noted that the starting percentile used for this method was the 50th percentile.

In general we found little difference in the estimated model parameters at each of the sampling frequencies. The consistent presence of these features at all these sampling rates allows us to conclude that such activities tend to take place at the high frequencies (< 1 second), providing some evidence that this may be prevalent on irregularly spaced time intervals, which will be discussed in the following chapters.

2.5 Goodness-of-Fit via variance weighted modified Kolmogorov-Smirnov test

In this section we consider each of the three fitted model estimations daily for 2010. On each day of estimation we perform a goodness of fit analysis using the specifically modified variance weighted Kolmogorov-Smirnov test for each model. Interestingly, we observed that when using the modified variance weighted Kolmogorov-Smirnov test, we found that the resulting test rejects the null hypothesis of the α -stable distribution providing a good fit to the data for all assets. Hence, we conclude that whilst the model was able to be estimated efficiently, the fit is not sufficiently reflective of the data. The modified Kolmogorov-Smirnov test rejects the null hypothesis of the generalized extreme value distribution providing a good fit to the data for all assets using the mixed L-moments method. However, the MLE method provides a good fit to the data for SP500 22% of the trading days for the bid side and 17% of the trading days for the ask side. For BOBL and NIKKEI, this provides a good fit less than 10% of the trading days on both the bid and ask side.

From Figure 2.13, we can see that the modified KS test does not reject the null hypothesis at a 5% level of significance for the generalized Pareto distribution using the MLE estimation method. For the 5YTN, the MLE estimation method provides a good fit for 72% of trading days for the bid and 76% of trading days for the ask. Whereas, the Pickands' method only provides a good fit for 8% of trading days on the bid and 6% of trading days on the ask for the 5YTN. These results are largely consistent with all other assets, however the generalized Pareto distribution MLE method provides a good fit for 95% of days for the bid and 88% of days for the ask for the SP500. Likewise, the Pickands' results were also better for the SP500, with a good fit being observed 17% of trading days on the bid and 10% on the ask side. We do not present the results for the empirical percentile method for the modified KS test, due to the estimation issues outlined above.

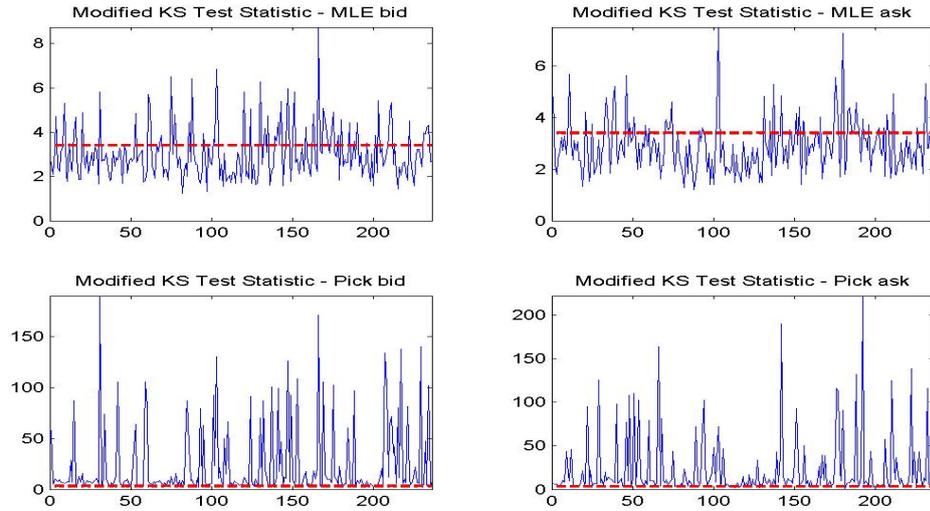


Figure 2.13: Modified KS-Test statistics for the 5YTN, for every trading day, bid and ask side, using generalized Pareto distribution, MLE and Pickands’ method.

We considered 6 different estimation methods across three different distributions, α -stable, generalized extreme value and generalized Pareto distribution. Table 2.5 shows the superior method for each of the distributions. Across all methods, the generalized Pareto distribution MLE method provided the best fit to the data across all assets. The parameter estimates did not vary significantly with a change in sampling rate. The weighted modified Kolmogorov-Smirnov test concluded the generalized Pareto distribution MLE method provided a good fit for LOB volume data over 70% of trading days, across all assets on the bid and ask side.

Table 2.5: Superior method fitted to volume data for six assets, 5YTN, BOBL, SP500, NIKKEI, GOLD and SILVER. The methods include, MLE, method of moments, McCullochs quantile based estimation (McCulloch, 1986), Pickands’ estimator (Pickands III, 1975) and empirical percentile method

Method	α -stable	General. extreme value	Generalized Pareto distribution
MLE		Best	Overall best
McCullochs	Best		
Mixed L-moments			
Pickands		Worst	Second
Emp. percent. method			Failed

2.6 Conclusion

The aim of this chapter was to create the building blocks for a dynamic model of the LOB, by considering the long memory in the dependence structure, the heavy tail features in the marginal distributions and the impact of sampling frequency. By implementing two procedures, the Hurst exponent and the more recent technique being the extremogram, this research confirmed the finding of long range dependence across all assets, with increasing dependence for shorter time increments. This forms a key consideration for modelling the dependence structure within the LOB volume processes. We further explored the importance of a study on heavy tailed marginal distributions in the context of the observed dependence. The large sample sizes on which to fit the extreme value distributions were sufficient to obtain consistent estimators of the marginal distribution and the heavy tailed features.

Researching the second objective, we considered 3 distributions across 6 futures assets for an entire year of trading (2010). From this analysis we attempted to model and assess the heavy tail volume process features present in the data, at each level of the bid and ask in the LOB. In addition to the different distributional fits considered, we explored many different parameter estimation methods within each distribution to ensure that we have a method that is robust to the ultra-high frequency data analysed, and a method that best captures the features present in the data.

There was statistical evidence for the presence of sub-exponential volume process right tails at different levels of the order book for all models considered, the α -stable, generalized extreme value, generalized Pareto distribution models. All models demonstrated consistent results for the presence of intra-day periods of heavy tails, in which the tail index parameter spiked in an asymmetric manner on level 1 of the bid and ask volume processes. With regard to the notion that perhaps the heavy tailed features may be exchange specific, due to possible effects related to specific exchange mechanisms or market participants in such exchanges, we found that the potential for heavy tailed features in the intra-day volume processes was present in all exchanges considered in the analysis (CBOT, EUREX, SGX, CME and COMEX). In addition, the asymmetry present in the times of occurrence of the heavy tailed features of the bid and ask volume processes was also not exchange specific. Interestingly, particular asset classes had a greater tendency to systematically display heavy tailed volume profiles at all levels on the bid and ask. For instance, we demonstrated that the precious metals, GOLD and SILVER, displayed heavy tailed intra-day behaviour systematically throughout the trading year, whereas the 5YTN and BOBL displayed such features intermittently throughout the year.

Due to the range of high frequency sampling rates considered at 10 seconds, 5 seconds, 2 seconds and 1 second intra-day for each trading day of 2010, we were able to disambiguate the real stochastic behaviour of the heavy tailed features of the LOB volume process structures from the high-frequency micro-structure noise (Dayri, 2012). We showed that for a given market and asset class, the presence of statistical features, such as the heavy tailed attributes, were persistent in the stochastic processes for the range of sampling resolutions. The presences of these features at the lowest resolution of 1 second,

is suggestive that a study across irregularly spaced time intervals is likely to exhibit similar features for volume based marks.

During non-heavy tailed volume process intra-day time periods, we found the bid and ask LOB marginal dynamics to generally follow a strong common trend. However, during heavy tailed events, there was an asymmetry in the volume processes on the bid and ask at all levels. It was also clear that there is strong statistical evidence from the model estimations undertaken to recommend that when developing parametric models for the volumes on the bid and ask, the models should incorporate both intra-day and inter-day dynamics for the model parameters, time varying volatility and heavy tailed features. This was clear, since all assets demonstrate time variation for the parameters across the year, with intra-day and inter-day variations present. For BOBL, SP500 and NIKKEI, the large number of times the shape parameter dropped below 1.8 was indicative of why we found the Normal and Gamma distributions to be a bad fit, even after simple transformations were applied. Skewness for the SP500 is aligned with the findings for 5YTN, with a $\beta = 1$ a large portion of the year. Both BOBL (Figure 2.10) and NIKKEI demonstrated left skewness, but less pronounced compared with 5YTN and SP500.

To assess the quality of the statistical model estimations, a variance weighted modified Kolmogorov-Smirnov test, which attributes appropriate weights to the sub-exponential tails of the model was considered. The generalized Pareto distribution using the MLE parameter estimation method was the best fitting model. The model provided a good fit of between 70%-100% of trading days for all assets on both the bid and ask side.

A dynamic model would need to incorporate a number of key components to capture all of the features observed in this analysis. Computationally efficient methods of parameter estimation are critical in the context of the large data sets. All other statistical approaches for modelling LOB volume processes in financial literature are invalidated with the findings of the high levels of skewness, kurtosis and the heavy tailed features in the marginal distribution. Dynamic parameter estimates would be required to capture the time varying inter day parameters if modelling across days. For volume based marks, and in the context of a Hawkes process, assessment of long memory and serial dependence observed in the upper tails will be necessary. In addition, the generalized Pareto distribution with an MLE estimation method is a likely candidate for the marginal distribution of volume based marks that exhibit heavy tails.

Chapter 3

Construction of the observed limit order book and event process

3.1 Physical, reported and observed data

This section provides detail of the irregularly spaced data that is used throughout the remainder of this research. Section 1.3 provides an introduction to the LOB. We begin by introducing the Physical LOB as a continuous time construction of prices and volumes. We then move into the Reported LOB, describing data types and specifically the product, Thompson-Reuters Tick History (TRTH). We review a number of studies that address the limitations of this product and the subsequent impact this may have on modelling the LOB. We then consider the Observed LOB and the three key steps of matching the market depth with trades, defining the event types and aggregating the data onto a unique time stamp. Finally, we consider the event process and mark features we extract from the Observed LOB for modelling purposes. Figure 3.1 provides a schematic of these four representations of the LOB, of which we describe each in detail.

3.1.1 The physical limit order book

A thorough overview of financial markets can be found in Harris (2003). Abergel et al. (2016) provides a comprehensive introduction to LOBs. The order driven market is an electronic platform that aggregates all available orders in a LOB. Orders are submitted in a double auction and they are matched as they arrive over time, subject to time and price priority rules (Abergel et al. (2016)).

In continuous time t , the LOB is comprised of prices $P_t^{(s,l)}$ and volumes $V_t^{(s,l)}$, at level $l \in \{1, \dots, d\}$ residing on the bid or ask side $s \in \{B, A\}$. The state of the LOB is modified by events such as limit orders, market orders and cancellations. These events can occur at any time and they can occur simultaneously. For example, a market order may alter the state of the LOB on the bid side price level 1, at the same time a limit order may be submitted on the ask side at level 3, and two cancellations made at level 1 and level 4 of the bid side.

3.1.2 The reported limit order book

Types of Data

The list below describes the types of data common to studies that utilize financial market data. At this stage it is important to distinguish between a price level 3 and Level III data. Price level 3 is the third price away from the best bid (or ask) in the LOB. Whereas, Level III data refers to the description of the LOB datasets and has nothing to do with the third price level. We formally define the matched LOB in Definition 5, however for the description below, let t_i be the event time, $l \in \{1, \dots, L\}$ be levels of the LOB for each side $s \in \{B, A\}$. For the majority of assets we consider, the number of LOB levels provided is $L = 10$, however some assets only have LOB levels $L = 5$ available.

- **Level III** data refers to LOB data that contains all transaction data, inclusive of limit orders, market orders and cancellations at the granularity of an individual order submission. Level III data contains enough information to replay the market. The TRTH database does not provide Level III data and it is not used in this research.
- **Level II or Market Depth** data refers to LOB data that contains all data for a pre-specified number of price levels, where for the bid side

$$P_{t_i}^{(B,l=1)} > P_{t_i}^{(B,l=2)} > \dots > P_{t_i}^{(B,l=L)},$$

and for the ask side

$$P_{t_i}^{(B,l=1)} < P_{t_i}^{(A,l=2)} < \dots < P_{t_i}^{(A,l=L)},$$

with a consolidation of volume at each price level. For example, price level $P_{t_i}^{(B,l=1)}$ may be made up of one or more individual orders at that price. The volume at that price level represents the sum of the volume of all individual orders submitted at that price level. Contrasting this to Level III data where all of the individual orders at that price level and all price levels that exist, not restricting levels specifically to, for instance, thresholds of $L = 5$, $L = 10$, which is common in industry practice for Level II data.

- **Level I** data refers to the best bid and best ask. Strictly speaking, the definition of Level I data is the very highest price and associated volume for the bid order, likewise, the very lowest price and associated volume for the ask order. Often the top level of the Level II data will be called the Level I data or BBO (best bid offer), even though it is inclusive of all orders at that price level.
- **Trade data**, also known as *Time and Sales* data refers to the orders that have being executed, realized as trades and forming the traded price of the instrument.
- **Matched data** is comprised of LOB data (either level I, level II or level III data) that is time matched with trade data. Because the data is event driven, there will not always be a new trade every time the LOB changes. If the LOB changes and a

new record is made, but no trade has been executed at that time, then nothing will be recorded for the trade. If a trade occurs, but there are no changes in the LOB at that time, the prevailing state of the LOB will be associated with that trade.

- **Transformed data** in context of this study, refers to LOB and trade data that has been altered in some way to construct variables or marks. An example may include the spread, event type volume or the last traded price over evenly spaced time intervals for the day.
- **Time series data** is a version of transformed data and consistent with the above example, it is data that is aggregated in some way over evenly spaced time intervals.
- **End of day data** is a summary over the entire day.

The data used in this study originates from the Thompson-Reuters Tick History (TRTH) database. Time is recorded to a minimum time granularity of one millisecond and is therefore recorded on a discrete grid. The market depth (LOB) file contains the Reported LOB and the trade file contains the reported market orders (executed trades). The market depth file is consolidated at the price level, that is, the volumes of all limit orders submitted at price level l , will be aggregated into a single line of total volume at that price level. This product provides data such as trades, level I and level II, although the level II data is consolidated by price and only a specified number of price levels are available for each market.

Filimonov and Sornette (2015) are the first to provide such a comprehensive review of the biases present when dealing with this type of data. They present a description of the grouping and the ‘bundling effect’, where the reported data results in identical timestamps. As Filimonov and Sornette (2015) outline, the origin of this is due to the FAST/FIX protocol bundling multiple updates within a single message by an algorithm designed by the exchange. The time stamps within the reported data are time stamps provided by the vendor when the data arrives, not the exchange time stamps when the event actually occurred. The exchanges may report one second time stamps and the vendor will attempt to enrich the data by providing a millisecond time stamp when the data arrives.

The discussion by Filimonov and Sornette (2015) outlines the uncertainty this process introduces into the datasets, such as packing/unpacking FAST/FIX package, latency from exchange to vendor and grouping of multiple events, with these uncertainties resulting in tens of milliseconds. To investigate the impact of possible bias due to bundling of events, using numerical simulations Filimonov and Sornette (2015) found an overestimation of the branching ratio, similar to the effect of outliers. They proposed that studies should be confined to a one second resolution timestamps of the exchange or be complemented with model calibration with different assumptions on interval durations. However, it was also highlighted that the negative impact of bundling events may be mitigated by the overhead for data processing, and the two biases may partially compensate each other in the calibration of the Hawkes model when using the TRTH database (Filimonov and Sornette

(2015)). With these data limitations in mind, we proceed with caution and with a research focus on careful calibration of the Hawkes process.

3.1.3 The observed limit order book

In view of comments in Section 3.1.2, without further processing the Reported LOB is not in a form that is suitable for modelling. To achieve a data set that contains the most accurate information possible about the state of the Physical LOB in any millisecond time window, additional stages of data processing are needed and these are summarized in Figure 3.1.

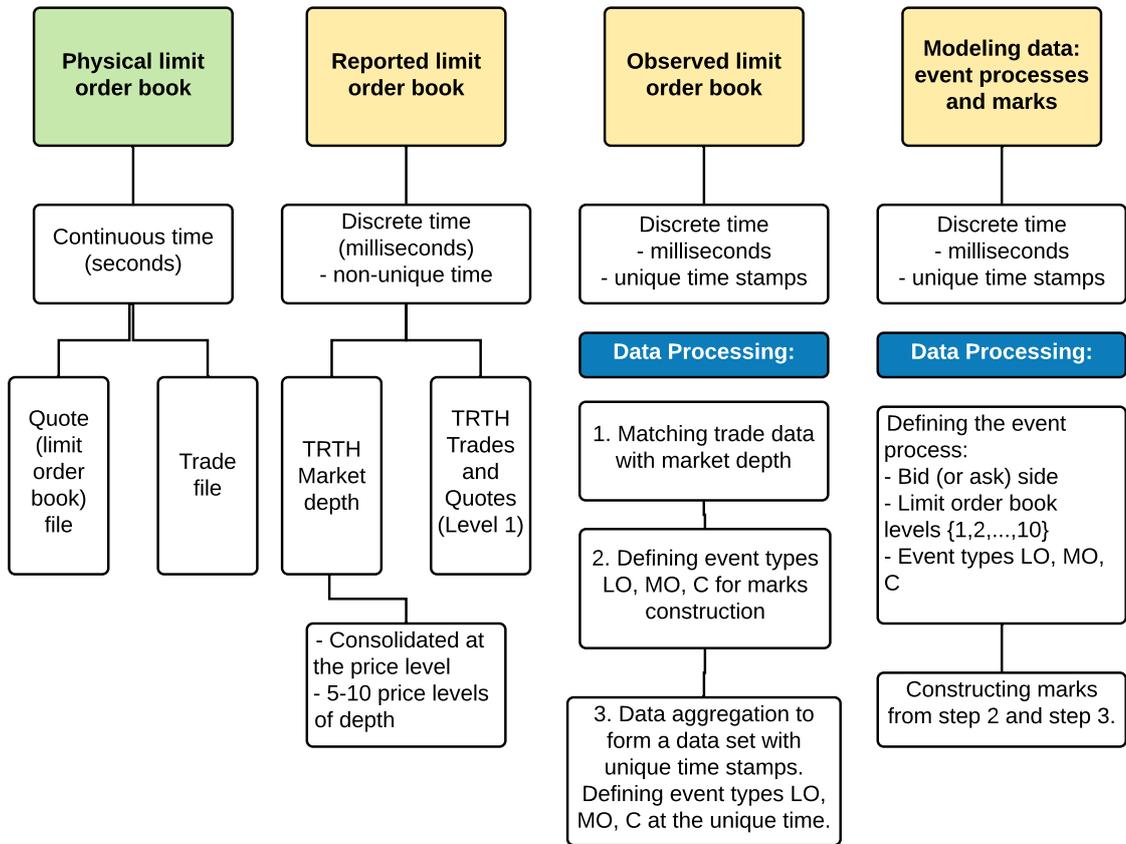


Figure 3.1: Flow chart of the various stages of data processing.

Stage 1: Matching limit order book and trade data

The purpose of matching the LOB and trade data, is to facilitate the classification of event type (limit order (LO), market order (MO) and cancellation (C)) and in addition, gather further information to define the event process and marks that may be attached to the events. Failure to match the data will result in the inability to decide whether a reduction in volume on the LOB was due to a market order or cancellation. The difference in impact that can arise from these two types of events in modelling the LOB is non-trivial.

Construction of matched data by example. We present a simple example of the matching processing using actual market depth, and trades and quotes data. In this simple example, we use an exact matching of time stamps and show the failure of this exact matching process. The example of matching uses the two tables presented below to show the market depth data for 2 price levels (Table 3.1), and the trades and quotes data (Table 3.2) that will be used to construct the final table of matched data (Table 3.3).

Table 3.1: *Market depth data.* MD=Market Depth, BP=Bid Price, AP=Ask Price, BV=Bid Volume, AV= Ask Volume, N=number.

Time	Type	L1 BP	L1 BV	L1 N. Buy	L1 AP	L1 AV	L1 N. Sell	L2 BP	L2 BV	L2 N. Buy	L2 AP	L2 AV	L2 N. Sell	
BHP	10:53.4	MD	44.54	307	3	44.55	2510	1	44.50	27084	12	44.56	3002	2
BHP	10:53.5	MD	44.54	303	2	44.55	2510	1	44.50	27084	12	44.56	3002	2
BHP	10:53.6	MD	44.54	303	2	44.55	2510	1	44.50	27084	12	44.56	3002	2
BHP	10:53.6	MD	44.54	303	2	44.55	2510	1	44.50	27084	12	44.56	3002	2
BHP	10:53.6	MD	44.54	303	2	44.55	2532	2	44.50	27084	12	44.56	3002	2
BHP	10:53.7	MD	44.54	303	2	44.55	2532	2	44.50	27084	12	44.56	3002	2
BHP	10:53.7	MD	44.54	232	1	44.55	2532	2	44.50	27084	12	44.56	3002	2
BHP	10:53.7	MD	44.50	22650	12	44.55	2532	2	44.49	1023	2	44.56	3002	2
BHP	10:53.8	MD	44.50	24850	13	44.55	2532	2	44.49	3223	3	44.56	3002	2
BHP	10:53.8	MD	44.50	24850	13	44.55	2532	2	44.49	3223	3	44.56	5202	3
BHP	10:53.8	MD	44.50	24850	13	44.55	2532	2	44.49	3223	3	44.56	5202	3
BHP	10:53.8	MD	44.52	3249	1	44.55	2532	2	44.50	24850	13	44.56	5202	3
BHP	10:53.8	MD	44.52	3249	1	44.55	2532	2	44.51	2200	1	44.56	5202	3
BHP	10:53.8	MD	44.55	12468	1	44.56	5202	3	44.52	3249	1	44.57	5000	1

Table 3.2: *Trades and Quotes.*

Instru- ment	Time	Record Type	Price	Volu- me	Value	Bid Prices	Bid Size	Ask Price	Ask Size	Qual- ifiers
BHP	10:53.5	TRADE	44.54	4	178.16					XT
BHP	10:53.5	QUOTE					303			
BHP	10:53.6	QUOTE						2532		
BHP	10:53.6	TRADE	44.54	71	3162.34					XT
BHP	10:53.7	QUOTE					232			
BHP	10:53.7	TRADE	44.54	232	10333.28					XT
BHP	10:53.7	TRADE	44.50	4434	197313.00					
BHP	10:53.7	QUOTE				44.5	22650			
BHP	10:53.8	QUOTE					24850			
BHP	10:53.8	QUOTE				44.52	3249			
BHP	10:53.8	TRADE	44.55	22	980.10					
BHP	10:53.8	TRADE	44.55	2510	111820.50					

Table 3.3: *Matched data*. MD=Market Depth, LO=Limit Order, B=Bid, A=Ask.

Instru- ment	Time	Origin	Price	Volu- me	Num- ber	Type	Levels	Qual- ifier	L1-B	L1-A	Note
BHP	10:53.5	TRADE	44.54	4	1	Trade	B1	XT	44.54	44.55	(1)
BHP	10:53.6	MD	44.55	22	1	LO	A1		44.54	44.55	(2)
BHP	10:53.6	TRADE	44.54	71	1	Trade	B1	XT	44.54	44.55	(3)
BHP	10:53.7	TRADE	44.54	232	1	Trade	B1	XT	44.54	44.55	(4)
BHP	10:53.7	TRADE	44.50	4434	1	Trade	B1		44.50	44.55	(5)
BHP	10:53.8	MD	44.50	200	1	LO	B1		44.50	44.55	(6)
BHP	10:53.8	MD	44.49	3000	1	LO	B2		44.50	44.55	(7)
BHP	10:53.8	MD	44.56	2200	1	LO	A2		44.50	44.55	(8)
BHP	10:53.8	MD	44.52	3249	1	LO	B1		44.52	44.55	(9)
BHP	10:53.8	MD	44.51	2200	1	LO	B2		44.52	44.55	(10)
BHP	10:53.8	TRADE	44.55	22	1	Trade	A1		44.52	44.55	(11)
BHP	10:53.8	TRADE	44.55	2510	1	Trade	A1		44.55	44.56	(12)
BHP	10:53.8	MD	44.55	12468	1	LO	B1		44.55	44.56	(13)

The commentary below is linked to the 'Note' Column of Table 3.3, giving a description of the events and how this might translate into the matched data.

1. The market depth data (Table 3.1) shows a decrease of 4 units on the bid. This could be the result of either a cancellation or trade. By checking the trades and quotes data (Table 3.2) we can see a trade occurred at the bid price of 4 units, thus concluding it was a trade.
2. Arrival of a limit order on the best ask of 22 shares at 44.55.
3. From the market depth data (Table 3.1), we can see that this trade occurs on the buy side. However, the time stamp on the market depth data does not match that of the trades and quotes data (Table 3.2), with a difference of 1ms.
4. Trade at 44.54 of the remaining 232 shares at the best bid, moving the best bid down to 44.50.
5. Trade at 44.5 at the new best bid price. The spread has widened by 4 cents.
6. Limit order of 200 shares is placed on the best bid.
7. Recorded in the same time stamp as (6) in the market depth data, a new limit order is placed on level 2 of the bid for 3,000 shares.
8. Limit order placed on level 2 of the ask for 2,200 shares.
9. Limit order of 3,249 shares placed inside the spread creating a new best bid of 44.52 and changing the spread from 5 cents to 3 cents.
10. Limit order of 2,200 placed on level 2 of the bid. The previous level 2 price was 44.52, it is now 44.51 with the orders at 44.52 forming the level 3 (not shown in this data).
11. The best bid was 44.52 and is now 44.55. Events (11), (12) and (13) create this change. The first event, the best ask was 44.55 and is now 44.56. The 2,532 shares that were on the best ask at 44.55 have being traded in two lots (22 + 2,510) shares.

12. The second of the two lot trades on the best ask (22 + 2,510) shares.
13. At the same time as the above two items (within the same line on the market depth data), a limit order has been submitted on the best bid for 12,468 shares.

Matching the LOB and trade data is complicated, due to the issues discussed in Section 3.1.2 and as shown in the previous example. Research presented by Toke (2016) outlines an algorithm used to match the trade data with the LOB data, specifically the reported TRTH data used in this research also. Research by Toke (2016) is one of the most comprehensive discussions we have found on the matching process and it presents the often undervalued, but extremely important practical choices one must make in the data preparation stages of researching the LOB.

Toke (2016) initially defines the limit order and cancellations. He then presents and compares three matching algorithms: 1. the perfect match; 2. grouping trades to match a single LOB change on the same time stamp; 3. group trades that do not have exactly the same time stamp to match the LOB change. The third matching algorithm is used to reassign cancellations to market orders where a trade has occurred and accounting for volume traded.

This differs to the method used in this research, namely LIFO (last-in-first-out) priority matching algorithm to match the trade and LOB data. For the case where the Reported LOB event may contain multiple reported events $j \in \{1, \dots, m_i\}$, on the same unique time stamp $t_{i,j}$, a trade takes on the closest forward LOB event time within certain time boundaries. If there are for example, two LOB events that occur at times $t_{i,j} = 10 : 30 : 01.100$ and $t_{i+1,j} = 10 : 30 : 01.115$, and multiple lines of trades occur between these two time points, the trades will all be reported at the last time $t_{i+1,j}$. We then assign event type of limit order, cancellation and market order to events as described below.

Figure 3.2 shows the performance of the matching process employed within our research. Across the assets we consider, the market depth events that can be matched with the trade data represents 95% of events. This method has a higher matching percentage compared with that employed by Toke (2016), which yields a matching of 92% of events on average. However, the datasets we use are vastly different to those used by Toke (2016), where the number of reported transactions of their sample of stocks listed on the Euronext, reaches a maximum of 12,000 transactions. The assets we consider are futures assets over various underlying instruments and they can yield up to 2.2 million events in a single day. It is unclear from the use of different instruments, which method is in fact superior.

A key difference between the method we employ and that of Toke (2016), is the retention of the remaining 5% of unmatched data, which is associated with trades. Whilst we cannot assign an event type to this data, we attached an indicator and retain this data as it still contains information about the LOB, despite the quality of data being lower.

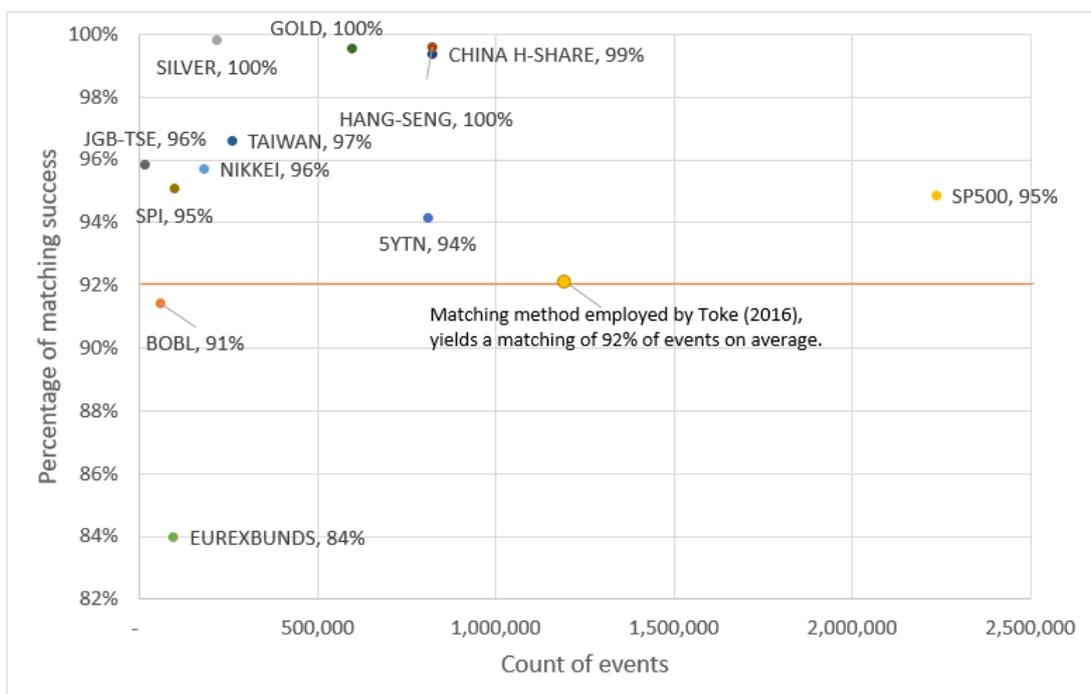


Figure 3.2: Percentage of matched trades, as a function of total number of reported events, across 10 trading days (July 2015), for each futures asset considered.

Special case: ‘Fill-down’ events. Occasionally many trades are reported at the same time stamp, but they cannot be matched to a corresponding change in the LOB, see Figure 3.3. An example of a fill-down event is when a number of smaller orders are traded against a single large order and this may lead to many trades reported on the same time stamp. The LOB may not be updated at each step. Toke (2016) suggests reported transactions may be executed against hidden liquidity, resulting in potentially a single update to the market, despite executing against more shares than displayed. Another explanation for this reporting, is the potential for ‘off the book’ trades that may not affect the quote file, but are reflected as trades in the trade file, for instance executions from a dark pool.

Rather than discarding this data, as done in Toke (2016) as failed matches, the proprietary matching algorithm we use takes a copy of the LOB, and this is replicated for each trade. A ‘copy’ of the LOB simply means that the LOB is replicated from the previous event and thus, does not show any changes from the previous time step. By comparison, for all other events that are not classified as ‘fill-down’ events, there is a change in the LOB from one event to the next (Figure 3.3). As mentioned in the previous section, this data is not discarded, rather it is retained and can be used within the model through the introduction of additional marks.

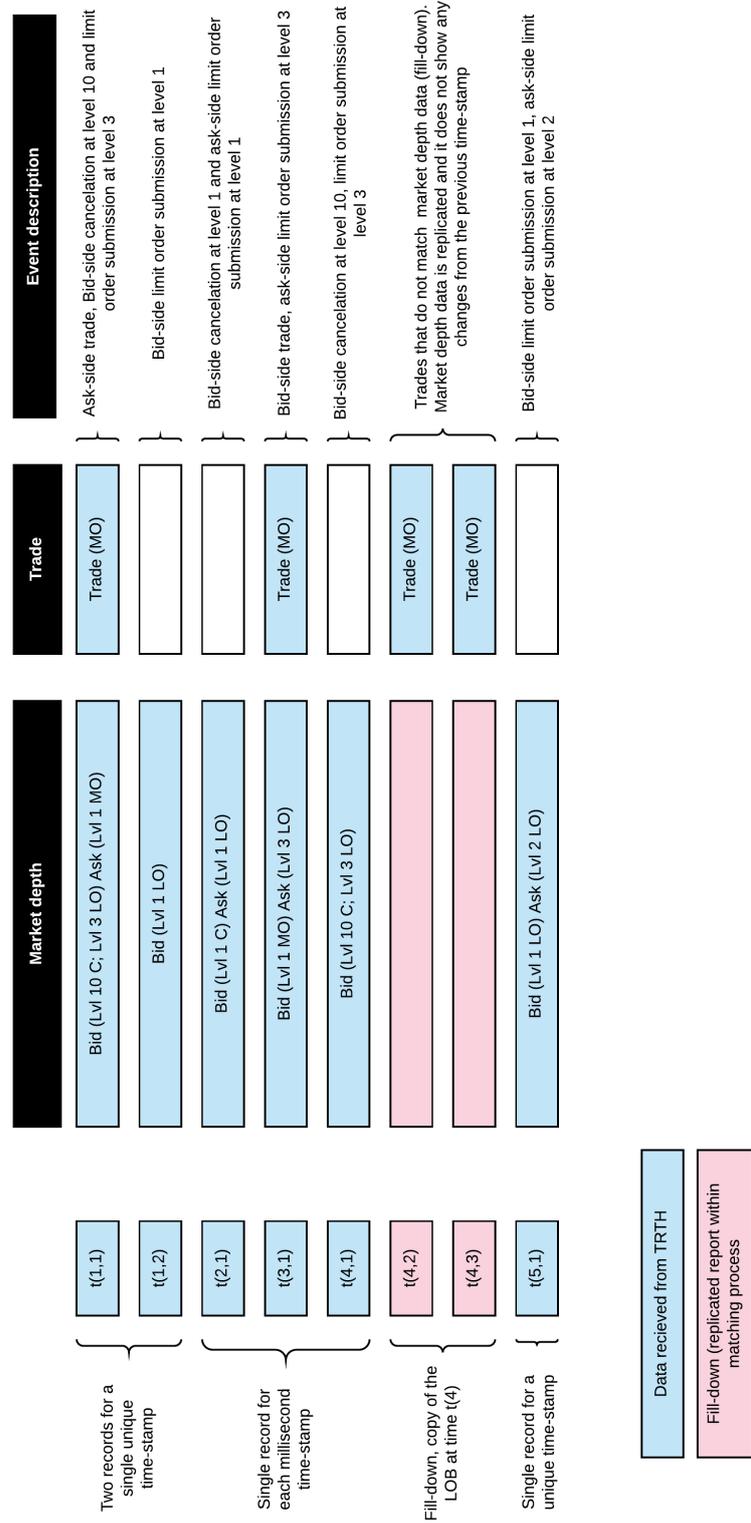


Figure 3.3: An example of the matched market depth and trade data, with an example of multiple records on a single time-stamp, and the case of fill-down, whereby the trade data does not match the market depth data.

For each event we denote the time-stamp as $t_{i,j} \in [0, T]$ and which takes the value 0, at $t = 0$, where $i \in \{1, \dots, n\}$ and $j \in \{1, \dots, m_i\}$. The j represents the events *within* a time stamp and thus, $t_{i=k,j=1} = t_{i=k,j=2} = \dots = t_{i=k,j=m_i}$. This means that one or more reported events j may occur on a single unique time-stamp t_i . Each level of the LOB is defined by a price $P_{i,j}^{(s,l)}$, on the side s , where B is the bid side and A is the ask side, at level l of the LOB. All submitted orders have an associated volume, and limit orders submitted at the same price level will be consolidated into a single volume $V_{i,j}^{(s,l)}$ for that price level. A trade, superscripted by ‘*’, is recorded with a price $P_{i,j}^*$ and volume $V_{i,j}^*$. A trade is a result of a market order transacting with a limit order. The use of ‘trade’ and ‘market order’ are used interchangeably.

Definition 5 (Matched LOB with unique time-stamps that contain one or more records per time-stamp). *A unique time stamp t_i , where $0 = t_0 < t_1 < t_2 < \dots < t_n \leq T$ is defined to be a time to the nearest millisecond, at which at least one event is observed in the Reported LOB. Multiple events $m_i \geq 1$, can be reported at t_i , so we represent the time of these events as t_{ij} , for $j = 1, \dots, m_i$. Note that the physical order of t_{ij} , recorded at the t_i millisecond, is $t_{i,1} = t_{i,2} = \dots = t_{i,m_i}$. Let $t_{i,j} \in \mathbb{R}$ be a unique time-stamp of events $i \in \{1, \dots, n\}$ and within each unique time stamp there may be multiple events $j \in \{1, \dots, m_i\}$. The Reported LOB is defined by the following components:*

- $V_{t_{i,j}}^{(s,l)}$ is the volume on the LOB at time $t_{i,j}$, for level $l \in \{1, \dots, d\}$ and bid or ask side $s \in \{B, A\}$;
- $P_{t_{i,j}}^{(s,l)}$ is the price on the LOB at time $t_{i,j}$, for level $l \in \{1, \dots, d\}$ and bid or ask side $s \in \{B, A\}$;
- $V_{t_{i,j}}^*$ is the volume of a trade at time $t_{i,j}$;
- $P_{t_{i,j}}^*$ is the price of a trade at time $t_{i,j}$;
- $D_{t_{i,j}} = \begin{cases} -1 & \text{if } P_{t_{i,j}}^* - P_{t_{i,j}}^{(B,1)} \leq P_{t_{i,j}}^{(A,1)} - P_{t_{i,j}}^* ; \\ +1 & \text{if } P_{t_{i,j}}^* - P_{t_{i,j}}^{(B,1)} > P_{t_{i,j}}^{(A,1)} - P_{t_{i,j}}^* . \end{cases}$
Trade direction is set to -1 if the traded price is closer to the best bid and +1 if the traded price is closer to the best ask.

Stage 2: Defining event types at time $t_{i,j}$

Due to the limitations of the data sets discussed above, we can only make an estimate of the types of events. Key to the definition of an event type is the net volume. As we will present below, we can define the net volume at each event time $t_{i,j}$ in (3.1). If a market order has originated on the best bid, yet the volume change from the previous event to the current event is greater than the market volume, the difference represents a both a submitted market order and a limit order. Conversely, if the volume change is less than the market volume, the difference represents a cancellation.

Algorithm 1 (Event classification $e \in \{LO, MO, C\}$ at time $t_{i,j}$). *The classification of event types follows.*

1. *The event is a market order (MO) if the volume of a trade $V_{t_{i,j}}^*$ and price of a trade $P_{t_{i,j}}^*$ is not empty. Further to the definition of a market order, we classify a market order more specifically to the origin of execution, either on the bid or ask side. If the trade direction is $D_{t_{i,j}} = -1$, then the market order originated at the bid and if $D_{t_{i,j}} = 1$, then the market order originated at the ask.*
2. *Net volume on the price level is the change in volume from the current event to the previous event, plus the volume of the market order at the current event, and defined as*

$$V_{t_{i,j}}^{net,s,l} = \begin{cases} V_{t_{i,j}}^{(s,l)} - V_{t_{i-1,j}}^{(s,l)} + V_{t_{i-1,j}}^* \mathbb{I}[D_{t_{i,j}} = -1], & \text{if } j = 1, s = B; \\ V_{t_{i,j}}^{(s,l)} - V_{t_{i-1,j}}^{(s,l)} + V_{t_{i-1,j}}^* \mathbb{I}[D_{t_{i,j}} = 1], & \text{if } j = 1, s = A; \\ V_{t_{i,j}}^{(s,l)} - V_{t_{i,j-1}}^{(s,l)} + V_{t_{i,j-1}}^* \mathbb{I}[D_{t_{i,j}} = -1], & \text{if } j \in \{2, \dots, m_i\}, s = B; \\ V_{t_{i,j}}^{(s,l)} - V_{t_{i,j-1}}^{(s,l)} + V_{t_{i,j-1}}^* \mathbb{I}[D_{t_{i,j}} = 1], & \text{if } j \in \{2, \dots, m_i\}, s = A. \end{cases} \quad (3.1)$$

3. *The event is a limit order (LO) if the net volume on the LOB is positive $V_{t_{i,j}}^{net,s,l} > 0$, for a specific level l and side s .*
4. *The event is a cancellation (C) if the net volume on the LOB is negative $V_{t_{i,j}}^{net,s,l} < 0$, for a specific level l and side s .*
5. *In the special case of ‘fill-down’, we cannot accurately determine the event type. A flag is attached to these specific events to identify them as ‘fill-down’ trades, triggering an exception in the event defining algorithm, thus not assigning limit orders or cancellations based on the prevailing LOB and the occurrence of the trade. We define an event to be a ‘fill-down’ if all of the following conditions are met on a single non-unique event:*

- *LOB change indicator = 0, which represents no limit orders or cancellations at that event;*
- *Trade indicator = 1, which represents a trade at that event;*
- *Time-stamp has more than one reported event = 1.*

6. *For each event, the prices at each level are compared with the prices at each level in the previous event. It is crucial to check that the prices at each level align, to ensure the correct volumes are being used in the net volume calculation. If $P_{t_{i,j}}^{(s,l)} \neq P_{t_{i-1,j}}^{(s,l)}$, then we search for the matching previous price before calculating the net volume.*

A practical example for [Step 6] in Algorithm 3.1.3, is a case where a trade has occurred at event time $t_{i,j}$ and the trade volume has consumed all of the volume on the best bid. The volume at price level 1 on the bid at time $t_{i,j}$, needs to then be compared with the volume on price level 2 on the bid at time $t_{i-1,j}$. If the volume at price level 1 at time

$t_{i,j}$ was instead compared with the volume on price level 1 at time $t_{i-1,j}$, an incorrect assignment of a limit order or cancellation would likely result.

Some examples of limitations of the method employed to define event types are listed below.

- *Amendments* will not be classified and they will be reflected as:
 - a reduction in the volume of an order will appear as a cancellation;
 - an increase in the volume of an order will appear as a limit order;
 - a price amendment will appear as a cancellation and then a new limit order submission.
- If a market order occurs, it is also possible for a limit order cancellation or submission on the price levels that were affected by the market order. This algorithm does not account for such occurrences.

Stage 3: Aggregating data such that each unique time stamp contains only one consolidated record per time stamp

As discussed in previous sections, the Hawkes process is a simple point process and this means that no two events can occur on the same time-stamp. To be able to model the data using a Hawkes process and to retain as much information within the data as possible, we aggregate the data according to Algorithms 2 and 3 below.

We consider events during *liquid market hours*, as described in Section 2.1.1. If we take the example of the NIKKEI on the trade date 21-July-2017, 308,676 events were reported. Considering liquid market hours, this reduces the number of events to 127,225. Aggregating the data so that all events have a single observed event for each unique time-stamp, reduces the total count of events to 53,253. This will of course reduce further when considering specific event types as the event process, such as bid side only and levels one to three, for example.

Algorithm 2 and 3 describe the process employed to aggregated the price and volume data, such that all time-stamps have a single reported event. It is worth noting, that the method for determining ‘trade direction’, defined in the Algorithm 2, was first presented by Hasbrouck (1988) to classify trades at the midpoint of the bid-ask spread. Alternatives to this were originally proposed by Lee and Ready (1991) and known as the Lee-Ready procedure. Toke (2016) provides a recent study of this procedure in relation to choices made about the matching process.

Algorithm 2 (Aggregating prices and volumes). *To aggregate the data, such that the time-stamps t_i have a single observed event, we create an index associated with the time-stamps with a single reported event, and copy the data directly from the LOB and market orders: $V_{t_i}^{(s,l)} = V_{t_{i,j=1}}^{(s,l)}$, $P_{t_i}^{(s,l)} = P_{t_{i,j=1}}^{(s,l)}$, $V_{t_i}^* = V_{t_{i,j=1}}^*$, $P_{t_i}^* = P_{t_{i,j=1}}^*$. LOB events, as defined in Definition 1, are also copied directly. Counts of events at that unique time stamp are set to one.*

For the case where $m_i > 1$, the following process is applied.

1. The final state of the LOB volume within the unique time-stamp t_i is

$$V_{t_i}^{(s,l)} = V_{t_i,j=m_i}^{(s,l)}. \quad (3.2)$$

2. The final state of the LOB price within the unique time-stamp t_i is

$$P_{t_i}^{(s,l)} = P_{t_i,j=m_i}^{(s,l)}. \quad (3.3)$$

3. The sum of the total traded volume at the unique time-stamp t_i is

$$V_{t_i}^* = \sum_{j=1}^{m_i} V_{t_i,j}^*. \quad (3.4)$$

4. The VWAP (volume weighted average price) at the unique time-stamp t_i is

$$P_{t_i}^* = \frac{1}{V_{t_i}^*} \times \sum_{j=1}^{m_i} \left(V_{t_i,j}^* \times P_{t_i,j}^* \right). \quad (3.5)$$

5. Trade direction is set to -1 if the VWAP price is closer to the best bid and +1 if the VWAP price is closer to the best ask, denoted by

$$D_{t_i} = \begin{cases} -1, & \text{if } P_{t_i}^* - P_{t_i}^{(B,1)} \leq P_{t_i}^{(A,1)} - P_{t_i}^*; \\ +1, & \text{if } P_{t_i}^* - P_{t_i}^{(B,1)} > P_{t_i}^{(A,1)} - P_{t_i}^*. \end{cases} \quad (3.6)$$

6. LOB events are summarized, such as indicators representing the event type at each level and side are recorded.

Algorithm 3 (Aggregating volumes for events at unique times t_i). To determine the volume of events $e \in \{LO\}$ and $e \in \{C\}$, we utilize $V_{t_i,j}^{net,s,l}$ in (3.1).

- The aggregated volume of event $e \in \{LO\}$ for a specified level l and side s is

$$V_{t_i}^{LO,s,l} = \sum_{j=1}^{m_i} V_{t_i,j}^{net,s,l} \mathbb{I}[V_{t_i,j}^{net,s,l} > 0] \quad (3.7)$$

- The aggregated volume of event $e \in \{C\}$ for a specified level l and side s is

$$V_{t_i}^{C,s,l} = \sum_{j=1}^{m_i} |V_{t_i,j}^{net,s,l}| \mathbb{I}[V_{t_i,j}^{net,s,l} < 0] \quad (3.8)$$

3.1.4 Assets and data attributes

We consider the LOB and trade data for 6 futures over equity, commodity and interest rate assets. The analysis is conducted on data collected during liquid market hours and from 01-Jan-2014 to 31-Jul-2015. The assets, markets and liquid market hours are presented in Table 3.4, utilizing descriptions from Section 2.1.1 (Richards et al., 2015).

Table 3.4: Asset description used in the analysis and modeling. Market hours refer to the liquid market hours in local trading time of the exchange.

Asset Name	Acronym	Liquid Market Hours (local time)	Exchange
Interest rate derivatives			
1. 5 Year T-Note	<i>5YTN</i>	7:30:00 to 14:00:00	CBOT
2. Euro-BOBL	<i>BOBL</i>	8:00:00 to 19:00:00	EUREX
Equity derivatives			
3. SIMEX Nikkei 225	<i>NIKKEI</i>	8:00:00 to 14:00:00	SGX
4. E-mini S&P 500	<i>SP500</i>	08:30:00 to 15:00:00	CME
Precious Metals			
5. Gold	<i>GOLD</i>	06:30:00 to 13:30:00	COMEX
5. Silver	<i>SILVER</i>	08:30:00 to 13:00:00	COMEX

Table 3.5 shows the mean percentage of events per day relative to the total number of events before the matched data is aggregated.

Table 3.5: *Matched data*. Mean values of counts of events divide by total events. The tick range within bid and ask, not including the spread, across 10 trading days (July 2015) for each futures asset considered.

Asset	LOB/ Events	Trade/ Events	Non-unique/ Events	Fill-down/ Events	Events	Tick range
5YTN	94.17%	8.34%	82.64%	5.83%	808575.0	{0.0078, 0.0467}
BOBL	89.94%	16.50%	11.13%	8.56%	59631.7	{0.0100, 0.0100}
GOLD	98.89%	4.96%	73.33%	0.45%	596337.7	{0.0100, 4.3000}
SILVER	99.39%	3.81%	69.62%	0.18%	217748.5	{0.0050, 0.3350}
SP500	94.87%	9.11%	91.81%	5.13%	2235120.6	{0.2500, 1.7500}
NIKKEI	95.71%	8.16%	59.86%	4.29%	180041.5	{5.0000, 15.0000}

Table 3.6 demonstrates the mean percentage of the aggregated data, whereby there are one record per time-stamp and the events per day relative to the total number of events.

Table 3.6: *Aggregated data*. Mean values of counts of events divide by total events. The spreads between the best bid and ask across a 10 trading days (July 2015) for each futures asset considered.

Asset	LOB / Events	Trade / Events	Events	Min. spread	Max. spread
5YTN	100.00%	7.36%	138962.5	0.00780	0.0078
BOBL	98.30%	8.88%	52929.4	0.0100	0.0100
GOLD	99.74%	14.09%	158822.9	0.1000	3.0000
SILVER	99.93%	9.97%	65325.0	0.0050	0.0400
SP500	100.00%	26.27%	181843.2	0.2500	2.0000
NIKKEI	100.00%	7.46%	67264.0	5.0000	15.0000

3.2 Event processes

The events times t_i , are irregularly spaced unique times within the aggregated data sets, where $t_i \geq \text{millisecond}$. Events occur when there is one or more changes in the LOB due to a limit order submission, cancellation, a market order (trade), or any combination of the above.

Definition 6 (LOB irregularly spaced unique time-stamps within the data sets of aggregated reported events). *For $i = 1, \dots, n$, the event times are the unique and strictly*

increasing millisecond time stamps t_i , defined above in constructing the Observed LOB for $0 = t_0 < t_1 < t_2 < \dots < t_n \leq T$ event times within a day and where $t_i \geq 1$ millisecond.

The combinations of event types and levels that will be consider for the most appropriate choice of event process are highly dependent on the hypothesis explored. For example, if we were to undertake a study on the impact various marks such as spread and event type volume has on the intensity of LOs and opposite side MOs, an appropriate choice of event process may include events times based on $e = LO$, $l \in \{1, \dots, 5\}$ and for each $s \in \{B, A\}$. In addition, $e = MO$, which may be studied for each $s \in \{B, A\}$. These event processes could be modelled in a the form of a univariate Hawkes process, with a model for each side.

3.2.1 Empirical studies of event processes

The studies that follow, address important questions about the appropriate type of Hawkes self-exciting point process that can be used to model the Observed LOB data. It should be noted, that we use Hawkes self-exciting point process and Hawkes process interchangeably throughout the remainder of this thesis. We consider the proportion of simultaneous events on the bid and ask side and how this might inform the selection of a univariate Hawkes process, rather than a multivariate Hawkes process, so that key assumptions of the model are not violated. We study the impact of including additional levels from the LOB, inspecting whether the additional information from the outer levels of the LOB is informative or contributing noise the to the process. Finally we consider the decay of the counts, after a fluctuation in event arrivals. The choices made on model type, level inclusion, event type inclusion and guidance for decay function type on high frequency event arrivals will have a direct impact on the chapters that follow, in defining the marks and the fitting of a Hawkes process with vector-valued marks.

Proportion of simultaneous events on bid and ask side

Recall that event times are a sequence of non-negative random variables, such that $\forall i \in \mathbb{N}, t_i < t_{i+1}$. If there are simultaneous event times on the bid and ask, that is, events occurring at the same millisecond on the bid and ask side, then these cannot be modelled as components within the multivariate Hawkes process setting. We begin this empirical study by investigating the frequency of simultaneous events and the occurrence of these simultaneous events based on our choice of event types $e \in \{LO, MO, C\}$, levels $l \in \{1, \dots, 10\}$, $l \in \{1, \dots, 5\}$, $l \in \{1, \dots, 3\}$, $l = 1$ and for each $s \in \{B, A\}$. Findings from this will be crucial in the important decisions of the type of Hawkes process that one can implement.

Tables 3.7, 3.8, 3.9, 3.10, 3.11 and 3.12 show the percentage of simultaneous events on the bid and ask, specifically the overlapping event times for various combinations of event types and levels. The left hand side of the table considers all event types LO, MO and C for levels 1:10, levels 1:5, levels 1:3 and level 1. The right hand side of the table excludes cancellations, C from the datasets. Even with the removal of cancellations

and considering price level 1 data only, almost all assets considered have greater than 5%, on average, overlapping events, with the SP500 demonstrating a sizable 33% overlap. Considering all event types and 5 price levels of data, there is between 20% and 40% of overlapping events, with SP500 presenting 74% of overlapping events. With these findings, it is clear that modelling this process as a bivariate multivariate Hawkes process, whereby the sides represent components is not possible. The most appropriate structure to consider would be the univariate Hawkes process for each side of the LOB.

Table 3.7: The percentage of simultaneous events occurring on the bid and ask side, and the mean number of events on the bid and ask, for various combinations of event types and levels, across 10 trading days (July 2015), for 5YTN.

	Event types: LO, MO, C				Event types: LO, MO			
	Lvl 1:10	Lvl 1:5	Lvl 1:3	Lvl 1	Lvl 1:10	Lvl 1:5	Lvl 1:3	Lvl 1
Mean	31.37%	28.82%	26.65%	21.63%	21.95%	17.01%	14.05%	8.89%
Median	31.09%	28.31%	26.33%	21.23%	21.84%	17.13%	13.72%	8.71%
Std	1.38%	1.34%	1.55%	1.35%	1.16%	1.15%	1.27%	0.90%
Bid	90808	84390	77982	63788	73161	63938	56843	43906
Ask	91560	85894	80677	63095	73262	64684	58227	42290

Table 3.8: The percentage of simultaneous events occurring on the bid and ask side, and the mean number of events on the bid and ask, for various combinations of event types and levels, across 10 trading days (July 2015), for BOBL.

	Event types: LO, MO, C				Event types: LO, MO			
	Lvl 1:10	Lvl 1:5	Lvl 1:3	Lvl 1	Lvl 1:10	Lvl 1:5	Lvl 1:3	Lvl 1
Mean	30.47%	28.95%	27.30%	21.52%	23.04%	19.19%	13.88%	5.66%
Median	30.05%	28.63%	26.85%	21.13%	22.81%	19.04%	13.74%	5.50%
Std	1.87%	1.92%	1.84%	1.52%	1.36%	1.52%	1.03%	0.57%
Bid	34174	32906	31286	25896	27671	25316	22331	16483
Ask	33463	32360	31107	26207	26603	24433	21768	16167

Table 3.9: The percentage of simultaneous events occurring on the bid and ask side, and the mean number of events on the bid and ask, for various combinations of event types and levels, across 10 trading days (July 2015), for GOLD.

	Event types: LO, MO, C				Event types: LO, MO			
	Lvl 1:10	Lvl 1:5	Lvl 1:3	Lvl 1	Lvl 1:10	Lvl 1:5	Lvl 1:3	Lvl 1
Mean	44.03%	38.16%	35.26%	23.61%	28.17%	20.95%	18.14%	10.03%
Median	44.18%	37.48%	34.64%	23.42%	28.25%	20.93%	17.83%	9.71%
Std	1.36%	1.92%	2.20%	1.30%	1.54%	1.25%	1.19%	1.24%
Bid	114868	102238	94620	69003	89763	75935	68757	47306
Ask	112503	99510	93142	68595	86946	73269	67505	46424

Table 3.10: The percentage of simultaneous events occurring on the bid and ask side, and the mean number of events on the bid and ask, for various combinations of event types and levels, across 10 trading days (July 2015), for SILVER.

	Event types: LO, MO, C				Event types: LO, MO			
	Lvl 1:10	Lvl 1:5	Lvl 1:3	Lvl 1	Lvl 1:10	Lvl 1:5	Lvl 1:3	Lvl 1
Mean	37.49%	34.71%	30.94%	21.93%	22.52%	17.75%	14.37%	9.23%
Median	37.42%	34.56%	30.99%	22.01%	22.42%	17.47%	14.53%	9.16%
Std	2.13%	2.08%	1.79%	1.63%	1.91%	1.63%	1.27%	0.74%
Bid	44963	42571	39470	29791	33325	29673	26410	18818
Ask	44682	42294	39280	30172	34044	30508	27393	20432

Table 3.11: The percentage of simultaneous events occurring on the bid and ask side, and the mean number of events on the bid and ask, for various combinations of event types and levels, across 10 trading days (July 2015), for SP500.

	Event types: LO, MO, C				Event types: LO, MO			
	Lvl 1:10	Lvl 1:5	Lvl 1:3	Lvl 1	Lvl 1:10	Lvl 1:5	Lvl 1:3	Lvl 1
Mean	74.23%	71.22%	68.47%	59.13%	56.0%	49.28%	44.06%	33.83%
Median	74.84%	71.83%	69.02%	59.57%	56.78%	49.78%	44.38%	34.12%
Std	2.18%	2.21%	2.13%	2.18%	2.51%	2.46%	2.05%	1.97%
Bid	158761	154747	151024	137840	137012	128968	122383	106612
Ask	158156	154206	150406	136740	136025	127568	120874	105545

Table 3.12: The percentage of simultaneous events occurring on the bid and ask side, and the mean number of events on the bid and ask, for various combinations of event types and levels, across 10 trading days (July 2015), for NIKKEI.

	Event types: LO, MO, C				Event types: LO, MO			
	Lvl 1:10	Lvl 1:5	Lvl 1:3	Lvl 1	Lvl 1:10	Lvl 1:5	Lvl 1:3	Lvl 1
Mean	20.31%	19.33%	18.04%	11.41%	11.11%	9.45%	7.17%	4.05%
Median	19.95%	19.17%	17.84%	11.21%	10.98%	9.47%	7.16%	3.89%
Std	3.16%	3.04%	2.73%	1.41%	2.24%	1.97%	1.30%	0.55%
Bid	40617	39049	36256	22937	26425	24844	22352	14556
Ask	40556	39006	36105	22428	26322	24754	22518	14034

Study of the levels and event type counts

The study that follows will explore the implications of the various combinations of event types, and levels for each side of the LOB. This will help inform the number of levels that should be included in the event process. We study in detail two assets, SILVER and NIKKEI, with the choice being motivated by their vastly different features, both being highly liquid and with differing underlying assets, being the silver commodity and the Nikkei-225 equity index, respectively.

We first consider the asset **SILVER** across 10 trading days. Table 3.13 presents the mean summary statistics for counts of events within evenly spaced, one minute time buckets for each level across the trading days considered. What is apparent, is the significant drop in counts of events in the outer levels of the LOB. On average we can see level 1 has 160 events occurring within a minute, compared to the 2 events per minute of level 10 of the LOB. However, we do see a reasonable proportion of events arising in the depth of the LOB, but restricted to the higher levels. It is therefore reasonable to expect that the event arrivals within the higher levels of the depth of the LOB will contribute to informing the intensity process of the LOB, rather than just the best bid and ask, which is commonly modelled in financial literature.

Table 3.13: The mean summary statistics of counts of events within evenly spaced time intervals of one minute, occurring on the bid and ask side, event types, LO, MO, C, across each level, for 10 trading days (July 2015), for SILVER.

	Lvl 1	Lvl 2	Lvl 3	Lvl 4	Lvl 5	Lvl 6	Lvl 7	Lvl 8	Lvl 9	Lvl 10
Mean	160.04	31.11	14.13	8.83	3.96	3.12	2.33	1.19	1.48	2.25
Median	128.65	26.05	11.70	7.10	3.00	2.40	1.55	0.60	0.80	1.40
Std	123.911	21.08	9.99	7.09	3.73	3.07	2.56	1.91	1.93	2.76
Min	12.00	0.70	0.30	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Max	1009.1	139.5	60.4	46.6	25.6	19.6	17.0	15.2	11.9	19.8

The distribution of events within the evenly spaced time buckets on each level, will help determine whether the outer levels, which have a significantly lower frequency of counts, are contributing additional information to the intensity process. For example, if level 1, which represents over 3/4 of all counts, is highly correlated with the increase in counts of events at the outer levels, then the inclusion of the outer levels may not be contributing additional meaningful information into the process.

Figure 3.4 displays the plots of counts across the evenly spaced time bins. We can see that for all levels, the spike in counts tends to occur at similar times for all 10 levels. This is further supported with high correlations of counts for each level in the matrix of correlations in Figure 3.5. For example, the correlation of level 1 with all other levels, ranges from $[0.47, 0.84]$. It is clear that both the general shape of the count of events does not vary greatly from one level to the next and the contribution of events drops significantly from price level 5 onwards.

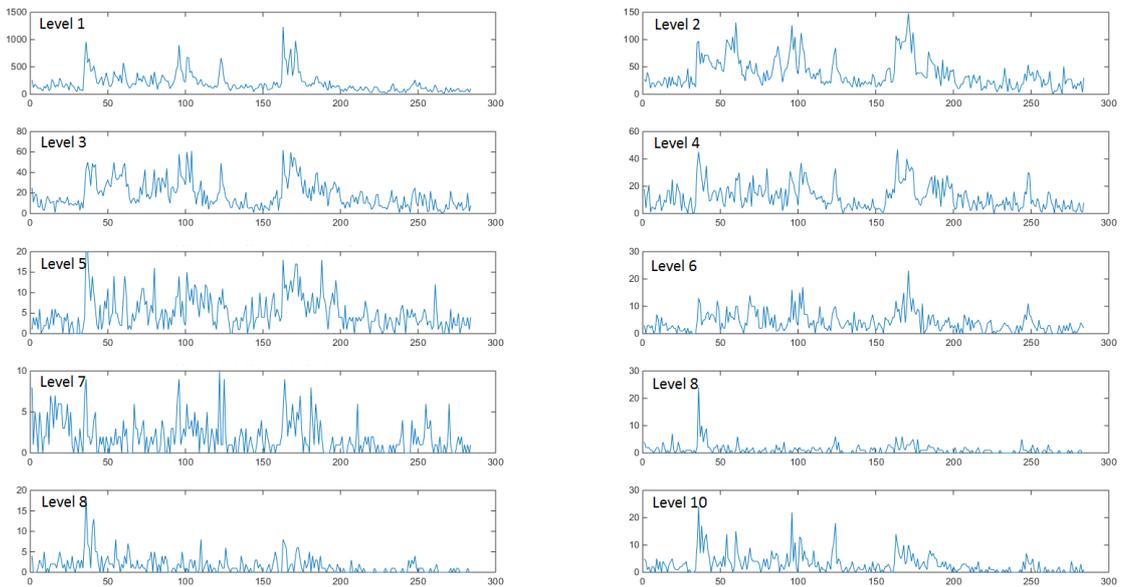


Figure 3.4: Count of events on the bid and ask side, LO, MO, C, for SILVER within 1 minute time bins, for the trading date 20-July-2015.

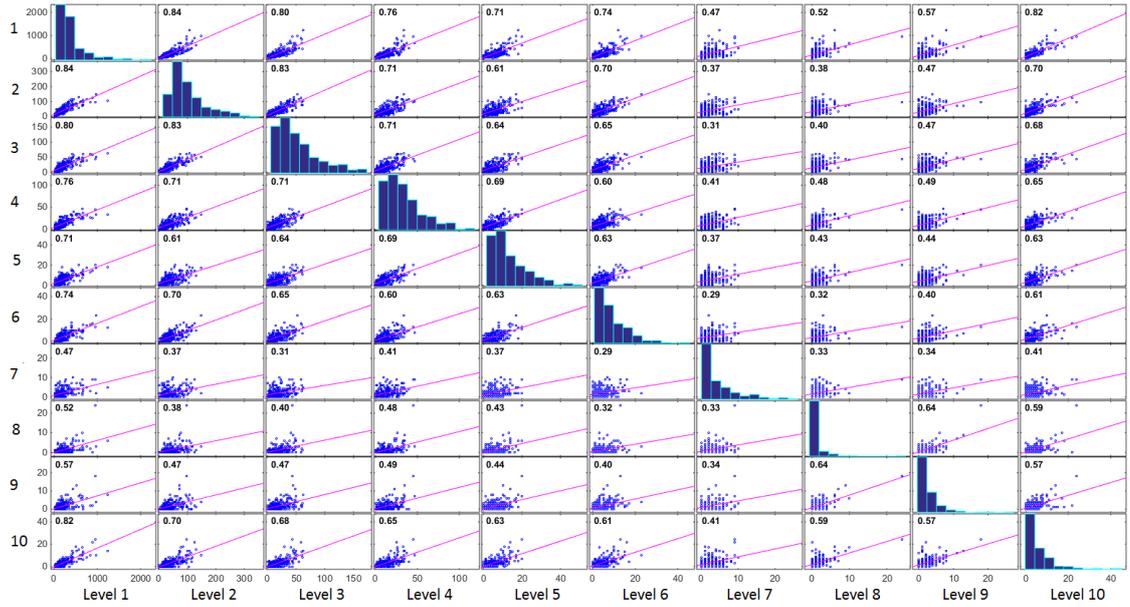


Figure 3.5: Correlation matrix of the count of events in each level 1:10, on the bid and ask side, LO, MO, C, for SILVER, within 1 minute time bins, for the trading date 20-July-2015.

We extend this study further to consider an additional futures asset, **NIKKEI** to assess whether the findings are asset dependent. Whilst the count of events in one minute time buckets is lower than what we've observed for the asset SILVER, similar patterns emerge in Table 3.14. The majority of events occur on the first few levels of the LOB, with a significant decrease in counts in the outer levels.

Table 3.14: The mean summary statistics of counts of events within evenly spaced time intervals of one minute, occurring on the bid and ask side, event types, LO, MO, C, across each level, for 10 trading days (July 2015), for NIKKEI.

	Lvl 1	Lvl 2	Lvl 3	Lvl 4	Lvl 5	Lvl 6	Lvl 7	Lvl 8	Lvl 9	Lvl 10
Mean	108.36	48.79	15.72	9.81	3.78	2.45	1.42	1.25	1.36	0.64
Median	93.30	40.70	8.60	8.20	2.60	1.60	0.70	0.50	0.70	0.10
Std	74.78	38.00	18.23	8.29	3.81	2.82	1.98	1.65	1.71	1.34
Min	4.30	0.80	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Max	478.90	175.70	103.10	55.50	26.4	19.6	15.70	10.70	11.80	14.30

Figure 3.6 shows the counts of all event types LO, MO, C for the bid and ask side within one minute time bins for each level. Similarly to SILVER, the general shape of the count of events remains similar, however as shown in Figure 3.6, this similarity across the levels is not as pronounced as we observed in SILVER. Level 2 events show a much more pronounced drop in events during the lunchtime period, when the underlying equity market is closed. However, this lunchtime effect is far less dramatic on the top level 1, compared with levels below level 2. Figure 3.7 displays the correlation and the contribution of events drops significantly from price level 5 onwards. Figure 3.7 reflects this observation, with a substantially lower correlation between levels, compared to what we observed in SILVER.

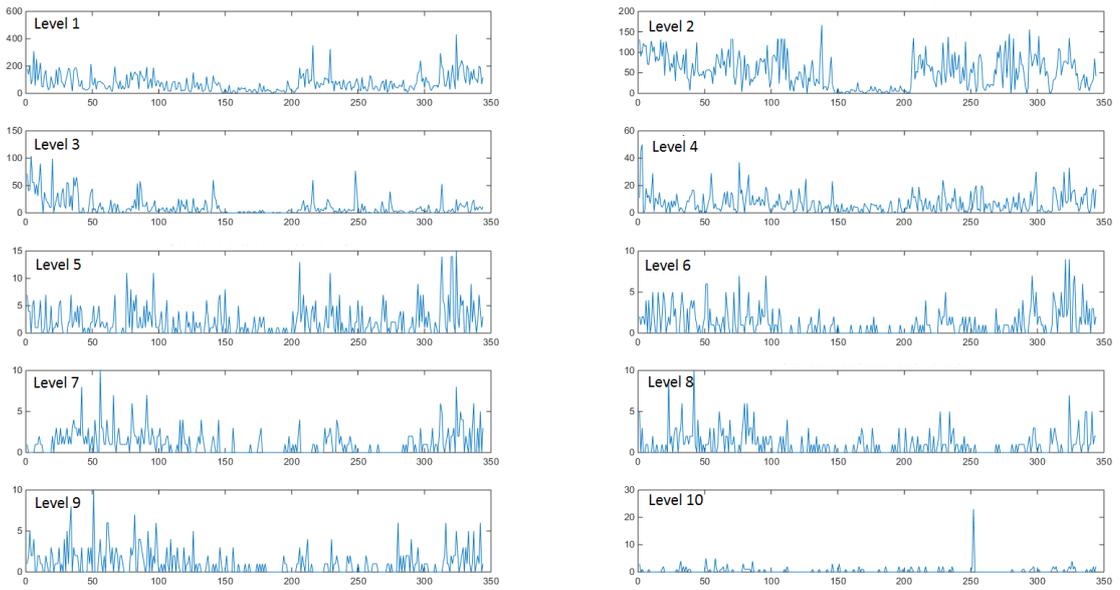


Figure 3.6: Count of events on the bid and ask side, LO, MO, C, for NIKKEI within 1 minute time bins, for the trading date 20-July-2015.

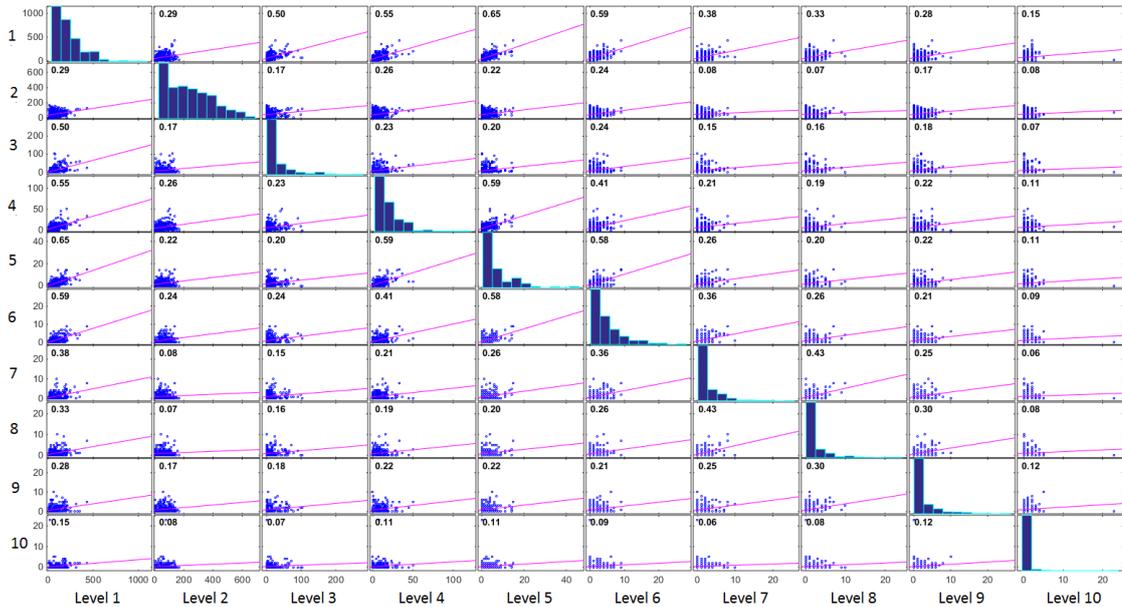


Figure 3.7: Correlation matrix of the count of events in each level 1:10, on the bid and ask side, LO, MO, C, for NIKKEI, within 1 minute time bins, for the trading date 20-July-2015.

Based on this preliminary analysis and taking a conservative approach of retaining maximal information, contribution of levels 1:5 provides a significant amount of information to inform the intensity process. The inclusion of deeper levels of the LOB beyond level 5 will not contribute meaningful information.

Speed of decay

As will be discussed in Chapter 4, there are multiple choices for the decay function of a Hawkes process, with the most common being the exponential decay function. An alternative that is used with far less frequency in financial literature, and has not yet been studied in detail within the financial application of Hawkes process with multivariate marks, is the power-law decay function. The power-law decay function was originally used in ETAS modelling for earthquake modelling in Ogata (1988) and is often referred to as Omori's law.

An advantage of the Hawkes process with an exponential decay function, which is a Markov process, is that the algorithm for calculating the intensity process is more efficient than that used by a Hawkes process with a power-law decay function. The non-recursive process required for estimation of a Hawkes process with a power-law decay is very computationally expensive for the vast data sets of the LOB data. As discussed earlier, Filimonov and Sornette (2015) presented a careful study on factors such as edge effects which impact simulation of long memory processes, creating uncertainty around research outcomes of work using a power-law decay function for high frequency data. This is a major consideration when using the power-law decay function on LOB data sets, given the size of the data and computational challenges it presents.

It is not enough to simply take guidance on appropriate decay functions from current literature, with no research to-date applying the Hawkes process to LOB data with 1:5 levels. Toke and Pomponio (2012) and Stindl and Chen (2018) studied a subset of market orders, utilizing information from the depth of the LOB down to level 2, to subset the market orders. Market orders only reflect a very small percentage of the intra-day data we consider and in addition, the subset of market orders determined via the LOB level 2 data, is an entirely different proposition to modelling LOB arrivals.

Very few papers consider the use of exponential versus a power-law decay function and those that do, utilize data that is much lower frequency than what we consider. Bacry and Muzy (2014) proposed a non-parametric technique for 'big' data sets to capture jointly, the mid-price upward and downward jumps and market order buys and sells, with kernels demonstrating a slow decrease, which are well described by power-law behaviour. As we will present in Table 5.2, market orders only constitute approximately 8% of the events within the event process that we consider. Whilst their data sets are large, they do not reflect the substantial size of the LOB data we utilize in this study, nor the frequency of event times considered.

Further to this, research by Filimonov and Sornette (2012, 2015), Hardiman et al. (2013) and Hardiman and Bouchaud (2014) debated the use of a power-law decay as represented by a sum of exponential functions, versus an exponential decay function, and with conflicting results about the most appropriate decay function. Their research was conducted on equity futures data across many years, and the event process consisted of mid-price changes. The intra-day frequency of mid-price changes is only a tiny fraction of the events that would be considered from an event process consisting of all event arrivals for the LOB with a depth of 5 levels. Finally, Lallouache and Challet (2016) proposed a

power-law decay, also represented by a sum of exponentials, as the appropriate choice for a Hawkes process applied to financial data. However, the data within their study was of hourly time frequency.

To gain an understanding of the rate of decay of the counting process, we create an evenly spaced time series of counts of events C_{t_z} , for each trading day, where $t_z = [z, 2z, \dots, nz]$ is the evenly spaced time, z is the time bin size $z \in \{1, 2, \dots, 360\}$ seconds (or 1 second through to 6 minutes), and n is the number of time bins of size z in a day. The events include all event types LO, MO, C for levels 1:5 for SILVER and NIKKEI. We consider increasing time bins sizes for two reasons, firstly we hypothesize that at very high frequencies, the decay rate of spikes in counts of events (as a proxy for intensity) decays far more rapidly than at lower frequencies. Secondly, if we find support for this hypothesis, then the limited research that has proposed a power-law decay function for event arrivals into the LOB, may not be relevant to this research given the lower frequency and different statistical properties that the data they've used presents.

We calculate the mean of the counts $\mathbb{E}[C_{t_z}]$, for a trading day, with the day sliced into time bins of size z . For each day we evaluate the time it takes for a sequence of exceedence (spike in the count of events) $C_{t_z} > \mathbb{E}[C_{t_z}]$, starting at $C_{t_z} \mathbb{I}[C_{t_z} > \mathbb{E}[C_{t_z}]] \mathbb{I}[C_{t_z-1} < \mathbb{E}[C_{t_z}]] \neq 0$, to revert back to the mean or lower counts. When the counts drop below the mean, we continue the search for the next 'spike' in counts and repeat the process until we have a sequence of times until reversion or durations r_z , where $z < r < nz$ for each time bin size z .

This forms an estimate of the inter-arrival times from a Hawkes process. If the duration is short, this is indicative of a rapid decay, which would be akin to the exponential decay of a Hawkes process, however if the durations are long, then it may be indicative of a power-law decay. Formal tests would involve parameter estimation via the full joint likelihood of both models to the event process data, and assessing the goodness-of-fit. However for reasons outlined in the remainder of this research, that is a non-trivial task, especially with the incorporation of marks. Thus, we aim to use empirical measures to guide the appropriate choice of decay function.

Figures 3.8 and 3.9 present the mean reversion times $\mathbb{E}[r_z]$, where $z = 1$ second, after a spike in counts of events above the mean counts across 10 trading days for SILVER and NIKKEI, respectively. Both assets present similar reversion durations, returning to mean levels or below within 1.7 seconds following a spike. The increase in activity does not continue beyond 40 seconds across the two assets and 10 trading days considered.

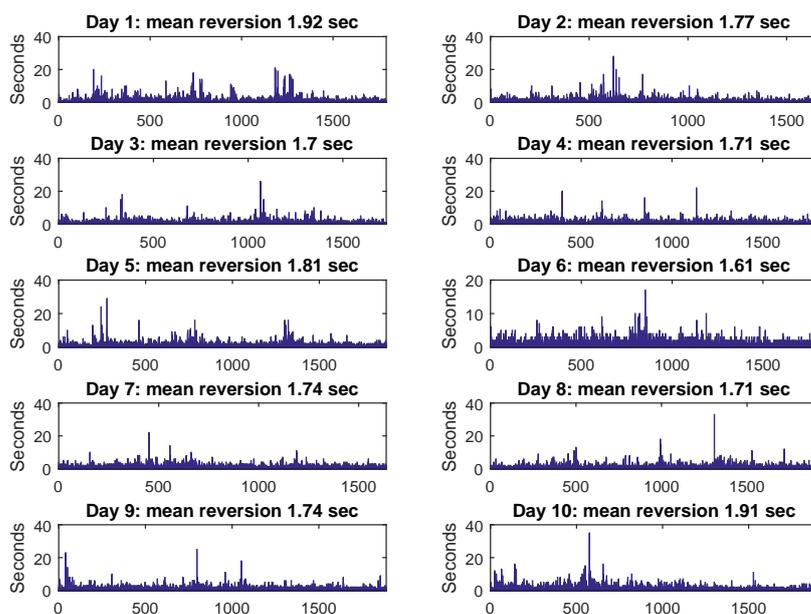


Figure 3.8: Sequence of reversion times r_z , when the count of events in a $z = 1$ second time bin have spiked above the mean level. We consider 10 trading days (20-31 July 2015), for SILVER, bid side only, 5 levels and for events, LO, MO, C.

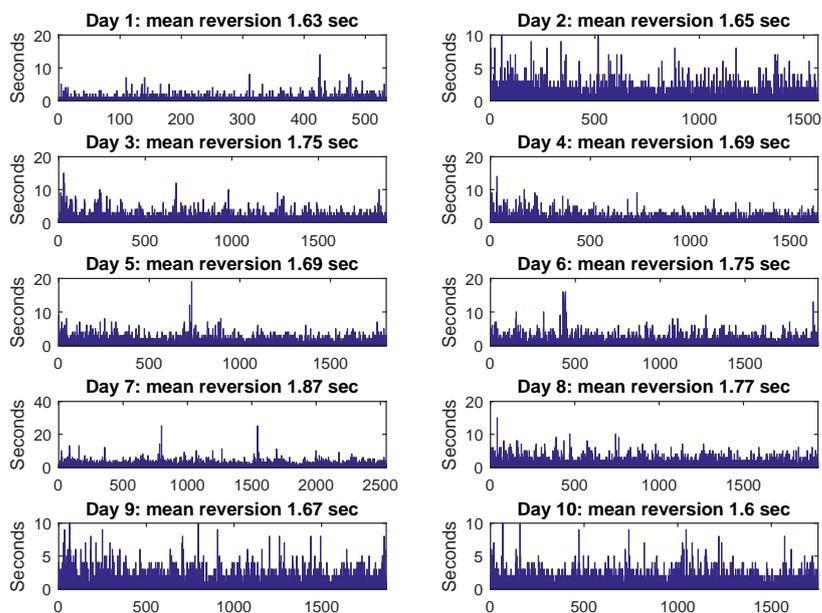


Figure 3.9: Sequence of reversion times r_z , when the count of events in a $z = 1$ second time bin have spiked above the mean level. We consider 10 trading days (20-31 July 2015), for SILVER, bid side only, 5 levels and for events, LO, MO, C.

We now study the impact that lower frequency observations has on the evaluated reversion time. We study time bin sizes $z \in \{1, 2, \dots, 360\}$ seconds and define the mean reversion time factor as $\mathbb{E}[r_z]/z$.

Figure 3.10 presents the moving average of the mean reversion time factors across 10 trading days for SILVER and NIKKEI, with increasing time bins. If the count data is observed every 6 minutes, the reversion back to the mean levels is a factor of $\mathbb{E}[r_z]/z \approx 3$ times the observed time, hence it will take 18 minutes before the counts revert back to, or below their mean level. Now of course, we cannot observe smaller than a 6 minute time window for reversion in this case, however if the reversion is indifferent to the size of the time bin, we would expect that the time reversion factor to be similar across time bins. As we observed in Figures 3.8 and 3.9 the 1 second time bin is much lower, with $\mathbb{E}[r_z]/z \approx 1.7$.

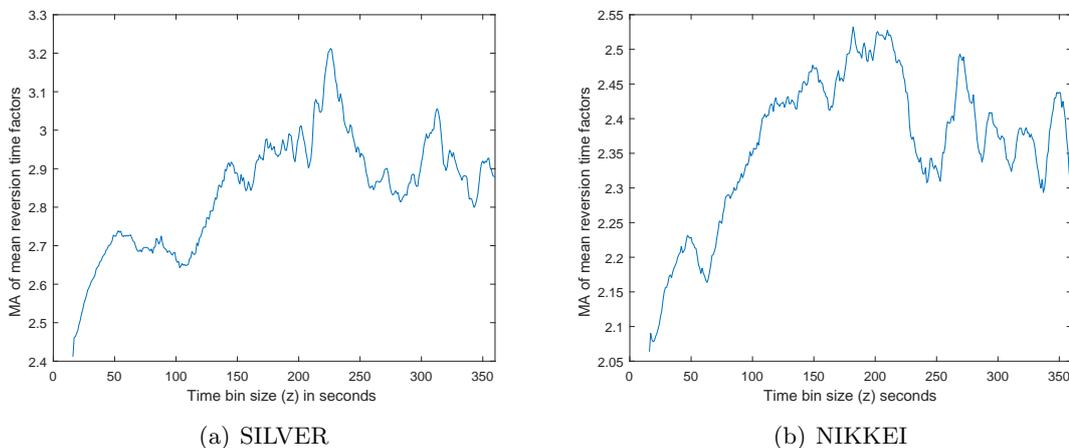


Figure 3.10: Moving average of a sequence of mean reversion time factors $\mathbb{E}[r_z]/z$, for increasing time bins, $z \in \{1, 2, \dots, 360\}$ seconds. We consider 10 trading days (20-31 July 2015), for SILVER and NIKKEI, bid side only, 5 levels and for events, LO, MO, C.

From the study of counts and mean reversion times across many time bin sizes, the reversion of counts to their mean level is fast for high frequency intervals of one second. There is a clear trend of slower decay as the time bins increase, when measured as a factor of the time bin size. Considering that the event times, which will be modelled by the Hawkes process, have a minimum time granularity of one millisecond, this empirical study provides support for an exponential decay function as the most appropriate. In addition, the added challenges of simulating and estimating a power-law decay function for ultra high frequency data add additional weight to this decision.

3.3 Conclusion

Within this chapter we provided a comprehensive description of the vast data sets that we intend to study within this research. We define the Physical LOB as a continuous time construction of prices and volumes. We then proceed to present the Reported LOB, describing the complexity of the data sets and the limitations and challenges with modelling

such a dataset.

This chapter has outlined the data processing steps, summarized in Figure 3.1 that are required to achieve a dataset that is both suitable for modelling, whilst retaining the majority of the important information about the event process and the state of the Physical LOB. The steps we outline include the non-trivial process of matching trade data with market depth data, defining the event types which are critical in the construction of marks and the final data aggregation to ensure that the data we model, fits within the assumptions of a simple point process, whereby no two events can occur at the same event time, t_i . The most comprehensive discussion found on the matching process, is that by Toke (2011). Whilst the datasets that we use differ to those used by Toke (2011), we achieve a higher matching percentage of 95% across all assets.

We formally define the event process in Definition 6. We consider the crucial question of the appropriateness of a multivariate Hawkes process versus a univariate Hawkes process with vector-valued marks for the modelling of LOB datasets. To guide this selection, we present an empirical study of the event process for a selection of futures assets. We initially consider the proportion of simultaneous events on the bid and ask side. We show the importance of constructing the Observed LOB from the Reported LOB, by demonstrating the simultaneous event times that exist on the bid and ask side. Even in the simple setting of a bivariate Hawkes process, where the model contains the bid and ask side of just level 1 data (best bid and best ask), we can see that the assumptions of non-simultaneous event times does not hold. Thus we conclude that the most appropriate model to consider is the univariate Hawkes process for each side of the LOB, with the detailed information about the LOB captured via vector-valued marks.

For both assets, SILVER and NIKKEI, we found that the majority of events reside on the upper levels of the LOB. Studying additional levels beyond the best bid and ask is informative to the intensity process, however those beyond level 5 are sparse and not contributing meaningfully to the process. We conclude that a suitable number of levels to model are levels 1:5 of the Observed LOB. Different combinations of event types were also studied, however the primary driver for this selection is the financial application in later chapters, where we consider the impact of all events in the LOB on the intensity process.

Finally, we consider an appropriate decay function with current literature providing very little guidance on the choice for ultra high frequency data. Upon inspection of reversion times of counts on evenly spaced time intervals, we note a fast decay for small time bins of one second. In addition, the computational challenges of fitting a Hawkes process with a power-law decay function to the large LOB datasets is non-trivial. For these reasons, we proceed with modelling the LOB data with a univariate Hawkes process with multivariate marks and an exponential decay function.

Chapter 4

Defining the Hawkes process

4.1 Defining a univariate Hawkes self-exciting marked point process

This section outlines key definitions required to specify the marked Hawkes self-exciting point process, using copula models for the joint distribution of the marks. We also define the residual process used to evaluate the goodness-of-fit of the Hawkes process.

We consider a univariate stationary Hawkes process, N_g observed over the interval $t \in [0, T]$, which takes the value 0 at $t = 0$. The observed points of this process are $\{(t_i, \mathbf{X}_i), i = 1, \dots, n\}$ $N_g \in \mathbb{N} \times \mathbb{X}$ with event times $0 = t_0 < t_1 < t_2 < \dots < t_n \leq T$ and vector of d marks $\mathbf{X}_i \in \mathbb{X} \subset \mathbb{R}^d$. The marked Hawkes process was originally introduced by Ogata (1988) as a model of earthquake occurrences, with the inclusion of the magnitude of the earthquake as a mark. In this research we consider the further generalization presented in Embrechts et al. (2011), which incorporates multivariate marks distributions with joint dependence. The scope of this research extends to the study of alternative boost functions that are additive and multiplicative functions of the marks, a broad range of discrete and continuous marks distributions and a range of copula models.

Embrechts et al. (2011) define two types of marked Hawkes process, a multivariate process with scalar valued marks impacting d intensity processes, and, a univariate point process with vector-valued marks. It is this latter process we consider in this thesis. Embrechts et al. (2011) and Liniger (2009) define the marked Hawkes process, N_g , on $t \in \mathbb{R}$ with intensity process given by

$$\lambda_g^\infty(t; \theta, \phi, \psi) = \eta + \vartheta \int_{(-\infty, t) \times \mathbb{X}} w(t-s; \alpha) g(\mathbf{x}; \phi, \psi) N_g(ds \times d\mathbf{x}), \quad t \in \mathbb{R}, \quad (4.1)$$

where $w : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ is a non-negative decay function satisfying

$$\int_0^\infty w(s; \alpha) ds = 1, \quad \int_0^\infty s w(s; \alpha) ds < \infty. \quad (4.2)$$

The cumulative decay function is defined as

$$W(t; \alpha) = \int_0^t w(s; \alpha) ds, \quad t \in \mathbb{R}_+. \quad (4.3)$$

The immigration rate is $\eta > 0$, the branching coefficient is $\vartheta > 0$ and the decay function parameter is α , not necessarily scalar. Note, that the use of the subscript g conveys the dependence on the marks through the boost function g and is not to be confused with the same notation used in Daley and Vere-Jones (2007) for the ground process.

The $d \times 1$ marks \mathbf{X} impact the intensity through the scalar valued boost function $g : \mathbb{R}^d \rightarrow \mathbb{R}_+$, which is parametrized with a vector ψ of length r , specifying the way in which marks enter the boost function and the parameters ϕ , required to specify the distribution of the marks. To obtain a stationary solution to (4.1), the boost function is normalized so that $\mathbb{E}_\phi[g(\mathbf{X}; \phi, \psi)] = 1$ for all ϕ and ψ .

Embrechts et al. (2011, Definition 3) assume that the marks are unpredictable as defined in Daley and Vere-Jones (2007, Definition 6.4.III(b)) so that the distribution of \mathbf{X}_i , the mark at time t_i is independent of previous event times and marks, i.e. of $\{(t_j, \mathbf{X}_j)\}$ for $t_j < t_i$. An example of unpredictable marks is where the marks are conditionally i.i.d. given the past of the process, but the marks may impact on the future of the intensity λ_g as in (4.1). Marks are assumed to have a multivariate density $f(\mathbf{x}; \phi)$, which could be further specified to have a copula structure as we do in some of the examples below. Under the conditions (4.2) on the decay function, the branching coefficient (which defines the spectral radius here) $\vartheta < 1$, and the normalizing condition $\mathbb{E}_\phi[g(\mathbf{X}; \phi, \psi)] = 1$, there exists a unique point process associated with intensity (4.1) (Embrechts et al., 2011, Proposition 1). Liniger (2009) remarks that the normalizing condition is designed such that the parameters ϑ appears explicitly in the model, allowing for the interpretation of whether the Hawkes process is well defined or not. The general normalized boost functions, which satisfy these requirements, can be constructed with some function $h(\mathbf{X}, \psi)$ depending on an r -dimensional parameter ψ , taking values in a suitable parameter space Ψ . We define the boost function as

$$g(\mathbf{X}; \psi, \phi) = \frac{h(\mathbf{X}; \psi)}{\mathbb{E}_\phi[h(\mathbf{X}; \psi)]}. \quad (4.4)$$

Note that the boost function g depends on the parameters ϕ of the marks distribution due to the normalization using $\mathbb{E}_\phi[h(\mathbf{X}; \psi)]$. We assume that h satisfies the property, $h(\mathbf{X}; 0) \equiv 1$. The property $h(\mathbf{X}; 0) \equiv 1$ is without loss of generality since the boost function is normalized. Based on the properties required of h , we have $\mathbb{E}_\phi[g(\mathbf{X}; \phi, \psi)] = 1$ for all $\psi \in \Psi$, $g(\mathbf{X}; \phi, 0) \equiv 1$. The requirements that $\mathbb{E}_\phi[h(\mathbf{X}; \psi)]$ exist imposes obvious conditions on the marginal distribution of \mathbf{X}_m . For example, if $h(\mathbf{X}; \psi)$ is a polynomial of degree p in \mathbf{X} then $\mathbb{E}_\phi[\mathbf{X}^p]$ needs to exist.

Many examples of boost functions including those presented in Liniger (2009) and those used in the simulations and applications in this research, satisfy these conditions.

For example, boost functions which are additive in functions of the marks specified as

$$h(\mathbf{X}; \psi) = 1 + \sum_{j=1}^r \psi_j h_j(\mathbf{X}), \quad (4.5)$$

or boosts which are multiplicative in the functions of the marks are specified as

$$h(\mathbf{X}; \psi) = \prod_{j=1}^r (1 + \psi_j h_j(\mathbf{X})). \quad (4.6)$$

To aide in the intuition behind the Hawkes process in the context of LOB modelling, the immigration intensity η , is the frequency with which new events arrive in the LOB. The intensity is increased proportionally by the branching coefficient ϑ , given an event has occurred. When new orders (immigrants) arrive in the LOB, we expect a clustering effect which results in an increase in secondary events called descendants. How fast this effect decays in time is governed by the decay function w . The boost function controls how strong the effect of the mark of some event is on the intensity and is the influence of both the time lag and the mark value after an event. The marks are assumed to be independent of the past intensities of the Hawkes process. The marks allow a complete specification of the dynamics of the Hawkes process (Liniger (2009)).

4.1.1 Copula models for distributions of the marks

In this section we present some results that are directly relevant to the application of modelling the joint dependence structures of marks within this research. For definitions of the two classes of copula models that we consider in this work, Gaussian and Archimedean see Appendix A.1. The descriptions in Appendix A.1 and the detail within this section on copulas, relies heavily on the work by Cruz et al. (2015).

We will consider the joint distribution of the marks constructed according to a meta-copula framework, see Balkema et al. (2010) and Cruz et al. (2015). Under such a framework, we consider developing the joint distribution of the marks random vector based on the second component of Sklars theorem. The joint multivariate distribution follows

$$F_{\mathbf{X}}(x_1, \dots, x_d) = C(F_1(x_1), F_2(x_2), \dots, F_d(x_d); \Upsilon), \quad (4.7)$$

for some choice of model for the copula C , parametrized by Υ and marginal distributions.

Definitions which are required for the boost normalization in the event of joint dependence follows. **In the case of the copula C , the off diagonal terms given when utilizing the bivariate Spearman's rank correlation for any copula C on the mark random vector are**

$$\mathbb{E}[(X_1 - \mu_{1,1})(X_2 - \mu_{1,2})] = \sqrt{\mu_{2,1}\mu_{2,2}} \varphi \left(\left[12 \int_{[0,1]} \int_{[0,1]} u_1 u_2 dC(u_1, u_2) - 3 \right] \right), \quad (4.8)$$

for $\varphi(\rho_s) = \rho$, where ρ denotes the linear correlation coefficient between X_1 and X_2 , $\varphi(\cdot)$ is the mapping and is known in closed form for many copula families.

It will also be valuable at this stage to introduce the definitions of Spearman's rank correlation in terms of a copula density in a bivariate settings. The bivariate Spearman's rank correlation can be expressed explicitly via the bivariate copula C according to

$$\rho_S(X_1, X_2) = 12 \int_{[0,1]} \int_{[0,1]} u_1 u_2 dC(u_1, u_2) - 3. \quad (4.9)$$

We utilize the rank transformation of the marks to obtain an expression for any copula for the Spearman's rank correlation from (4.9), to obtain the off diagonal terms explicitly. Note, we use

$$\begin{aligned} \rho_S(X_1, X_2) &= 12 \int_{[0,1]} \int_{[0,1]} u_1 u_2 dC(u_1, u_2) - 3 \\ &= 12\mathbb{E}[U_1 U_2] - 3 \\ &= \frac{\mathbb{E}[U_1 U_2] - 1/4}{1/12} \\ &= \frac{\text{Cov}[U_1, U_2]}{\sqrt{\text{Var}[U_1] \text{Var}[U_2]}} \\ &= \rho(F_1(X_1), F_2(X_2)), \end{aligned} \quad (4.10)$$

hence, we need to introduce the mapping $\rho = \varphi(\rho_S)$ to adjust for the rank correlation versus the linear correlation.

For many copula families, the explicit solution for the copula based expression for the Spearman's rank correlation is known explicitly in terms of the copula parameters. Furthermore, in many copula families the mapping $\varphi(\cdot)$ is also known in closed form. In the case of bivariate Gaussian copula with correlation parameter ρ , and using Cruz et al. (2015, equation 10.51) the following relation is true

$$\rho(F_1(X_1), F_2(X_2)) = \varphi(\rho_S(X_1, X_2)) = 2 \sin\left(\frac{\pi}{6} \rho_S(X_1, X_2)\right). \quad (4.11)$$

4.1.2 Quasi-likelihood

The log-likelihood and associated asymptotic statistical properties for the unmarked Hawkes process has a long history, see Ozaki (1979), Ogata (1978) or Anderson et al. (1996) and Clinet and Yoshida (2017) for example. The log-likelihood in Embrechts et al. (2011) for the stationary marked case with infinite past history is formally defined in Liniger (2009) and Embrechts et al. (2011). In practice, the point process and marks are only observed over the interval of time $[0, T]$, and the stationary intensity process defined by (4.1) cannot be computed because it relies on the infinite past. For computations, a truncated version must be used as presented in Liniger (2009), which calculates the likelihood over available event times. For the unmarked case, Ozaki (1979) also approximates the stationary intensity process

$$\lambda^\infty(t; \theta) = \eta + \vartheta \int_{(-\infty, t)} w(t-s; \alpha) N(ds), \quad t \in \mathbb{R}, \quad (4.12)$$

by one which is non-stationary presented below. Ogata (1978) shows that the asymptotic properties of the likelihood estimates based on the stationary and non-stationary version are equivalent for the exponential decay Hawkes process. Likewise Clinet and Yoshida (2017) consider estimation based on the quasi likelihood constructed using observations available over $[0, T]$ and the non-stationary intensity process

$$\lambda(t; \theta) = \eta + \vartheta \int_{[0, t)} w(t - s; \alpha) N(ds), \quad t \in \mathbb{R}. \quad (4.13)$$

The equivalent non-stationary formulation for the marked process intensity in (4.1) is required here

$$\lambda_g(t; \theta, \phi, \psi) = \eta + \vartheta \int_{[0, t) \times \mathbb{X}} w(t - s; \alpha) g(\mathbf{x}; \phi, \psi) N_g(ds \times d\mathbf{x}), \quad t \in \mathbb{R}. \quad (4.14)$$

For the unmarked and exponential decay Hawkes process Clinet and Yoshida (2017, Proposition 4.4) show, using results of Bremaud and Massoulié (1996), that a suitable probability space exists on which a stationary version can be defined, and for which the non-stationary version in (4.14) converges suitably to the stationary version.

To proceed to the definition of the quasi log-likelihood, let $\theta = (\eta, \vartheta, \alpha) \in \Theta$, $\phi \in \Phi$ and $\psi \in \Psi$ for some parameter spaces Θ , Φ and Ψ . Let $\nu = (\theta, \phi, \psi) \in \Theta \times \Phi \times \Psi$ be the collection of all parameters for the marked process with intensity function (4.14).

The quasi-likelihood version of their definition, which is very similar to that in the practical implementation of Liniger (2009), specifies the quasi-log-likelihood for the observed event times and associated marks as

$$\begin{aligned} l_g(\nu) &= \int_{[0, T] \times \mathbb{X}} \ln \lambda_g(t; \nu) N_g(dt \times d\mathbf{x}) - \Lambda_g(T; \nu) \\ &\quad + \int_{[0, T] \times \mathbb{X}} \ln f(\mathbf{x}; \phi) N_g(dt \times d\mathbf{x}), \end{aligned} \quad (4.15)$$

in which the compensator at T is

$$\Lambda_g(T; \nu) = \int_{[0, T]} \lambda_g(t; \nu) dt. \quad (4.16)$$

This likelihood is derived under the assumption that the marks are conditionally independent given the event times. Later in Chapter 6 we generalize this to allow for serial dependence in the marks. Note that all three components of $l_g(\nu)$ are functions of the mark density parameters ϕ , because the boost function normalization requires moments of the functions of the marks. This means that the quasi maximum likelihood estimates are obtained as the values of ν , which optimise (4.15). From here-on we will drop the qualifier ‘quasi’ and refer to $l_g(\nu)$ as the log-likelihood, and to the estimates obtained from it as the maximum likelihood estimators.

4.1.3 Residual process

Works by Daley and Vere-Jones (2007) and Liniger (2009) present the random time change theorem and the residual analysis theorem that follows. The random time transformation $\tau(t; \nu) = \Lambda(t; \nu)$ takes the point process with conditional intensity function $\lambda(t; \nu)$, assuming $\lambda(t; \nu) > 0$ and $\Lambda(t, \nu) < \infty$ over $[0, T]$, into a Poisson process with unit rate. The random time change transformation underpins the goodness-of-fit test based on the residual point process. For a conditional intensity function, and hence a compensator which is explicitly known, the transformed sequence $\{\tau(t_1, \nu), \tau(t_2, \nu), \dots\}$ is a realization of a Poisson process with unit rate, if and only if the original sequence $\{t_1, t_2, \dots\}$ is a realization from the point process defined by $\Lambda(t, \nu)$.

4.2 Numerical algorithms

4.2.1 Estimation of the Hawkes process

The section that follows describes the numerical algorithms for calculating the log-likelihood in (4.15), specification of decay and boost functions and optimization of the log-likelihood. For the description of the algorithms for the marked Hawkes process, we rely considerably on Liniger (2009) with modifications for extensions.

The algorithm for the log-likelihood function, derived assuming the mark vectors are independent and identically distributed, expresses the integrals in (4.15) by summations over the events as follows.

Algorithm 4 (Hawkes process likelihood function). *Let the observation period be over the interval $t \in (0, T]$ and $\mathbf{X}_i \in \mathbb{X} \subset \mathbb{R}^d$ be a vector of d marks. In the case where the components of the mark vector are independent, the likelihood function is given by*

$$\hat{l}(\nu) = \sum_{i=1}^n \ln \hat{\lambda}(t_i; \nu) - \hat{\Lambda}(T; \nu) + \sum_{i=1}^n \sum_{j=1}^d \ln f_j(x_i; \phi). \quad (4.17)$$

In the case of jointly dependent marks the likelihood function is given by

$$\hat{l}(\nu) = \sum_{i=1}^n \ln \hat{\lambda}(t_i; \nu) - \hat{\Lambda}(T; \nu) + \sum_{i=1}^n \sum_{j=1}^d \ln f_j(x_i; \phi) + \sum_{i=1}^n \ln c(F_1(x_{i1}), \dots, F_d(x_{id}); \Upsilon), \quad (4.18)$$

using a copula specification of the joint dependence.

Algorithm 5 (Hawkes intensity process). *To evaluate Algorithm 4, we require the calculation of the intensity process defined in (4.14), which in summation form is*

$$\hat{\lambda}(t_i; \nu) = \eta + \vartheta \sum_{i=1}^{i-1} w(t_i - t_m) g(\mathbf{x}_i; \phi, \psi). \quad (4.19)$$

Algorithm 6 (Compensator). *The compensator in (4.16) at time T , required for Algorithm 4, can be estimated by*

$$\hat{\Lambda}(t, \nu) = \eta T + \vartheta \sum_{i=1}^n W(t_n - t_i; \alpha) g(\mathbf{x}_i; \phi, \psi). \quad (4.20)$$

Recall from (4.3) $W(t; \alpha)$ is the cumulative decay function.

Some notes on the evaluation of copula models

The notes below are presented assuming bivariate marginal mark distributions F_1 and F_2 . Descriptions of functions are in MATLAB, however equivalent functions will be readily available in other statistical software. For the coding examples assume $\mathbf{u} = (F_1(x_1), F_2(x_2))$.

- Estimate one of the rank correlations, Spearman's rank, $\rho_s(F_1(x_1), F_2(x_2))$ (e.g. `corr(u, 'type', 'Spearman')`) or Kendall's tau, $\rho_\tau(F_1(x_1), F_2(x_2))$ (e.g. `\rho_s = corr(u, 'type', 'Spearman')`).
- The rank correlation is used to evaluate the copula parameter (e.g. `\rho = copulaparam('Gaussian', \rho_s, 'type', 'spearman')`).
- For the Archimedean copulas that do not have an analytic formula for the calculation of the copula parameter from Spearman's rank, it is advised to use Kendall's tau, despite MATLAB providing an approximation to Spearman's rank.
- Evaluate the probability density function of the bivariate copula (e.g. `c(F_1(x_1), F_2(x_2)) = copulapdf('Gaussian', u, \rho)`).
- The generic function, 'copulapdf' within MATLAB does not evaluate the probability density function for dimensions $d > 2$.
- The resulting likelihood estimation is achieved by maximizing the log-likelihood, given by $\sum_{i=1}^n \ln c(F_1(x_1), F_2(x_2))$.
- The jointly dependent marks enter the intensity of the likelihood function via the boost function. In the case of jointly dependent marks, an adjustment is required to the normalization of the boost function. This will be covered in detail in Section 4.2.1.

Decay functions

The non-negative decay function w , is often presented as an exponential decay and less frequently as a power-law decay function. Due to the computational challenges associated with the implementation of a Hawkes process with a power-law decay, mixture of exponential decay functions have also been proposed in recent literature, see Hardiman et al. (2013) and Filimonov and Sornette (2015). In Section 3.2.1 we established that an exponential decay function was the most appropriate for this research, however for completeness

we will present the description of both forms here.

For $t \geq 0$ and parameter α , the exponential decay function is

$$w(s; \alpha) = \alpha e^{-\alpha s}, \quad s > 0, \quad \alpha > 0, \quad (4.21)$$

and the associated cumulative decay function W is

$$W(s; \alpha) = 1 - e^{-\alpha s}. \quad (4.22)$$

For $t \geq 0$ and parameters α and β , the power decay function is

$$w(s; \alpha, \beta) = \frac{(\alpha - 1)\beta}{(1 + \beta s)^\alpha}, \quad s > 0, \quad \alpha > 2, \quad \beta > 0, \quad (4.23)$$

and the associated cumulative decay-law function W is

$$W(s; \alpha, \beta) = 1 - \frac{1}{(1 + \beta s)^{\alpha-1}}. \quad (4.24)$$

The requirements of (4.2) for both the exponential and power-law decay functions clearly hold.

The algorithm used for calculating the intensity process is more efficient when the Hawkes process has an exponential decay function. Recall that a process is a Markov process if it has the property, conditional on the present, that the future is independent of the past. When the decay function is exponential then the intensity process is a Markov process, as shown by Ogata (1981). If we consider a Hawkes process with an exponential kernel in (4.21) for intensity (4.14), then $(N(t), \lambda(t))$ is a Markov process of the form $d\lambda(t) = \lambda(t)dt + \alpha dN(t)$. This property is extended to the case of mixture exponential decay functions. For the power-law decay function in (4.23), this does not hold and all past history is required to perform simulation and estimation.

If the decay function is exponential and to evaluate Algorithm 4, we use Algorithm 7, instead of Algorithm 5.

Algorithm 7 (Hawkes intensity process with an exponential decay function). *Assume the decay function in (4.21), the value $\lambda(t_{i-1})$ is known for some $t_{i-1} \in \mathbb{R}$ and $0 < t_{i-1} < t_i < \dots < t_n \leq T$. The first value of the intensity process is estimated by $\hat{\lambda}(t_1; \nu) = \eta$. Evaluate the following equation for $2 \leq i \leq m$*

$$\hat{\lambda}(t_i; \nu) = \eta + e^{-\alpha(t_i - t_{i-1})} [\lambda(t_{i-1}) - \eta] + \vartheta \alpha e^{-\alpha(t_i - t_{i-1})} g(\mathbf{x}_i; \phi, \psi). \quad (4.25)$$

Boost Functions

This general specification of mark vector and associated boost function in (4.4) accommodates many standard examples occurring in the literature Liniger (2009); Embrechts et al. (2011); Fauth and Tudor (2012), as well as allowing extensions to new specifications. Special cases of boost functions, such as polynomials, exponential or power laws can be applied to each individual mark in \mathbf{X}_n and these can be combined multiplicatively or additively.

Polynomial Boost. This includes linear, quadratic and higher order polynomials

$$g(x; \phi, \psi) = \frac{1 + \sum_{j=1}^p \psi_j x^j}{1 + \sum_{j=1}^p \psi_j \mathbb{E}[X^j]}. \quad (4.26)$$

Power law boost

$$g(x; \phi, \psi) = \frac{x^\psi}{\mathbb{E}[X^\psi]}. \quad (4.27)$$

Exponential Boost

$$g(x; \phi, \psi) = \frac{\exp(\psi x)}{\mathbb{E}[\exp(\psi X)]}. \quad (4.28)$$

We now illustrate some examples of how boost functions combine the impact of the individual marks multiplicative and in the following ways.

Example 1: where g_1 is a quadratic function of x_1 and g_2 is a linear function of (x_2, x_3) .

$$g(\mathbf{x}; \phi, \psi) = g_1(x_1; \phi, \psi)g_2(x_2, x_3; \phi, \psi),$$

$$g_1(x_1; \phi, \psi) = \frac{1 + \psi_{1,1}x_1 + \psi_{1,2}x_1^2}{1 + \psi_{1,1}\mathbb{E}_\phi[X_1] + \psi_{1,2}\mathbb{E}_\phi[X_1^2]},$$

where, $\psi_{1,1}$ and $\psi_{1,2}$ are the linear and quadratic parameter for x_1 .

$$g_2(x_2, x_3; \phi, \psi) = \frac{1 + \psi_{2,3}x_2 + \psi_{3,4}x_3}{1 + \psi_{2,3}\mathbb{E}[X_2] + \psi_{3,4}\mathbb{E}[X_3]},$$

where, $\psi_{2,3}$ and $\psi_{3,4}$ are the linear parameter for x_2 and x_3 , respectively. For this example we have a collection of four boost parameters to estimate for three marks, $\{\psi_{1,1}, \psi_{1,2}, \psi_{2,3}, \psi_{3,4}\}$.

Example 2: where g_1 is a linear function of x_1 , g_2 is a linear functions of x_2 and x_1 and x_2 are jointly dependent.

$$\begin{aligned} g(\mathbf{x}; \phi, \psi) &= g_1(x_1; \phi, \psi)g_2(x_2; \phi, \psi) \\ &= \frac{1 + \psi_1 x_1 + \psi_2 x_2 + \psi_1 \psi_2 x_1 x_2}{1 + \psi_1 \mathbb{E}[X_1] + \psi_2 \mathbb{E}[X_2] + \psi_1 \psi_2 \mathbb{E}[X_1 X_2]}. \end{aligned}$$

In practice, the distribution of the marks could be specified and fit so that the theoretical expectations $\mathbb{E}[X^j]$ could be evaluated. We consider up to the order four, $j \in \{1 \dots 4\}$ for the moments. See Appendix A.2 for the identities linking marginal central moments with non-central (raw) moments.

Normalization of boost function in the presence of joint dependence

We will demonstrate the normalization using the multiplicative form of the boost function in (4.6) with bivariate marks. For the example that follows, the notation is simplified to ψ_1 and ψ_2 , being the linear boost parameters for x_1 and x_2 . We consider the independence copula W and marginal mark distributions F_1 and F_2 with finite moments up to at least the second order, such that $\mu'_{2k} < \infty$ for $k \in \{1, 2\}$. The boost function g is constructed

by starting with some function $h(\mathbf{X}; \psi)$, which is linear for this example, such that

$$h(\mathbf{X}; \psi) = \prod_{j=1}^2 (1 + \psi_j h_j(\mathbf{X})) = (1 + \psi_1 X_1)(1 + \psi_2 X_2).$$

The boost function is normalized using $\mathbb{E}[h(\mathbf{X}; \psi)]$, where

$$\begin{aligned} \mathbb{E}[h(\mathbf{X}; \psi)] &= \mathbb{E}[(1 + \psi_1 X_1)(1 + \psi_2 X_2)] \\ &= 1 + \psi_1 \mu_{1,1} + \psi_2 \mu_{1,2} + \psi_1 \psi_2 \mathbb{E}[X_1 X_2]. \end{aligned}$$

In the case of jointly dependent marks $\mathbb{E}[X_1 X_2] \neq 0$ and

$$\mathbb{E}[X_1 X_2] = \mathbb{E}[(X_1 - \mu_{1,1})(X_2 - \mu_{1,2})] + \mu_{1,1} \mu_{1,2},$$

where

$$\mathbb{E}[(X_1 - \mu_{1,1})(X_2 - \mu_{1,2})] = \rho \sqrt{\mu_{2,1} \mu_{2,2}},$$

where ρ is the pairwise linear correlation between X_1 and X_2 . Recall in (4.8), utilizing the bivariate Spearman's rank correlation for any copula C on the mark random vector we get

$$\mathbb{E}[(X_1 - \mu_{1,1})(X_2 - \mu_{1,2})] = \sqrt{\mu_{2,1} \mu_{2,2}} \varphi \left(\left[12 \int_{[0,1]} \int_{[0,1]} u_1 u_2 dC(u_1, u_2) - 3 \right] \right),$$

for $\varphi(\rho_s) = \rho$.

For the Gaussian copula, we can obtain the pairwise linear correlations by transforming the Spearman's rank correlation ρ_s in (4.11), obtaining

$$\mathbb{E}[(X_1 - \mu_{1,1})(X_2 - \mu_{1,2})] = \sqrt{\mu_{2,1} \mu_{2,2}} 2 \sin \left(\frac{\pi}{6} \rho_S(X_1, X_2) \right).$$

Therefore, the normalization in the presence of joint dependent bivariate marks that are modelled via a Gaussian copula is

$$\begin{aligned} \mathbb{E}[h(\mathbf{X}; \psi)] &= \mathbb{E}[(1 + \psi_1 X_1)(1 + \psi_2 X_2)] \\ &= 1 + \psi_1 \mu_{1,1} + \psi_2 \mu_{1,2} + \psi_1 \psi_2 \sqrt{\mu_{2,1} \mu_{2,2}} 2 \sin \left(\frac{\pi}{6} \rho_S(X_1, X_2) \right) + \psi_1 \psi_2 \mu_{1,1} \mu_{1,2}. \end{aligned} \tag{4.29}$$

For copula models that do not have a closed form solution for calculating the linear correlation, such as the Gumbel and Clayton, we use a Monte-Carlo method as presented in Algorithm 8. Alternative methods to this may include quadrature, or the use of Kendall's correlation ρ_τ as a robust (but biased) estimator of the linear correlation. The linear correlation coefficient based on the covariance of the two variables is not preserved by copulas. However, the Kendall's correlation is a constant of the copula, see (Cruz et al., 2015, Chapter 10).

4.2.2 Simulation algorithm for the univariate Hawkes process

To simulate a series of random events according to the specification of a given Hawkes process, we utilize Ogata’s modified thinning algorithm (Ogata, 1981), which can be applied to any choice of decay function. For the general framework and notation, we rely on the algorithm description by Liniger (2009, Algorithm 1.21), however extensions are made for the inclusion of joint dependence structures for the marks.

The intensity process λ , is left continuous with right-hand side limits. For the purposes of simulation, we define λ^+ as the right-hand side limit after time t , that is, after the intensity is boosted. We denote λ^- as the left-hand-side limit, also known as the *hazard rate* (Liniger, 2009).

The algorithm is comprised of an outer and inner loop. The outer loop iteration variable is $i \in \{2, \dots, n\}$ and the inner loop is indexed by $r \geq 1$. For brevity we present only the procedure for the recursive method below. The recursive and non-recursive methods are presented together in Appendix A.3, with a brief description of the vectorization of the intensity function to enhance the speed of the simulation.

Algorithm 8 (Hawkes process simulation). *Initialization:*

$$i = 2; \quad \lambda^+(1) = \eta/(1 - \vartheta).$$

Monte-Carlo method for the case of jointly dependent marks.

Examples of some functions in MATLAB are provided below.

1. For $Y_{i,j} \sim F_{i,j}$, where $j \in \{2, \dots, d\}$ are jointly dependent random variables with copula W and marginal distributions $F_{i,1}, \dots, F_{i,d}$, let $U_{i,1}, \dots, U_{i,d}$ have joint distribution C . For a very long time series $i = 1 \dots n$ (i.e. $n = 20,000$):

(a) Simulate a random variable $u_{i,1} \sim U[0, 1]$,

- with example code in MATLAB, $\rho = \text{copulaparam}(\text{'Gaussian'}, \rho_s, \text{'type'}, \text{'spearman'})$ and $[u_1, u_2] = \text{copularnd}(\text{'Gaussian'}, \rho_s, 1)$;

(b) Simulate a random variable $u_{i,2}$ from $C_{i,2}(\cdot | u_{i,1})$. Continue simulating, such that you simulate a random variable $u_{i,d}$ from $C_{i,d}(\cdot | u_{i,d-1})$;

(c) Sample $(Y_{i,1}, \dots, Y_{i,d})$ from $F^{-1}(u_{i,1}), \dots, F^{-1}(u_{i,d})$,

- with example code in MATLAB: $y_{i,j} = \text{gpinv}(u_{i,j}, \zeta_j, \delta_j, 0)$.

2. Define $y_{i,j} := Y_{i,j}$ and calculate the long run empirical moments and correlation in the event of joint dependence, which we will consider below and as required for the specification of boost function, for example, $\mathbb{E}[Y_j]$, $\text{Var}[Y_j]$ and $\rho(Y_1, \dots, Y_d)$.

Outer loop

1. *Initialization:*

$$\tau_1 := t_{i-1}; \quad \lambda_r^- := \lambda_{i-1}^+.$$

Inner loop

(a) Calculate a new time τ_r

$$\tau_r = \begin{cases} t(i-1) + E/\lambda^+(i-1), & \text{if } r = 1; \\ \tau_{r-1} + E/\lambda^+(i-1), & \text{otherwise,} \end{cases}$$

where $E \sim \text{Exp}(1)$.

(b) Calculate the left hand limit intensity λ_r^- . For the case of an exponential decay function in (4.21), we can utilize the more efficient recursive expression for the simulation

$$\lambda_r^- = \eta + e^{-\alpha(\tau_r - t_{i-1})} [\lambda_{i-1}^+ - \eta].$$

(c) Sample a standard uniform random variable $U_r \sim U(0, 1)$ and let

$$u_r := U_r \lambda_{i-1}^+.$$

Check the condition

$$u_r \leq \hat{\lambda}_r^-.$$

If this condition is true, exit the inner loop. Failing this condition being met, we start the inner loop again, sampling a new τ_r where $r = r + 1$.

Exit the inner loop.

2. Define $t_i := \tau_r$.
3. Define $\lambda(i) := \lambda^-(r)$.
4. In the case of univariate marks or multi-dimensional marks that are independent, sample a random variable from $X_{i,j} \sim F_{i,j}$, where $j \in \{1, \dots, d\}$ and define $x_{i,j} := X_{i,j}$.
5. In the case of $d \geq 2$ jointly dependent marks with copula W and marginal mark distributions $F_{i,1}, \dots, F_{i,d}$, let $U_{i,1}, \dots, U_{i,d}$ have joint distribution C . Follow the method outlined in ‘Monte-Carlo method for the case of jointly dependent marks’ [Step (a)], to obtain samples $(X_{i,1}, \dots, X_{i,d})$ from $F^{-1}(u_{i,1}), \dots, F^{-1}(u_{i,d})$.
6. Calculate the boost function $g(\mathbf{x}; \phi, \psi)$, using the normalization adjustment in Section 4.2.1 in the event of jointly dependent marks. For the evaluation of $\mathbb{E}[X_1 X_2] \neq 0$, we require the estimation of the linear correlation. As described in Section 4.2.1, if this cannot be calculated explicitly, we utilize the long run empirical linear correlation from the ‘Monte-Carlo method for the case of jointly dependent marks’ above, $\rho(Y_1, \dots, Y_d)$.
7. Define $\lambda^+(i) = \lambda_r^- + \vartheta \alpha g(\mathbf{x}; \phi; \psi)$.
8. Let, $i=i+1$.

For the implementation of the algorithm, we need to define an end point for the iteration variable r , which we set to 200.

4.2.3 Goodness-of-fit algorithms

As we presented in Section 4.1.3 a point process can be characterized by the form of its compensator, which is a deterministic function. The random time transformation takes the point process with conditional intensity function $\lambda(t)$, into a unit-rate Poisson process. Therefore, the durations of the integrated process has a unit exponential distribution (i.e. with unit mean). Recall that the residual process is the transformed times and can be evaluated by Algorithm 6

The hypothesis that the estimated residual process comes from a unit rate Poisson process can be tested using a Kolmogorov-Smirnov (KS) statistic. Conditioned on the number of events, the location of the events is uniformly distributed in the observation interval (Liniger, 2009).

Algorithm 9 (Goodness-of-fit based on the residual point process). *The residual process can be evaluated by the following steps:*

1. Form the sequence of transformed times τ_i , as described in Algorithm 6;
2. Rescale the time domain to the unit interval $\tau_i^{norm} = \tau_i/\tau_n$;
3. Plot the counting function consisting of points $(\tau_i^{norm}, \frac{i}{n})$;
4. Calculate the upper and lower limits $\frac{i}{n} \pm k$, where k is the critical value from the Komogorov-Smirnov Test Statistic.

4.2.4 Computational considerations

- *Computational time.* The log-likelihood calculation for the Hawkes process can be computationally expensive. It should be noted that the computational complexity, which is often mentioned in literature pertaining to the log-likelihood calculation of the Hawkes process, relates to the multivariate Hawkes process where there are multiple components and thus multiple decay functions within the process. This results in a double summation in the likelihood function intensity component and the running time of calculations will have a quadratic complexity $O(N_T^2)$. For the Hawkes process with an exponential decay, which utilizes the recursive procedure has complexity of $O(N_T)$ (Ogata, 1981). In Appendix A.3 we discuss the computational challenges, with recommendations for the Hawkes process with a power-law decay function.
- *Initialization.* We assume that the initial value of the intensity is $\lambda(t_0) = C$ for some specified value of C . A sensible selection for the initialization of the intensity is

$$C = \mathbb{E}[\lambda(t)] = \eta/(1 - \vartheta), \quad (4.30)$$

which is the theoretical long run average for a stationary Hawkes process (Laub et al., 2015). Note also that this intensity function is not defined using events prior to the observation period where $t < 0$, in practice (4.14) is used for the computation of the likelihood.

- *Edge effects.* The Hawkes process is observed over the interval $t \in [0, T]$. The Hawkes process is comprised of clustered events, where there is a primary (parent) event, followed by secondary (descendant) events. As described in Daley and Vere-Jones (2007) and Rasmussen (2011) it is possible that the secondary events are observed within the interval $t \in [0, T]$, however the primary event occurred before the observation period and this is called the ‘edge effect’. This would mean that the contribution to the event intensity would not be counted. As described by Filimonov and Sornette (2015), the edge effect results in an underestimated branching coefficient and an overestimated immigration intensity, due to events occurring before the time window not being accounted for. For larger memory processes, such as an intensity function with a power law decay, the edge effect becomes more pronounced. Common practice is to set the initial value of the intensity to the immigration intensity (Daley and Vere-Jones, 2007), which is also the approach in this research. It is also noted in Rasmussen (2011), that for large samples the edge effect will be negligible.
- *Burn-in.* To limit the impact of edge effects of the thinning algorithm described above, a burn-in period for the Hawkes process is required. Filimonov et al. (2015) generate 100 replicates of sample size $T = 3500$ each. The first 500 points ($\approx 14\%$) are discarded to reduce the edge effects. A significantly larger burn-in is required for the Hawkes process with a power-law decay function, whereby Filimonov and Sornette (2015) suggest 1,000 times T . As noted in Section 3.2.1, the data used by Filimonov and Sornette (2012, 2015) considers an event process consisting of mid-price changes at a minimum time granularity of 1 second. This represents a tiny fraction of the events that are considered from an event process consisting of all event arrivals for the LOB with a depth of 5 levels for the same end time, T . Whilst the research by Filimonov and Sornette (2015) provides an excellent discussion of the key challenges when fitting a Hawkes process to financial data sets, a direct comparison of the two counting processes is not reasonable. Upon inspection, a burn-in of 1% for the high frequency data sets considered in this research is found to be sufficient for an exponential decay function and is used throughout. Additional challenges would need to be addressed for larger burn-ins required for a power law decay function, which is computationally burdensome as the model complexity increases.
- *Global versus local maxima.* The joint likelihood function of the Hawkes process is complex and the intricacies are further increased with the introduction of marks and marks with joint dependence. The joint likelihood function may not be globally convex and it follows that the log-likelihood estimation may result in multiple stationary points corresponding to local minima and/or maxima. For increasing

complexity of Hawkes process in Section 4.5, we study the impact of varying initializations to give confidence that the global maximum has been achieved with the method implemented. In the application of the Hawkes process to real data, we suggest performing the log-likelihood estimation with varying initializations and consider various optimization routines to ensure the parameter estimates are reliable.

4.3 Software for modelling the Hawkes process

To model the Hawkes process, we require a software package that is able to simulate and model the range of specifications we consider. The software package must also provide the flexibility to easily extend the existing code to a broader range of model specifications, such as heavy tailed distributions, discrete distributions and dependence features of the marks. Upon review of existing software, software developed in R by Liniger (2009) was the most comprehensive available and it was initially investigated as a potential platform. This software supports several continuous mark distributions, exponential, gamma, inverse gamma, log-normal, Pareto, Rayleigh, Weibull, normal and Student distributions. It also supports various boost functions, polynomial, power and exponential.

Early within the research it became apparent that this existing software was not adequate and extensions to the code were not easily achieved, as the majority of the source code is in C++. However, this code provided an excellent platform for verification of the simpler models, which we subsequently developed in MATLAB. Using the sample of DAX data provided in the R software package by Liniger (2009), we verified the accuracy of the estimation of the MATLAB code and in addition, to assess the robustness of our parameter estimations. In all cases the MATLAB optimized likelihood values match those in R. In all cases the MATLAB parameter estimates are more stable than those estimated in R. In all cases the parameter estimates in MATLAB and R are relatively close and improved for an increasing sample size.

It is also worth highlighting a limitation of the simulator developed in R. The simulator takes a model object, which is first obtained via log-likelihood estimation. It is not possible to directly specify the parameters for simulation. The purpose of this simulator is to first assume that a Hawkes model has been fitted to some data and then one wants to simulate a Hawkes process according to the estimated parameters. Passing the same model object to the simulator in R, produces identical results for replicates > 1 , suggesting the simulator is seeded, but without confirmation. In light of these limitations, the development of a simulation procedure for the Hawkes process was built in MATLAB. We also compared our simulation with that of HAWKESDEMO Dimitri Shvorob's simulator, for a Hawkes process with exponential decay and constant boost, which was also built in MATLAB. However, the simplicity of the model specification meant that extensive comparisons were not possible.

The models developed in MATLAB are extensive in the choices of parametric distributions, boost functions, various forms of combining boost functions for multiple marks, power-law and exponential kernels, extensions to copulas for joint dependence and finally as presented in Chapter 7, the construction of a decoupled approximate likelihood

method. In addition, the development of supporting software for the Hawkes process includes: simulation procedures for all possible combinations of models; extensions include serial dependence within the marks; the ability to assess the fit of a model through residual plots; and visualization tools to present the modelled intensity process of the Hawkes process. With modular programming and vectorized operations, the reliability, extendibility and speed are enhanced and allow for future extensions to be easily applied within this flexible framework.

Tables 4.1 and 4.2 presents the possible model combinations available in the MATLAB code. It is worth noting, that extensions to additional marks distribution, alternative boost functions and copulas are easily done, however we only list those that are used within this thesis and that have been tested extensively.

- Method of simulation: recursive or non-recursive.
- Kernel functions: exponential or power-law.
- Copula functions: Gaussian, Gumbel or Clayton.
- Marks dimension with joint dependence: for bivariate marks only.
- Marks dimension with no joint dependence: multi-dimensions.
- Serial dependence: simulation only and induced via an autoregressive model introduced to create a time varying parameter within the distribution.
- Likelihood methods: quasi-likelihood or decoupled approximate likelihood methods.
- Methods to combine the boost functions: multiplicatively, additive or jointly additive.

Table 4.1: *Simulation*. Combinations of continuous and discrete mark distributions, boost functions and dependence structures available for simulation.

Marks distribution	Constant	Boost: poly	Boost: power	Multi-dimension	Marks Copula	Serial dependence
Exponential	✓	✓	✓	✓	✓	✓
Gamma	✓	✓	✓	✓	✓	×
GPD	✓	✓	N/A	✓	✓	✓
Weibull	✓	✓	✓	✓	✓	×
Normal	✓	✓	N/A	✓	✓	✓
Location–Scale	✓	✓	N/A	✓	✓	✓
Poisson	✓	✓	×	✓	✓	✓
Negative Binomial	✓	✓	×	✓	✓	✓

Table 4.2: *Maximum likelihood estimation.* Combinations of continuous and discrete mark distributions, boost functions and dependence structures available for likelihood estimation.

Marks distribution	Constant	Boost: poly	Boost: power	Multi-dimension	Copula
Exponential	✓	✓	✓	✓	✓
Gamma	✓	✓	✓	✓	✓
GPD	✓	✓	N/A	✓	✓
Weibull	✓	✓	✓	✓	✓
Normal	✓	✓	N/A	✓	✓
Location–Scale	✓	✓	N/A	✓	✓
Poisson	✓	✓	×	✓	✓
Negative Binomial	✓	✓	×	✓	✓

4.4 Optimization: Hessian and bootstrapping

Optimization of the objective function, which in this case is the log-likelihood function, can be done many ways. Some methods require computation of the first and possibly the second derivatives of the log-likelihood function with respect to the parameters. In earlier research, Ozaki (1979) derived the gradient, and by taking second derivatives, the Hessian for the Hawkes likelihood function in the case of a scalar unmarked Hawkes process with an exponential decay function. However, for this optimization of the likelihood in the marked case, formulae for the first and second derivatives with respect to all parameters, intensity parameters θ , marks parameters ϕ , and boost parameters ψ , not just θ as in Ozaki (1979) are needed. In particular, the derivatives and cross derivatives with respect to ϕ specifying the marks distribution, vary depending on the particular parametric or copula density that is used and they are much more complicated than the formulae given by Ozaki (1979) for the unmarked process. For every combination of boost function and marks density specification formulae, derivatives would need to be derived. Given the complexity of our applications, the amount of derivation work required for this would be prohibitive. We therefore decided to use an optimization algorithm that only requires the objective function to be calculated and any required derivatives would be calculated numerically.

The algorithm *fmincon* provided in the MATLAB Optimization Toolbox (MathWorks, 2018) is the chosen optimization procedure for evaluating the log-likelihood function. This is a non-linear programming solver that attempts to find the minimum of a function, subject to linear inequalities. There are five algorithm options available in *fmincon*, of which we select the quadratic programming algorithm *interior-point*. For a description of quadratic program algorithms, see Schittkowski (1986). The motivation for this selection is due to characteristics, such as low memory usage and the ability to solve large problems quickly (MathWorks, 2018). For a discussion on the advantages and disadvantages of the various algorithm options, see MathWorks (2018, Section 2-9).

The Hessian

One way to calculate the standard errors of the Hawkes process parameters, is by taking the inverse of the Hessian matrix to calculate the covariance matrix of the parameters.

The Hessian estimated in the function *fmincon* is the Hessian of the Lagrangian function, i.e. incorporating the constraints and the objective function. Whilst we perform point estimation of the MLE via an interior point method, it was found to be numerically robust to utilize a different class of optimization package based on numerical gradient methods for the approximation of the Hessian. As described in MathWorks (2018, Section 6-37), at each major iteration, an approximation is made of the Hessian of the Lagrangian function using a quasi-Newton updating method, which is the BFGS method (Broyden, 1970; Fletcher, 1970; Goldfarb, 1970; Shanno, 1970).

Quasi-Newton methods are widely employed and use the gradient vector at each iteration to construct a sequence of matrices, which approximate the Hessian of the objective function (or its inverse). Solvers require accurate gradients to converge correctly, however they do not usually require precise Hessians. In fact the approximation can be inaccurate as an approximation of the Hessian and yet still be useful in deciding the length of the step in the direction of the gradient that takes the algorithm to an optimal solution. However, for the evaluation of standard errors, a precise Hessian is required. Whenever there are constraints, the Hessian of the Lagrangian does not typically match the true Hessian of the objective function in every component, but only in certain subspaces (MathWorks, 2018). Whilst it can be suitable for point estimation in an MLE procedure, it will not produce reliable and practically useful information for the standard errors in the estimation parameters. This is due to the local curvature around the maximum of the likelihood potentially being grossly under or over estimated in sub-spaces of the parameter space, depending on the likelihood profile in each subspace. Therefore, the Hessian that *fmincon* returns can be inaccurate, which is also confirmed by MathWorks (2018).

Bootstrapping

An alternative method for estimating the standard errors is via a bootstrapping technique described in Algorithm 10. Using the fitted parameter estimates obtained from modelling the observed data, we simulate a series of random events according to the specification of Hawkes process that was used to fit the model to the observed data, using Algorithm 8. This is repeated many times to obtain an empirical distribution for each parameter. This method is feasible for a Hawkes process with multivariate marks when the marks are i.i.d, however it is computationally expensive. Extensions would be required for non i.i.d. marks using time series bootstrapping methods (Liu, 1988; Berkowitz and Kilian, 2000; Goncalves and Kilian, 2004).

In Section 4.5.5 we consider an alternative *approximate likelihood method by decoupling the marks parameters*, which replaces the theoretically determined moments for normalizing the boost by empirical estimates. Recall, that the normalization is required to define the process intensity $\lambda(t)$, and which appears in the first two components of the log-likelihood in (4.15). For the approximate likelihood method we use $\tilde{g}(\mathbf{x}; \psi) = \frac{h(\mathbf{x}; \psi)}{\bar{h}(\mathbf{x}; \psi)}$, where $\bar{h}(\mathbf{x}; \psi)$ is the function h evaluated using empirical moments in place of theoretical moments.

We will discuss the advantages of this method later in this chapter, however, relevant

to bootstrapping is the computational advantages of this method, with the decoupling resulting in a faster evaluation of the likelihood function. This makes it more feasible to increase the number of replicates for bootstrapping, thus reducing the simulation error. Whilst we introduce this approximate likelihood method here for discussions on bootstrapping, it should be noted that the initial simulation studies that follow, will use theoretical moments.

Algorithm 10 (Model based bootstrapping for standard errors). *The bootstrapping procedure follows.*

1. *Fit the Hawkes process using Algorithm 4 and retain the fitted parameter estimates. Note, this step is evaluated once for this procedure.*
2. *Simulate the Hawkes process with Algorithm 8, using the fitted model specifications from Step 1. The parameters specified in the simulation are the fitted parameter estimates from Step 1. The specified length of the simulate is equal to the length of the original data n from Step 1. This simulate will form the bootstrap replication, which is then treated like the original data.*
3. *Refit the model with Algorithm 4 to the bootstrap replication from Step 2, retaining the new parameter estimates as the bootstrap sample estimators.*
4. *Repeat Steps 2 and 3 many times, i.e., 1,000.*
5. *Estimate the standard errors by the sample standard error of the parameter estimates of the replicates.*

4.5 Simulation Experiments

The simulation studies that follow, demonstrate both the flexibility and the challenges of fitting the Hawkes process, with increasingly complex model formulations. We consider a range of parametric distributions for the marks, increasing complexity of the models to multivariate marks and extensions to the Hawkes process to use copula methods to model joint dependence between the marks. We also consider different formulations of the boost function, including adjustments required for the normalization of the boost function in the presence of joint dependence. For the estimation of the moments required to normalize the boost function, we use theoretical moments. Introduced in the previous section, we also consider an alternative *approximate likelihood method by decoupling the marks parameters*, which replaces the theoretically determined moments for normalizing the boost, by empirical estimates.

We highlight the challenges with utilizing the inverse of the Hessian matrix to calculate standard errors, proposing alternative methods to establish reliable standard error estimates. Studies using different optimization initializations and stability of parameter estimates across different sample sizes will provide guidance about the robustness of the Hawkes process.

For the simulation studies that follow, we present five case studies of increasing complexity of simulated data, to test the robustness of the log-likelihood method. For each of the simulation experiments, we simulate 1,000 replicates each with a sample size of $n = 1,000$, unless otherwise stated. We will fit the intensity process in which the marks are linearly boosted

$$h(\mathbf{x}; \psi) = 1 + \psi \mathbf{x}. \quad (4.31)$$

The intensity parameter specified in the simulation are: $\eta = 0.0020$, $\vartheta = 0.7000$, $\alpha = 0.0100$, and with a boost function parameter $\psi = 0.5$. Whilst this chapter is presented prior to our study of the score test and application of the Hawkes process to real data, for continuity reasons, it should be noted that this selection of simulation parameters was in fact guided by model fits conducted in Chapter 7.

Throughout all simulation studies, we assess the quality of the fit using an estimate of *bias* of the parameter estimates. This is estimated as a percentage difference of the mean of the parameter estimates across all replicates and their true value. For example, the bias of the immigration intensity is estimated by $(\mathbb{E}[\hat{\eta}_r] - \eta)/\eta$, where $r \in \{1, \dots, 1,000\}$ replicates. We assess the *variability ratio* of the parameter estimates, by the ratio of the standard deviation of the parameter estimates, divided by their true values. For example, the variability of the immigration intensity is estimated by $\sqrt{\mathbb{V}[\hat{\eta}_r]}/\eta$.

A brief outline of the studies follows.

1. **Simulated Hawkes process with a constant boost function.** We explore the estimation of the intensity parameters and robustness across different optimization initializations and sample sizes. This section introduces the challenges of using the Hessian matrix for the calculation of standard errors. A proposed bootstrap method is presented for the estimation of the standard errors for the calculation of p-values used to evaluate the parameter estimates.
2. **Simulated Hawkes process with a univariate mark.** We explore a range of parametric distributions for the mark, both continuous and discrete. We assess the quality of the parameter estimation, the sensitivity to the initialization of the optimizer and the robustness of parameter estimation across varying sample sizes. The parametric distributions and specified parameters for the simulations are:
 - Exponential distribution: $X \sim \text{Exp}(\lambda = 1.00)$;
 - Generalized Pareto distribution: $X \sim \text{GPD}(\zeta = 0.05, \delta = 1.00)$;
 - Poisson discrete distribution: $X \sim \text{Pois}(\mu = 1.50)$.
3. **Simulated Hawkes process with bivariate marks.** We assess parameter estimation for the Hawkes process with a pair of linearly boosted independent marks $X_i \in \mathbb{R}^2$ in (4.31), with the following distributions:
 - Exponential distribution: $X_i \sim \text{Exp}(\lambda = 1.00)$;
 - Generalized Pareto distribution: $X_i \sim \text{GPD}(\zeta = 0.05, \delta = 1.00)$;
 - Poisson discrete distribution: $X_i \sim \text{Pois}(\mu = 1.50)$.

The boost function for these studies are multiplicative in the functions of the marks. We also consider a case study of a boost function, which is additive in functions of the marks.

4. **Simulated Hawkes process with higher dimensional marks.** This study presents the parameter estimations across 1,000 replicates for a Hawkes process with four dimensional marks $X_i \in \mathbb{R}^4$, following the generalized Pareto distribution, $X_i \sim \text{GPD}(\zeta = 0.05, \delta = 1.00)$.
5. **Simulated Hawkes process with marks that are jointly dependent.** The case studies within this group consider three copula models to model the joint dependence, Gaussian, Gumbel and Clayton. The simulations are specified with a Spearman's rank correlation of $\rho_s = 0.50$. The marginal distribution for the marks $X_i \in \mathbb{R}^2$ are the exponential distribution, $X_i \sim \text{Exp}(\lambda = 1.00)$ and two formulations of a generalized Pareto distribution, $X_i \sim \text{GPD}(\zeta = 0.05, \delta = 1.00)$ and $X_i \sim \text{GPD}(\zeta = 0.49, \delta = 1.00)$. Finally, we compare the log-likelihood method with the alternative approximate likelihood method by decoupling the marks parameters.

4.5.1 Case Study 1: Hawkes process with a constant boost function

Figure 4.1 presents the parameter estimates for the intensity parameters of the Hawkes process. The boxplots presents 1,000 replicates each and the red horizontal line reflects the true value of each parameter. From Figure 4.1, we can see that the intensity function parameters are well estimated by the log-likelihood method, with an upward bias in the immigration intensity of only 1.96%, downward bias of the branching coefficient of -0.26% and an upward bias of the decay function parameter of 0.53%. We observe only a few outliers and a low variability ratio for each parameter: immigration intensity 0.0913, branching coefficient 0.0473 and decay function parameter 0.0871.

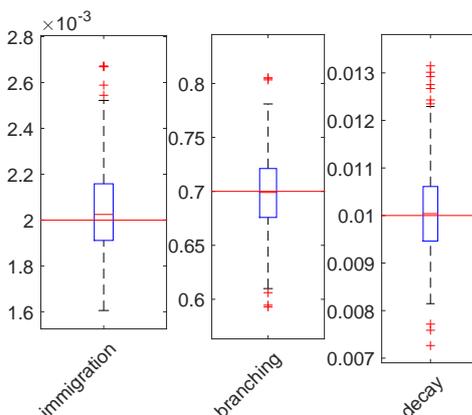


Figure 4.1: Boxplot of parameter estimates for a Hawkes process, with a constant boost function, a sample size $n = 1,000$, for 1,000 replicates. The solid red line represents the true value of each parameter.

Sensitivity to varying optimization initialization

In Section 4.2.4, we presented the challenge of the joint likelihood function being complex and that it may not be globally convex. There is a possibility that the log-likelihood estimation may result in a local maximum, rather than a global maximum. For that reason we assess the sensitivity of the optimization to different initializations. If the resulting estimates uncover the true value consistently across different initializations, this will provide confidence that we are achieving the global maximum.

Three initializations for the optimization are considered. The first initialization sets the parameters at the true value used in the simulation. The second and third introduce an additive stochastic perturbation to the true values. For the second initialization we introduce an error that represents 10% of the true value, randomized by a uniform random variable with support $U[-1, 1]$. For example, the initialization used for the immigration intensity $\eta_0^{(2)}$ is

$$\eta_0^{(2)} = \eta_0^{(1)} + \eta_0^{(1)} \times 0.1 \times Z, \quad (4.32)$$

where $Z \sim U[-1, 1]$. For the third initialization we introduce an error that represents 50% of the true value, and again randomized by a uniform random variable with support $U[-1, 1]$. For example, the initialization for the immigration intensity $\eta_0^{(3)}$ is

$$\eta_0^{(3)} = \eta_0^{(1)} + \eta_0^{(1)} \times 0.5 \times Z. \quad (4.33)$$

Regardless of the initial condition, the iterations converge to the same optimization value for the Hawkes process with a constant boost. Comparing the parameter estimates using the initialization of the true value, versus the introduction of 50% error on the initialization, gives a percentage deviation between the parameter estimates of, immigration intensity $6.0559e - 06\%$, branching coefficient $2.8554e - 06\%$ and decay function parameter $5.4280e - 06\%$.

Robustness across sample sizes

Figure 4.2 presents charts of the percentage bias and the variability ratio of the parameter estimates as the sample size increases, $n \in \{200, 400, \dots, 2,000\}$. The aim is to assess the robustness of the estimates across sample size. For all sample sizes, as the sample size increases, the bias decreases at a rate of roughly $1/\sqrt{n}$ for the immigration intensity and the decay function parameter. The branching coefficient has only a very small bias of -2.85% for the smallest sample size of $n = 200$ assessed, and the bias approaches zero as the sample size increases.

Figure 4.2(b) presents the variability of the parameter estimates, and similarly we observe a drop in the variability as the sample size increases. Sample size $n = 400$ and greater appears to produce consistent parameter estimates.

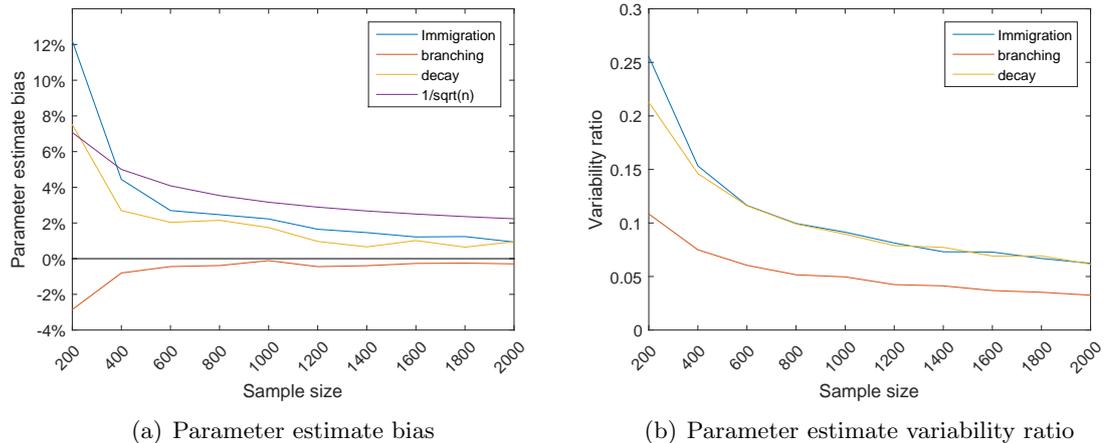


Figure 4.2: Parameter estimate bias and the variability ratio for a Hawkes process, with a constant boost function, as the sample size of the simulation (1,000 replicates each) increases $n \in \{200, 400, \dots, 2,000\}$.

Estimating standard errors

A common way to calculate standard errors of parameter estimates is by taking the inverse of the Hessian matrix. In Section 4.4 we introduced the method of optimization in MATLAB and the approximation of the Hessian, which is calculated during the optimization procedure. It was noted that the Hessian of the Lagrangian function may not match the true Hessian (MathWorks, 2018). To assess the reliability of the estimated Hessian matrix, we estimate the standard errors using the estimated Hessian for all replicates. Figure 4.3(a) displays the range of standard errors calculated via the approximate Hessian matrix. We see that the Hessian does not in fact provide a reasonable estimate of the standard error.

To assess this further, we consider the convergence rate, which is determined by how many iterations the optimizer has performed and this depends on the ratio of the smallest to largest eigenvalue of the Hessian. For symmetric positive matrices this is the condition number of the matrix. The median condition number across 1,000 replicates is $2.6e+6$, which shows that the matrix is ill-conditioned. For more complex models that we consider later, for example a Hawkes process with bivariate marks with a marginal generalized Pareto distribution, the mean condition number across a 1,000 replicates is significantly larger, $2.3e+14$. The inversion of the estimated Hessian will be numerically unstable, since the smallest singular values blow up the inversion. As a result, we observe very large and unreliable standard errors when trying to obtain the standard errors numerically via this classical approach, as shown in Figure 4.3(a). Whilst this is an appealing method to avoid developing formulae based Hessians (and information matrices), the Hessians computed numerically by MATLAB for this application are completely unreliable, being severely ill-conditioned, thus resulting in standard errors of parameter estimates that are not usable for inferential purposes.

In light of the unreliable estimated Hessian matrices, an alternative method for computing standard errors must be developed and used in practice for these complicated

marked processes. Although it is feasible to do with enough effort, we have already ruled out developing formulae for typical combinations of boost functions and marks densities relevant to our applications. This is due to the complexity of the derivations and programming of the results for each special case, which would make it not practically appealing or even feasible. We suggest the use of the bootstrap described in Section 4.4. Bootstrapping for the simulation experiments that follow contain only 200 replicates. For a more accurate estimation of standard errors, a larger simulation should be conducted.

Figure 4.3 displays the standard errors across 1,000 replicates for a Hawkes process with a constant boost function, using the inverse of the Hessian matrix obtained from the optimization of the likelihood function (Figure 4.3(a)) and the bootstrapping procedure described in Section 4.4 (Figure 4.3(b)). The bootstrapping procedure produces a considerable improvement to the estimate of the standard errors, providing more reliable and stable standard errors, but at the cost of substantial computational effort.

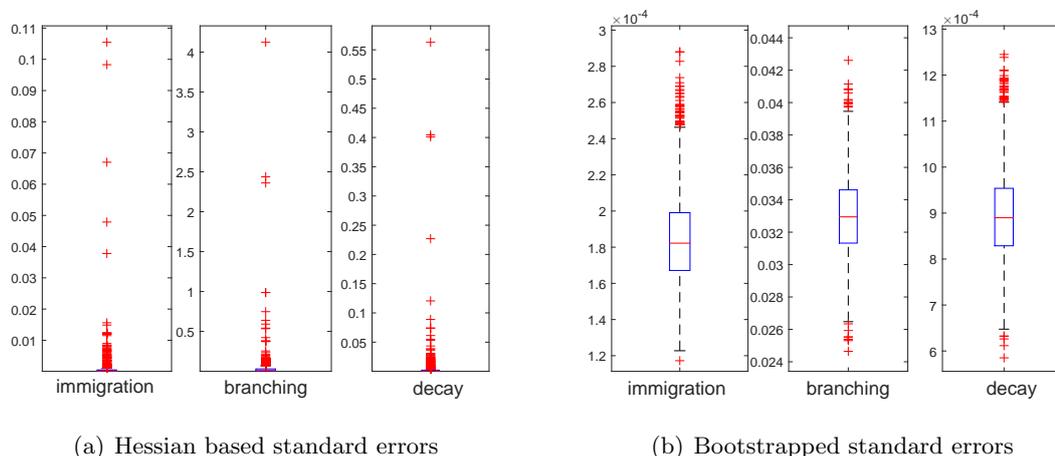


Figure 4.3: Boxplots of the standard errors of each estimated parameter for a Hawkes process, with a constant boost function, using the inverse of the Hessian obtained from the optimization procedure and the bootstrapping method. The bootstrapped standard error estimates each use 200 replicates. Each simulation contains 1,000 replicates.

4.5.2 Case Study 2: Hawkes process with a univariate i.i.d. mark with a range of parametric distributions

The simulation experiments that follow, consider a Hawkes process with a linear boost function in (4.31) and a univariate mark. We will consider a range of parametric distributions for the mark. The studies will assess the reliability of the parameter estimates via the log-likelihood method, the exploration of the initialization conditions for the optimization procedure and the impact of sample size on parameter estimates.

Exponential distribution

Figure 4.4 displays the parameter estimates for a Hawkes process with a univariate exponentially distributed mark. Consistent with the previous example, the parameters for the intensity process are well estimated, with an upward bias in the immigration intensity of

only 1.90%, downward bias of the branching coefficient of -0.25% and an upward bias of the decay parameter of 1.11%. The scale parameter of the marks distribution has a low bias of -0.006% . The boost parameter has a larger bias of 5.18% and presents a higher variability ratio of 0.4262. The boost parameter is statistically significant for the majority of the replicates, however there are some cases where this does not hold true.

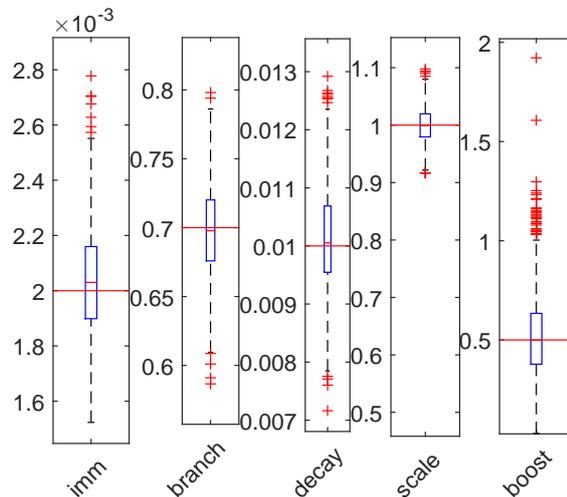


Figure 4.4: Boxplot of parameter estimates for a Hawkes process, with a sample size of $n = 1,000$, for 1,000 replicates. The mark is exponentially distributed $X \sim \text{Exp}(\lambda = 1)$, with a linear boost and with estimated theoretical moments. The solid red line represents the true value of each parameter.

With the introduction of marks into the log-likelihood function, increasing the randomization of the initial conditions, still produces reliable results in the optimization. Comparing the parameter estimates using the initialization of the true value versus the introduction of 50% error on the initialization, we observe only a tiny percentage deviation between the parameters estimated with each initialization, immigration intensity $8.5573E - 07\%$, branching coefficient $-1.8202E - 06\%$, decay function parameter $-7.0445E - 06\%$, mark parameter $5.10793E - 06\%$ and boost parameter $1.3787E - 04\%$.

Figure 4.5) shows an increase in the sample size $n \in \{200, 400, \dots, 2,000\}$, presents a consistent level of bias for the intensity parameters compared with the constant boost case. However, the boost function parameter has a large bias of 54% for the smallest sample size of $n = 200$. This improves significantly as the sample size increases. We observe a similar pattern in the variability ratio of the parameter estimates relative to their true values. Even with large sample sizes, the variability ratio of the boost function parameter remains much greater than the other parameters, however, it drops to 0.2743 for the largest sample size considered.

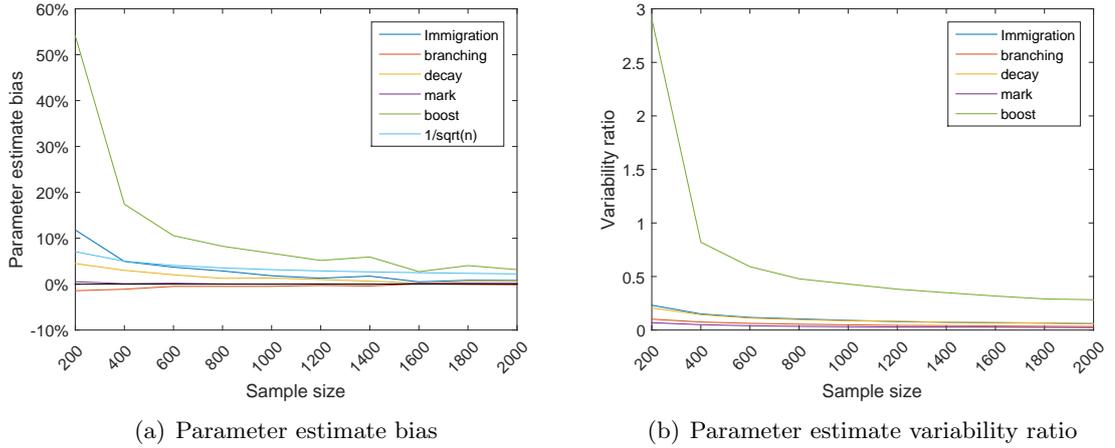


Figure 4.5: Parameter estimate bias and variability ratio for a Hawkes process, with a linear boost function, with an exponentially distributed mark $X \sim \text{Exp}(\lambda = 1)$, as the sample size of the simulation (1,000 replicates each) increases $n \in \{200, 400, \dots, 2,000\}$.

Generalized Pareto distribution

From Figure 4.6, we can see that the parameter estimates for the Hawkes process, with univariate generalized Pareto distributed marks, are well estimated with the intensity and mark parameters having a low bias from the true values. The bias for each is, immigration intensity 2.28%, branching coefficient -0.53% , decay function parameter 1.60%, shape parameter -2.73% and scale parameter 0.07%. The boost function parameter has a higher bias of 6.62%, which is consistent with what we observed for the model with an exponentially distributed mark.

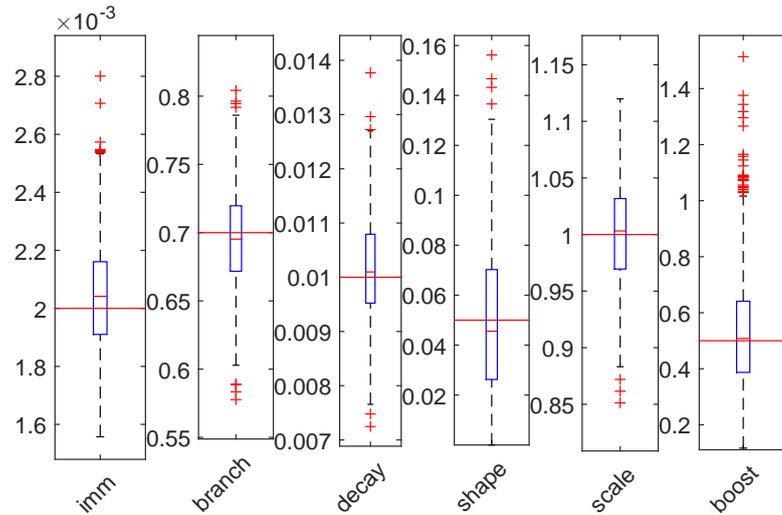


Figure 4.6: Boxplot of parameter estimates for a Hawkes process, with a sample size of $n = 1,000$, for 1,000 replicates. The mark is GPD $X \sim \text{GPD}(\zeta = 0.05, \delta = 1.00)$, with a linear boost and with estimated theoretical moments. The solid red line represents the true value of each parameter.

Varying the initial conditions used in the optimization procedure for the Hawkes process with a linear boosted generalized Pareto distributed mark, has almost no effect on the parameter estimates. The percentage deviation between the parameter estimates using initial conditions at the true value and those with an error of 50% are, immigration intensity $-1.1410E - 04\%$, branching coefficient $-1.0000E - 05\%$, decay function parameter $-8.3586E - 05\%$, shape parameter $5.009E - 03\%$, scale parameter $-3.6148E - 04\%$ and boost parameter $4.116E - 04\%$.

As we increase the sample size, displayed in Figure 4.7, the intensity and boost function parameter estimates present a similar decrease in bias and variability ratio that was observed in the previous case study. The generalized Pareto distribution shape parameter displays a much larger variability (Figure 4.7(b)) than the other parameters. The level of bias for the shape parameter shows only a limited reduction as the sample size increase from $n = 400 \rightarrow n = 2,000$ (Figure 4.7(a)).

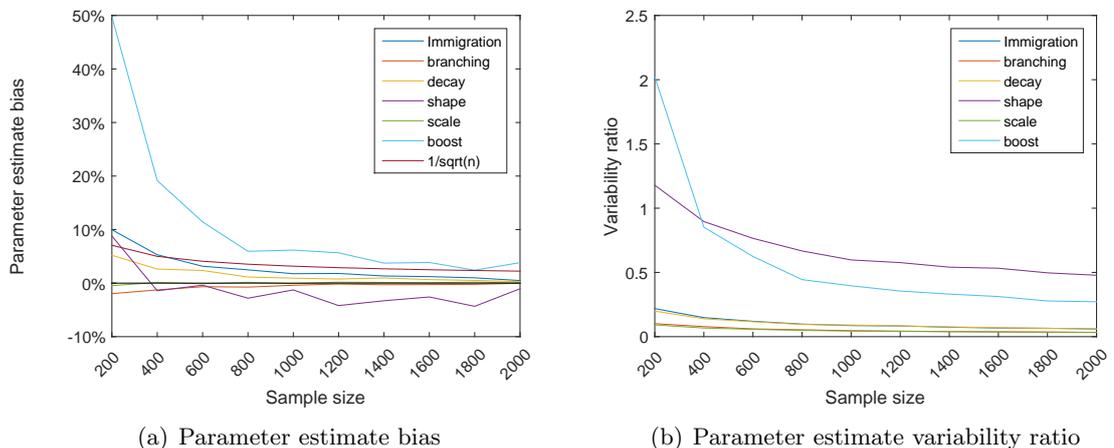


Figure 4.7: Parameter estimate bias and variability ratio for a Hawkes process, with a linear boost function, with a GPD mark $X \sim \text{GPD}(\zeta = 0.05, \delta = 1.00)$, as the sample size of the simulation (1,000 replicates each) increases $n \in \{200, 400, \dots, 2,000\}$.

As presented in Section 2.3.3, and researched in Grimshaw (1993), the shape parameter for the generalized Pareto distribution can be challenging to estimate. To address this, one should consider re-parametrizing $(\zeta, \delta) \rightarrow (\zeta, \tau) = \left(\zeta, \frac{\zeta}{\delta}\right)$ to compute the log-likelihood estimates. We then substitute τ in the expression below, into the log-likelihood function for the generalized Pareto distribution, creating the profile log-likelihood of τ

$$\hat{\zeta} = \frac{1}{T} \sum_{i=1}^J \ln(1 - \tau x_i).$$

The re-parametrized log-likelihood is given by

$$\ln l(\tau; \mathbf{X}) = -T - \sum_{i=1}^T \ln(-\tau X_i) - T \ln \left(-\frac{1}{T} \sum_{i=1}^T \frac{1}{\tau} \ln(1 - \tau X_i) \right). \quad (4.34)$$

This log-likelihood is estimated in two steps, with the first being the numerical computa-

tion, where the likelihood function is maximized subject to $\tau < 1/x_{(T)}$ and $\zeta \geq -1$. The estimated $\hat{\tau}$ is used to calculate explicitly $\hat{\zeta} = -\frac{1}{T} \sum_{i=1}^T \ln(1 - \hat{\tau}x_i)$. The final step is to solve for the original parametrization $\hat{\delta} = \frac{\hat{\zeta}}{\hat{\tau}}$.

This previously described method is for the estimation of the generalized Pareto distributed marginal parameters (Section 2.3.3). For estimation of the Hawkes process with generalized Pareto distributed marks, we need to incorporate this change in parametrization into the likelihood function for the Hawkes process. The numerical maximization will be dependent on only one parameter τ for the generalized Pareto distributed marks. The calculation of the shape parameter ζ and conversion back to the original parametrization will be required for the estimation of the theoretical moments, which are required for the boost function. For this hybrid method, the properties would need to be studied in the context of the log-likelihood method for the Hawkes process. For this research, given the shape parameter is on average estimated well with only a small downward bias of -2.73% , we will continue with the one stage estimation procedure. The incorporation of the hybrid method into the Hawkes process, should however be explored in future research.

Poisson discrete distribution

Figure 4.8 displays the parameter estimates for a Hawkes process with a linear boost function with a univariate Poisson distributed mark. The bias in the parameter estimates are, immigration intensity 1.71%, branching coefficient -0.27% , decay function parameter 0.73%, rate parameter -0.02% and boost parameter 7.33%.

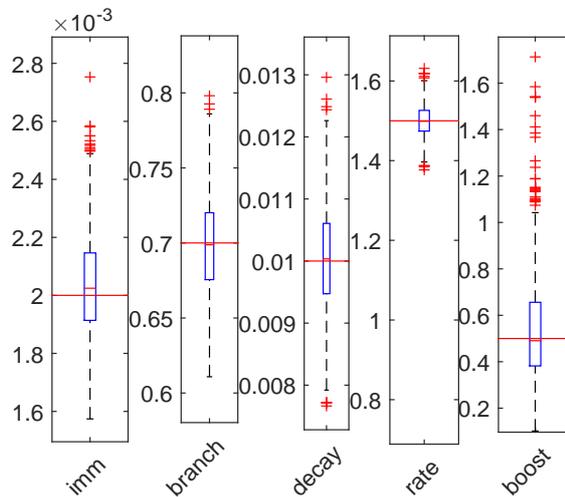


Figure 4.8: Boxplot of parameter estimates for a Hawkes process, with a sample size of $n = 1,000$, for 1,000 replicates. The mark is Poisson distributed $X \sim \text{Pois}(\mu = 1.50)$, with a linear boost and with estimated theoretical moments. The solid red line represents the true value of each parameter.

The bias in all parameter estimates decreases as the sample size increases (Figure 4.9(a)), however for sample size $n = 400$ and lower, the boost function parameter estimates result in some extreme outliers. Sample sizes $n = 800$ and greater, produce reasonable boost parameter estimates. We observe a similar pattern for the variability

ratio, with sample sizes of $n = 800$ and greater resulting in less variability of the estimates (Figure 4.9(b)).

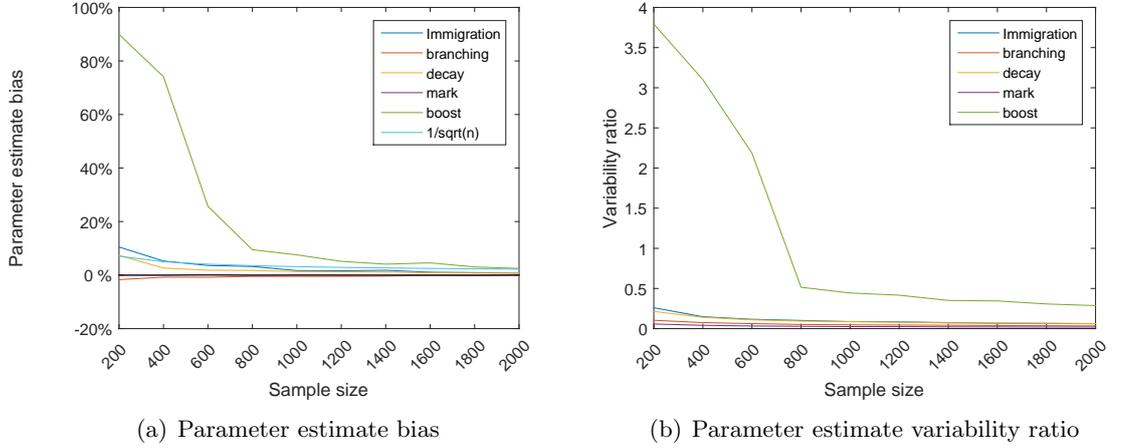


Figure 4.9: Parameter estimate bias and variability ratio for a Hawkes process, with a linear boost function, with a Poisson distributed mark $X \sim \text{Pois}(\mu = 1.50)$, as the sample size of the simulation (1,000 replicates each) increases $n \in \{200, 400, \dots, 2,000\}$.

4.5.3 Case Study 3: Hawkes process with independent bivariate marks with a range of parametric distributions

The simulation experiments that follow, assess the parameter estimation for the Hawkes process with a pair of linearly boosted independent marks $X_i \in \mathbb{R}^2$ in (4.31). The following parametric marks distributions are considered, exponential $X_i \sim \text{Exp}(\lambda = 1.00)$ (Figure 4.10), generalized Pareto distribution $X_i \sim \text{GPD}(\zeta = 0.05, \delta = 1.00)$ (Figure 4.11) and Poisson $X_i \sim \text{Pois}(\mu = 1.50)$ (Figure 4.12). For each model we simulate 1,000 replicates of sample size $n = 1,000$ each.

Table 4.3 presents the parameter estimate bias and variability ratio for each model. Comparing the bias of the estimated intensity parameters $\theta = \{\eta, \vartheta, \alpha\}$ for the univariate versus the bivariate models, the bias has not changed with the introduction of an additional mark for the three models considered. The boost parameter bias is very similar, approximately 5 – 6% for both the univariate and bivariate models with exponentially distributed marks and generalized Pareto distributed marks. The Hawkes process with Poisson distributed marks shows an increase in the bias of the boost parameter estimate from 7.33% in the univariate model, to 9.27% and 9.47% for the bivariate case.

The variability ratio is closely aligned across all three models for the intensity parameters and the boost parameters. The exception is the shape parameter for the generalized Pareto distribution, which shows a much larger variation in estimation of 0.6124 and 0.6258 and visually observable in Figure 4.11. However, this is consistent with our earlier findings in the univariate case.

In summary, these findings are suggestive that an increase in marks dimension does not affect the reliability of the parameter estimates for a range of models considered, with only a small increase in the bias of the boost parameter estimates for the model

with Poisson distributed marks. As we will note in Section 4.5.5, the increased bias in the boost parameter estimation for the Poisson distribution may be reduced by using the approximate likelihood method by decoupling the marks parameters.

Table 4.3: Parameter estimate bias (%) and variability ratio for three bivariate Hawkes process models, with a linear boost function and marks $X_i \in \mathbb{R}^2$ distributed, $X_i \sim \text{Exp}(\lambda = 1.00)$, $X_i \sim \text{GPD}(\zeta = 0.05, \delta = 1.00)$ and $X_i \sim \text{Pois}(\mu = 1.50)$, respectively. The simulations each have 1,000 replicates, with a sample size of $n = 1,000$ each.

Estimated parameter bias (%)									
Marks dist.	imm	branch	decay	mark 1		mark 2		boost 1	boost 2
				para. 1	para. 2	para. 1	para.2		
Exp	2.24%	-0.29%	1.31%	-0.10%		0.01%		5.22%	6.29%
GPD	1.88%	-0.41%	0.51%	-7.18%	0.38%	-1.18%	0.10%	6.84%	5.16%
Pois	1.62%	-0.66%	0.67%	-0.01%		-0.14%		9.27%	9.47%

Variability ratio									
Marks dist.	imm	branch	decay	mark 1		mark 2		boost 1	boost 2
				para. 1	para. 2	para. 1	para.2		
Exp	0.0867	0.0495	0.0886	0.0308		0.0309		0.4211	0.3948
GPD	0.0887	0.0507	0.0900	0.6124	0.0440	0.6258	0.0458	0.3955	0.3872
Pois	0.0871	0.0500	0.0872	0.0260		0.0258		0.4544	0.4685

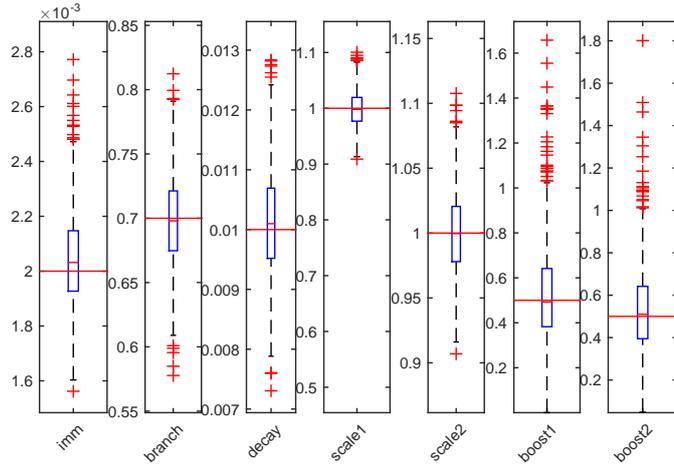


Figure 4.10: Boxplot of parameter estimates for a Hawkes process, with linearly boosted, bivariate exponentially distributed marks, where $X_i \in \mathbb{R}^2$ and $X_i \sim \text{Exp}(\lambda = 1.00)$, with a sample size of $n = 1,000$, for 1,000 replicates. The solid red line represents the true value of each parameter.

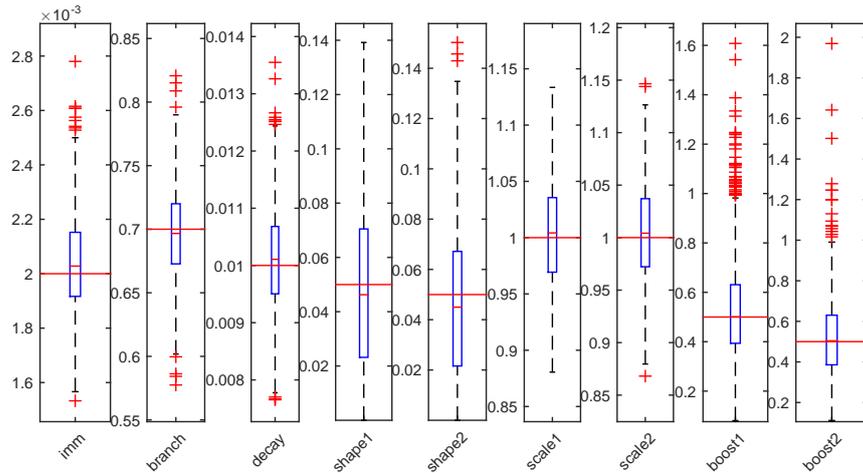


Figure 4.11: Boxplot of parameter estimates for a Hawkes process, with linearly boosted, bivariate GPD marks, where $X_i \in \mathbb{R}^2$ and $X_i \sim \text{GPD}(\zeta = 0.05, \delta = 1.00)$, with a sample size of $n = 1,000$, for 1,000 replicates. The solid red line represents the true value of each parameter.

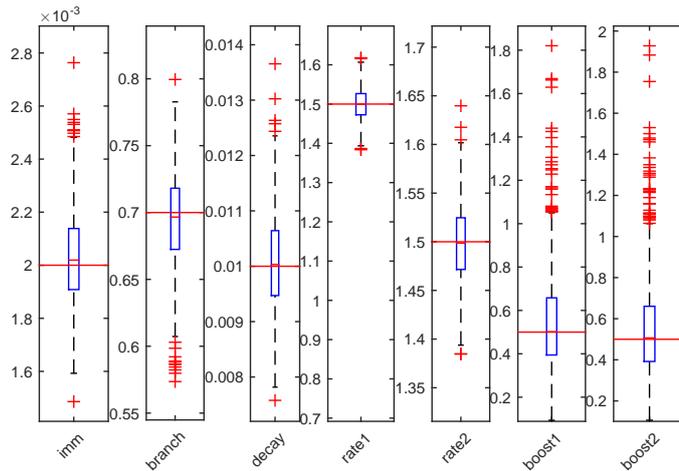


Figure 4.12: Boxplot of parameter estimates for a Hawkes process, with linearly boosted, bivariate Poisson distributed marks, where $X_i \in \mathbb{R}^2$ and $X_i \sim \text{Pois}(\mu = 1.50)$, with a sample size of $n = 1,000$, for 1,000 replicates. The solid red line represents the true value of each parameter.

Investigating additive and multiplicative boost functions

The mark vector and associated boost function accommodates many standard parametric distributions. Recall, that by starting with a function $h(\mathbf{X}, \psi)$, the boost function in (4.4) can be combined multiplicatively in (4.6) or additively in (4.5). We also present a third case, which is referred to as ‘jointly additive’, with the functional form presented below.

For the illustration that follows, we consider a Hawkes process with linearly boosted, bivariate marks $X_i \in \mathbb{R}^2$, which have a generalized Pareto distribution $X_i \sim \text{GPD}(\zeta = 0.05, \delta = 1.00)$. The boost function combines the impact of the individual marks in the following ways:

1. Multiplicatively,

$$g(\mathbf{x}; \phi, \psi) = g_1(x_1; \phi_1, \psi_1)g_2(x_2; \phi_2, \psi_2) \\ = \frac{1 + \psi_1 x_1 + \psi_2 x_2 + \psi_1 \psi_2 x_1 x_2}{1 + \psi_1 \mathbb{E}[X_1] + \psi_2 \mathbb{E}[X_2] + \psi_1 \psi_2 \mathbb{E}[X_1] \mathbb{E}[X_2]};$$

2. Additively,

$$g(\mathbf{x}; \phi, \psi) = \frac{g_1(x_1; \phi_1, \psi_1) + g_2(x_2; \phi_2, \psi_2)}{2} \\ = \left[\frac{1 + \psi_1 x_1}{1 + \psi_1 \mathbb{E}[X_1]} + \frac{1 + \psi_2 x_2}{1 + \psi_2 \mathbb{E}[X_2]} \right] / 2;$$

3. Jointly additively,

$$g(\mathbf{x}; \phi, \psi) = \frac{1 + \psi_1 x_1 + \psi_2 x_2}{1 + \psi_1 \mathbb{E}[X_1] + \psi_2 \mathbb{E}[X_2]}.$$

Figure 4.13 displays the kernel densities for three replicate boost functions for each of the three different methods of combining the boost, multiplicatively, additively and jointly additively. The impact of heavy tails will be greater for the multiplicative form of combining the boost functions. The additive boosts will tend to concentrate more around the mean, compared with the multiplicative form, which is observed in Figure 4.13. As the marks drop below their mean, the multiplicative form of combining the boosts will result in a lower value than the additive, assuming all else is equal. This effect can be seen in the spike in the kernel densities just below the true mean for the additive cases in Figure 4.13, and the higher concentration at the beginning of the kernel density for the multiplicative case.

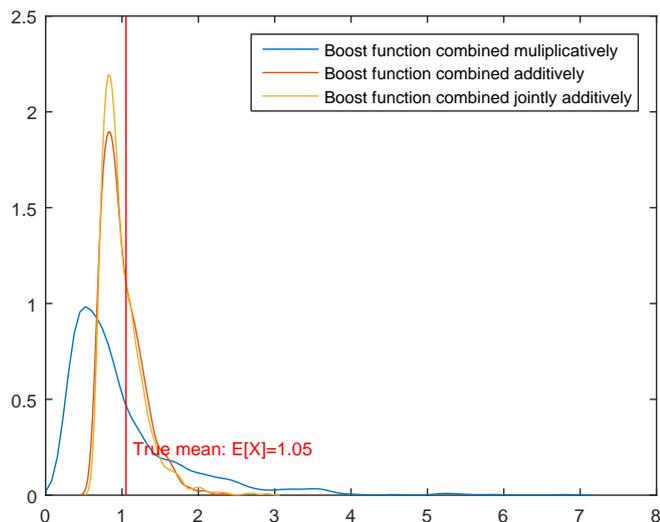


Figure 4.13: Kernel densities of the boost functions for a Hawkes process, with linearly boosted, bivariate marks $X_i \in \mathbb{R}^2$, distributed $X_i \sim \text{GPD}(\zeta = 0.05, \delta = 1.00)$. Each replicate combines the boost function multiplicatively, additively and jointly additively. The sample size of each replicate is $n = 1,000$.

Using the Hawkes process specifications noted above, this study is extended across three simulations each with a 1,000 replications, with a sample size of $n = 1,000$. Table 4.4 presents the bias of the parameter estimates using the three different methods of combining the boost functions. In each case, the same method is applied within the simulation of data and the fitting of the Hawkes process. The effects noted in the boost functions in Figure 4.13, leads to unreliable parameter estimates observed in Figure 4.14 and 4.15, and therefore a significant degradation in the reliability of the parameter estimates. The bias for the boost function parameter estimates increases substantial to 22.81% and 18.18% for the additive boost, and further to 29.31% and 30.58% for the jointly additive boost. The variability in the boost function parameter estimates has increased to an average of 1.5 times the true value. This provides evidence that the boost functions should be combined multiplicatively.

Table 4.4: Parameter estimate bias (%) and variability ratio for three bivariate Hawkes process models, with linear boost functions, with marks $X_i \in \mathbb{R}^2$ distributed $X_i \sim \text{GPD}(\zeta = 0.05, \delta = 1.00)$. The three different models combine the boost functions, multiplicatively, additively and jointly additively. The simulations each have 1,000 replicates, with a sample size of $n = 1,000$ each.

Estimated parameter bias (%)									
Boost combine	imm	branch	decay	parameter 1		parameter 2		boost 1	boost 2
				mark 1	mark 2	mark 1	mark 2		
Multiple	1.88%	-0.41%	0.51%	-7.18%	-1.18%	0.38%	0.10%	6.84%	5.16%
Additive	2.11%	-0.28%	1.06%	-4.47%	-1.20%	0.22%	0.24%	22.82%	18.81%
Joint. Add.	2.52%	-0.73%	1.66%	-5.33%	-3.28%	-0.21%	0.07%	29.31%	30.58%

Variability ratio									
Boost combine	imm	branch	decay	parameter 1		parameter 2		boost 1	boost 2
				mark 1	mark 2	mark 1	mark 2		
Multiple	0.0887	0.0507	0.0900	0.6124	0.6258	0.0440	0.0458	0.3955	0.3872
Additive	0.0943	0.0474	0.0873	0.6090	0.6403	0.0446	0.0436	1.4604	1.6228
Joint. Add.	0.0938	0.0474	0.0851	0.6314	0.6231	0.0463	0.0438	1.4785	1.1309

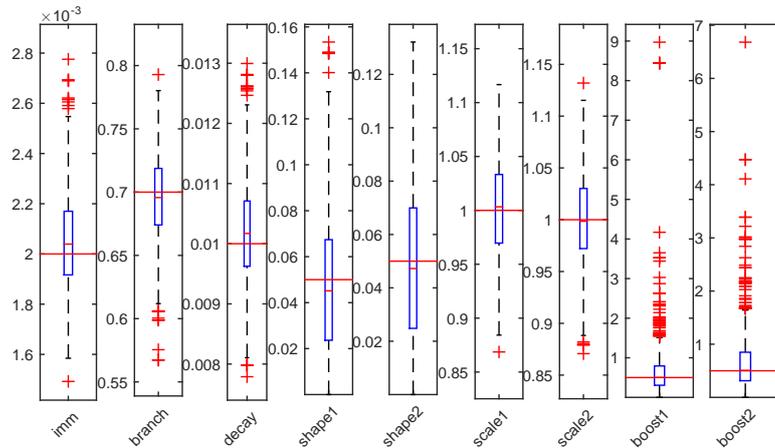


Figure 4.14: Boxplot of parameter estimates for a Hawkes process, with linearly boosted, bivariate GPD marks, where $X_i \in \mathbb{R}^2$ and $X_i \sim \text{GPD}(\zeta = 0.05, \delta = 1.00)$, combined with an *additive* boost function. The sample size is $n = 1,000$, for 1,000 replicates. The solid red line represents the true value of each parameter.

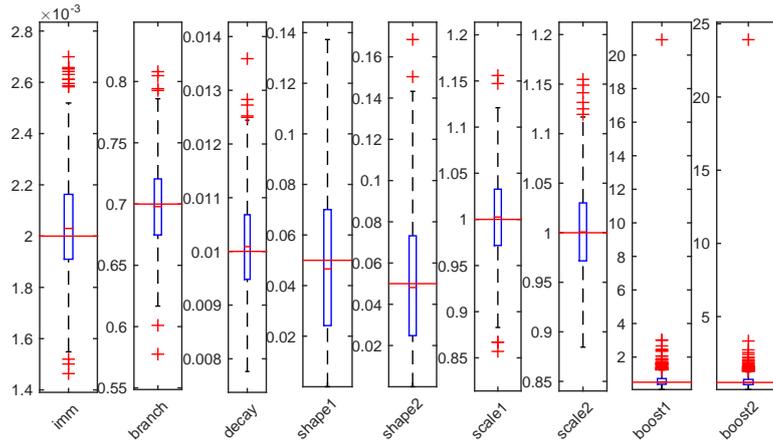


Figure 4.15: Boxplot of parameter estimates for a Hawkes process, with linearly boosted, bivariate GPD marks, where $X_i \in \mathbb{R}^2$ and $X_i \sim \text{GPD}(\zeta = 0.05, \delta = 1.00)$, combined with a *jointly additive* boost function. The sample size is $n = 1,000$, for 1,000 replicates. The solid red line represents the true value of each parameter.

4.5.4 Case Study 4: Hawkes process with marks in higher dimension

To further assess the robustness of the estimation methods, we consider a Hawkes process with linearly boosted, four dimensional marks $X_i \in \mathbb{R}^4$, with a generalized Pareto distribution $X_i \sim \text{GPD}(\zeta = 0.05, \delta = 1.00)$. Figure 4.16 presents the boxplots for each parameter estimate simulated across 1,000 replicates of sample size $n = 1,000$ each.

For comparison, we will present the results for the Hawkes process with generalized Pareto distributed bivariate marks (Table 4.3) in brackets. The bias in the intensity parameters are, immigration intensity 1.71% (1.88%), branching coefficient -0.33% (-0.41%) and decay function parameter 0.72% (1.31%). The boost parameter estimate bias is $\psi_i \in \{5.42\%, 7.32\%, 6.24\%, 6.64\%\}$, which is closely aligned with the bias presented in the Hawkes process with bivariate marks (Table 4.3). The bias of the marks distribution parameter estimates is also similar to what we observed for the bivariate case, however the highest observed bias for the shape parameter in this study is 4.56%. The variability ratio of the true values is also aligned to the variability ratios we observed in both the univariate and bivariate case studies. It is clear from this study that there is no degradation in the reliability of the parameter estimates as we increase the Hawkes process to include higher dimensional marks.

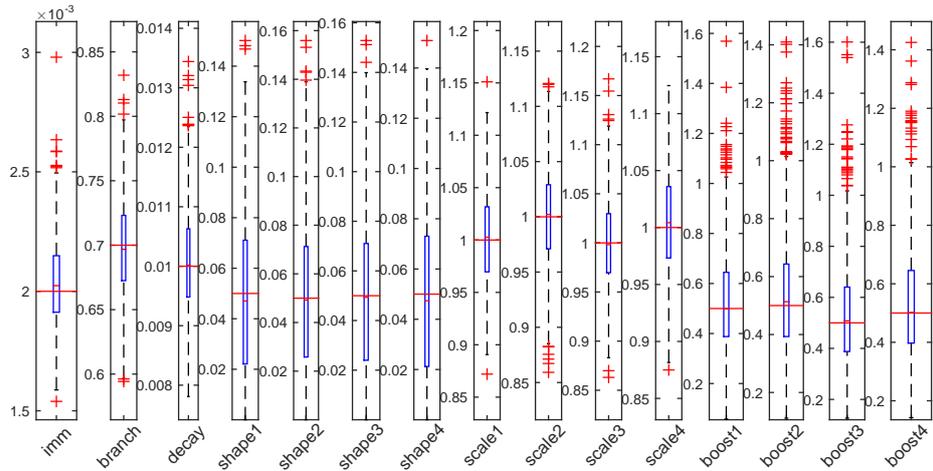


Figure 4.16: Boxplot of parameter estimates for a Hawkes process, with linearly boosted, four dimensional GPD marks, where $X_i \sim \text{GPD}(\zeta = 0.05, \delta = 1.00)$, with a sample size of $n = 1,000$, for 1,000 replicates. The solid red line represents the true value of each parameter.

4.5.5 Case Study 5: Hawkes process with jointly coupled marks

In the case studies that follow, we consider bivariate marks that are jointly dependent and with a linear boost function that is multiplicatively combined as in (4.2.1). The adjustment made to the normalization of a linear boost function in the presence of joint dependence, requires that the second moments exist. For marks that follow a generalized Pareto distribution, the shape parameter is required to be $\zeta < 0.5$ for second moments to exist. We will present a light tailed example, followed by a detailed study of the implications of the second moment requirement in the heavy tailed case, and the numerical challenges of fitting a Hawkes process this presents.

We will introduce a practical approach of replacing theoretical moments with empirical moments in the calculation of the normalization of the boost function required to define the process intensity $\lambda(t)$, appearing in the first two components of the log-likelihood in (4.15). That is, we use $\tilde{g}(\mathbf{x}; \psi) = \frac{h(\mathbf{x}; \psi)}{\bar{h}(\mathbf{x}; \psi)}$, where $\bar{h}(\mathbf{x}; \psi)$ is the function h evaluated using empirical moments in place of theoretical moments. This method, which we introduced briefly in Section 4.4, is an **approximate likelihood method by decoupling the marks parameters**, as it decouples the marks density parameters component of the likelihood from the Hawkes process component of the likelihood. This leads to nice outcomes such as the boost normalization for the intensity process not relying on the parametric specification of the marks. In addition, the strict adherence for moments existing is relaxed. There are also computational advantages for this method, with the decoupling leading to faster evaluation of the likelihood function, and as described in Section 4.4, a superior method for estimating standard errors via bootstrapping.

Joint dependence modelled via a Gaussian copula

Table 4.5 presents the parameter estimate bias and variability ratio for the four simulation studies. The light tailed case is represented by a Hawkes process with a linear boost and exponentially distributed marks. We present both the bias from the study above, where the marks are independent, and this study where the marks are jointly coupled via a Gaussian copula. The results presented here are consistent with the bias that was originally observed in the independence case study, with exception to the boost parameters, which show an increase from 5.22% and 6.29%, to 9.67% and 12.70% in bias for the model with jointly dependent marks. However, as presented in Figure 4.17, this may be due to some outliers.

Table 4.5 shows an inflation in the bias of the boost function parameter 1 estimate of 12.54% for the model with marks that are jointly coupled by a Gaussian copula. This is due to a concentration of outliers above $\psi = 1.4$, shown in Figure 4.18. There is a slight increase in the variability ratio for the boost parameter estimates, from 0.3955 and 0.3872, to 0.5582 and 0.4584. All other results are consistent with what was observed in the model with independent marks.

Table 4.5: Parameter estimate bias (%) and variability ratio for four bivariate Hawkes process models, with a linear boost function, with marks $X_i \in \mathbb{R}^2$ distributed: $X_i \sim \text{Exp}(\lambda = 1.00)$ (labelled ‘Exp’); marginal $X_i \sim \text{Exp}(\lambda = 1.00)$ and jointly coupled via a *Gaussian copula* where $\rho_s = 0.50$ (labelled ‘Exp-Cop’); $X_i \sim \text{GPD}(\zeta = 0.05, \delta = 1.00)$ (labelled ‘GPD’); and marginal $X_i \sim \text{GPD}(\zeta = 0.05, \delta = 1.00)$ and jointly coupled via a *Gaussian copula* where $\rho_s = 0.50$ (labelled ‘GPD-Cop’). The simulations each have 1,000 replicates, with a sample size of $n = 1,000$ each.

Estimated parameter bias (%)									
Marks dist.	imm	branch	decay	mark 1		mark 2		boost 1	boost 2
				para. 1	para. 2	para. 1	para.2		
Exp	2.24%	-0.29%	1.31%	-0.10%		0.01%		5.22%	6.29%
Exp-Cop	2.26%	-0.56%	1.21%	-0.06%		0.10%		9.67%	12.70%
GPD	1.88%	-0.41%	0.51%	-7.18%	0.38%	-1.18%	0.10%	6.84%	5.16%
GPD-Cop	2.21%	-0.22%	0.56%	-3.68%	0.23%	-2.61%	0.21%	12.54%	5.99%
Variability ratio									
Marks dist.	imm	branch	decay	mark 1		mark 2		boost 1	boost 2
				para. 1	para. 2	para. 1	para.2		
Exp	0.0867	0.0495	0.0886	0.0308		0.0309		0.4211	0.3948
Exp-Cop	0.0914	0.0577	0.0874	0.0481		0.0483		0.5581	0.5292
GPD	0.0887	0.0507	0.0900	0.6124	0.0440	0.6258	0.0458	0.3955	0.3872
GPD-Cop	0.0844	0.0505	0.0819	0.6224	0.0438	0.6042	0.0427	0.5582	0.4584

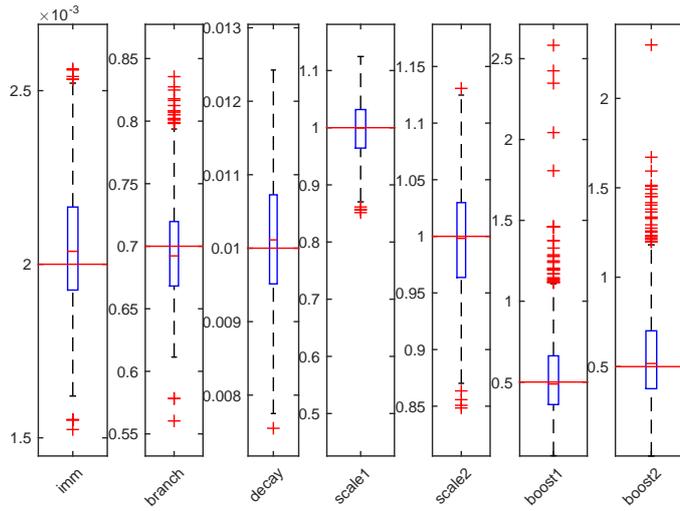


Figure 4.17: Boxplot of parameter estimates for a Hawkes process, with linearly boosted, bivariate exponentially distributed marks $\text{Exp}(\lambda = 1.00)$, with joint dependence modelled by a *Gaussian copula*, where $\rho_s = 0.50$. The sample size is $n = 1,000$, for 1,000 replicates. The solid red line represents the true value of each parameter.

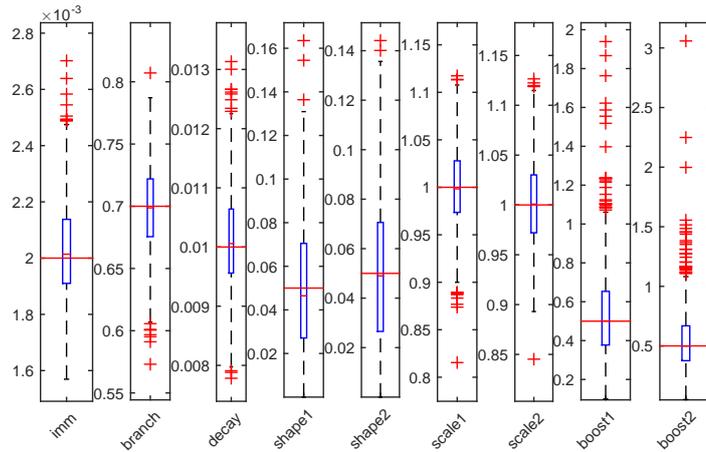


Figure 4.18: Boxplot of parameter estimates for a Hawkes process, with linearly boosted, bivariate GPD marks $\text{GPD}(\zeta = 0.05, \delta = 1.00)$, with joint dependence modelled by a *Gaussian copula*, where $\rho_s = 0.50$. The sample size is $n = 1,000$, for 1,000 replicates. The solid red line represents the true value of each parameter.

Robustness in the presence of joint dependence

Intuitively, when heavy tailed marks are jointly dependent and combined multiplicatively, the tail events are more likely to occur together, due to the correlation between the marks. This results in a greater impact on the intensity process compared with bivariate independent marks with the same marginal parameter estimates.

Recall, the normalization of a linear boost function with independent bivariate marks

$$\mathbb{E}[h(\mathbf{X}; \psi)] = 1 + \psi_1 \mathbb{E}[X_1] + \psi_2 \mathbb{E}[X_2] + \psi_1 \psi_2 \mathbb{E}[X_1] \mathbb{E}[X_2]. \quad (4.35)$$

The normalization of a linear boost function with jointly dependent bivariate marks

$$\mathbb{E}[h(\mathbf{X}; \psi)] = 1 + \psi_1 \mathbb{E}[X_1] + \psi_2 \mathbb{E}[X_2] + \psi_1 \psi_2 \mathbb{E}[X_1 X_2], \quad (4.36)$$

with the full specification for the normalization with a Gaussian copula model in (4.29).

Figure 4.19 presents a much larger boost function for the jointly dependent marks, compared with the independent marks, despite the increase in normalization. This is consistent with our intuition of the impacts of jointly dependent marks on the Hawkes intensity function.

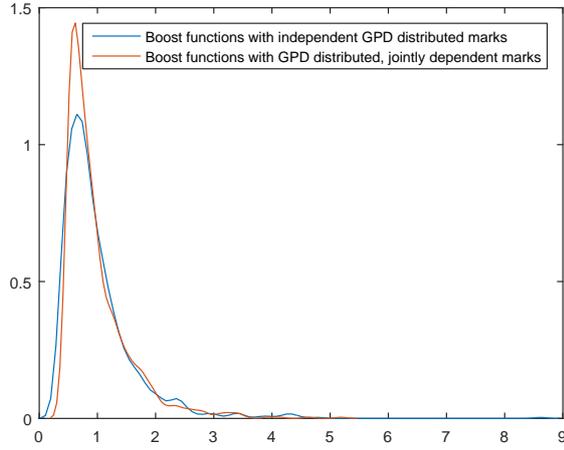


Figure 4.19: Kernel densities of the boost functions from a Hawkes process, with linearly boosted marks $X_i \in \mathbb{R}^2$ distributed $X_i \sim \text{GPD}(\zeta = 0.05, \delta = 1.00)$. The first case considers independent marks, the second case considers jointly coupled marks via a *Gaussian copula*, where $\rho_s = 0.50$. The densities have a sample size of $n = 1,000$.

We now consider the robustness of the estimation method in the presence of joint dependence by specifying the generalized Pareto distribution shape parameter $\zeta = 0.49$, which is very close to the boundary of second moments not existing. Figure 4.20 presents the normalization of the boost function for combinations of shape $\zeta \in \{0, 0.04, \dots, 0.48\}$ and scale $\delta \in \{0.01, 0.06, \dots, 0.80\}$, for a fixed $\psi = 0.5$. As the shape parameter approaches 0.5, there is a significant increase in the normalization in the case of joint dependent marks due to the variance increasing significantly.

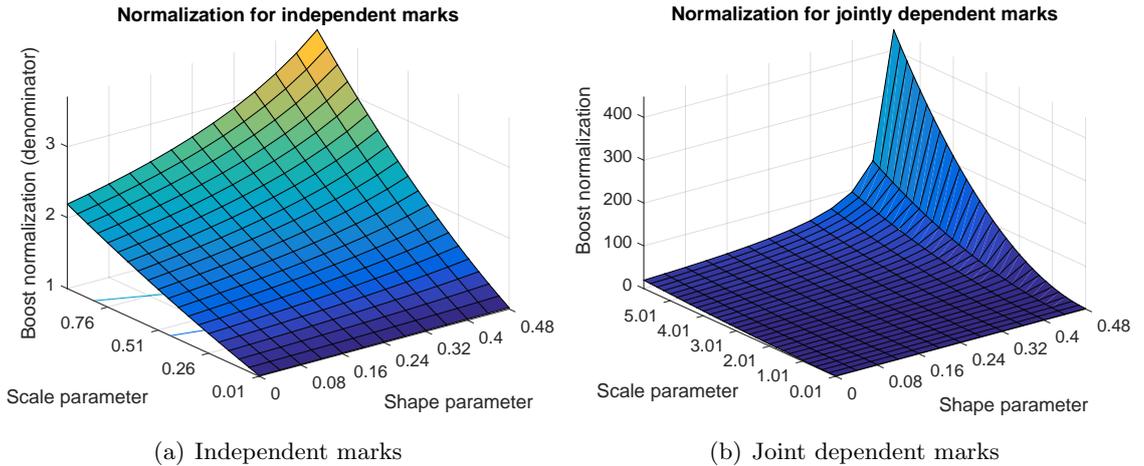


Figure 4.20: Normalization (denominator) of the boost function, assuming GPD marks, for combinations of shape $\zeta \in \{0, 0.04, \dots, 0.48\}$ and scale $\delta \in \{0.01, 0.06, \dots, 0.80\}$, for a fixed $\psi = 0.5$. The first plot presents the case of independent marks and the second plot presents the normalization for jointly dependent marks, where $\rho_s = 0.50$.

Considering marks $X_i \in \mathbb{R}^2$, that are generalized Pareto distributed $X_i \sim \text{GPD}(\zeta = 0.49, \delta = 1.00)$, the theoretical first and second central moments are $\mu_1 = 1.96$ and $\mu_2 = 192$. Despite the significant increase in the normalization of the boost with the jointly dependent marks, the impact of correlated tail events will lead to a boost function with much longer tails, compared with the independent case.

Table 4.6 presents the estimated parameter bias for three different simulation studies of 1,000 replicates of sample size $n = 1,000$. Each simulation study considers a bivariate Hawkes process models with linear boost functions, with marks $X_i \in \mathbb{R}^2$ distributed $X_i \sim \text{GPD}(\zeta = 0.49, \delta = 1.00)$. The marks are jointly coupled and modelled via a *Gaussian copula* with $\rho_s = 0.50$. It should be noted that whilst we consider a *Gaussian copula* for this study, similar results are observed for both the *Gumbel copula* and *Clayton copula* models. The first study labelled *Theo.* in Table 4.6 uses theoretical moments to normalize the boost function in (4.36). The second study labelled *UB.* in Table 4.6 imposes an upper bound on the shape parameters of 0.4999. The third study labelled *Emp.* in Table 4.6 uses empirical moments to calculate the normalization of the boost function in (4.36) and will be described in more detail below.

Figure 4.21 presents the parameter estimates for the study that uses theoretical moments to normalize the boost in the evaluation of a Hawkes process, with linearly boosted bivariate generalized Pareto distributed marks, $\text{GPD}(\zeta = 0.49, \delta = 1.00)$, and with joint dependence modelled by a *Gaussian copula*, where $\rho_s = 0.50$. Figure 4.21 and Table 4.6 show a large bias for all parameter estimates. Given the true value of $\zeta = 0.49$, and taking into account simulation variation, the estimated shape parameter will often be greater than the required value for theoretical moments to exist. In addition, the optimization procedure will search over a range of shape parameters that will breach this requirement, resulting in unreliable estimates. The boost function parameter estimates are often reaching the upper bound of the optimization, which is set to 100 in this study.

Table 4.6: Parameter estimate bias (%) and variability ratio for three bivariate Hawkes process models, with linear boost functions, marks $X_i \in \mathbb{R}^2$ distributed $X_i \sim \text{GPD}(\zeta = 0.49, \delta = 1.00)$. The marks are jointly coupled and modelled via a *Gaussian copula* with $\rho_s = 0.50$. The three different models are, theoretical moments (*Theo.*), theoretical moments with an upper bound (*UB.*) of 0.4999 for the optimization procedure and empirical moments (*Emp.*). The simulations each have 1,000 replicates, with a sample size of $n = 1,000$ each.

Estimated parameter bias (%)									
Boost	imm	branch	decay	parameter 1		parameter 2		boost 1	boost 2
				mark 1	mark 2	mark 1	mark 2		
Theo.	2534.42%	-46.40%	104.72%	-4.10%	-4.11%	1.49%	1.50%	1033.92%	979.54%
UB.	0.20%	-47.97%	2.54%	-4.82%	-4.66%	1.54%	1.45%	1059.43%	864.97%
Emp.	0.51%	1.68%	5.74%	-0.79%	-0.25%	0.52%	0.23%	-5.13%	-6.90%

Variability ratio									
Boost	imm	branch	decay	parameter 1		parameter 2		boost 1	boost 2
				mark 1	mark 2	mark 1	mark 2		
Theo.	801.34	0.23	31.69	0.07	0.07	0.05	0.05	40.43	39.32
UB.	0.04	0.22	0.22	0.06	0.06	0.05	0.05	41.54	36.63
Emp.	0.06	0.18	0.11	0.09	0.09	0.06	0.05	0.46	0.55

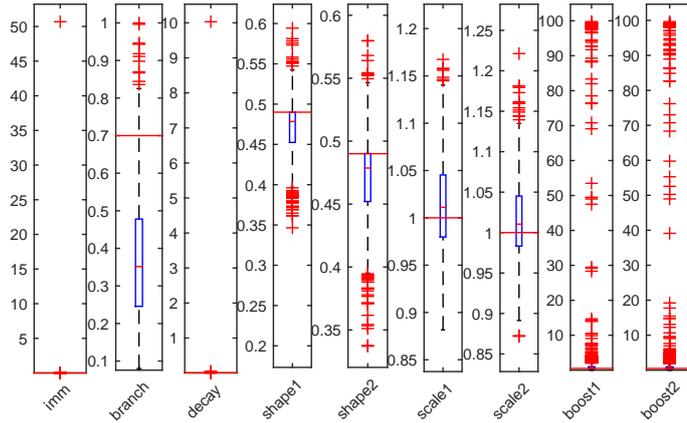


Figure 4.21: *Theoretical moments*. Boxplot of parameter estimates for a Hawkes process, with linearly boosted, bivariate GPD marks $\text{GPD}(\zeta = 0.49, \delta = 1.00)$, with joint dependence modelled by a *Gaussian copula*, where $\rho_s = 0.50$. The sample size is $n = 1,000$, for 1,000 replicates. The moments for the boost normalization are estimated theoretically. The solid red line represents the true value of each parameter.

A slightly better fit is achieved by imposing an upper bound on the shape parameters of 0.4999 in the optimization to ensure that second theoretical moments exist (Figure 4.22). However, there are still many cases where the boost parameters reach the upper bound of the optimization, resulting in greatly inflated boost parameters and a significant downward bias in the branching coefficient of -47.97% (Table 4.6).

Finally we consider a pragmatic approach of replacing theoretical moments with empirical moments in the calculation of the normalization of the boost function. Figure 4.23 demonstrates that the parameter estimates match closely to the true values by using the *approximate likelihood method by decoupling the marks parameters*. We observe one outlier for the second boost function parameter estimate out of 1,000 replicates, and a slight upward bias for the decay parameter of 5.74% (Table 4.6). Despite this, using empirical

moments in the event that the generalized Pareto distribution shape parameter is close to, or exceeding the theoretical second moments existing, appears to provide a robust alternative for estimation.

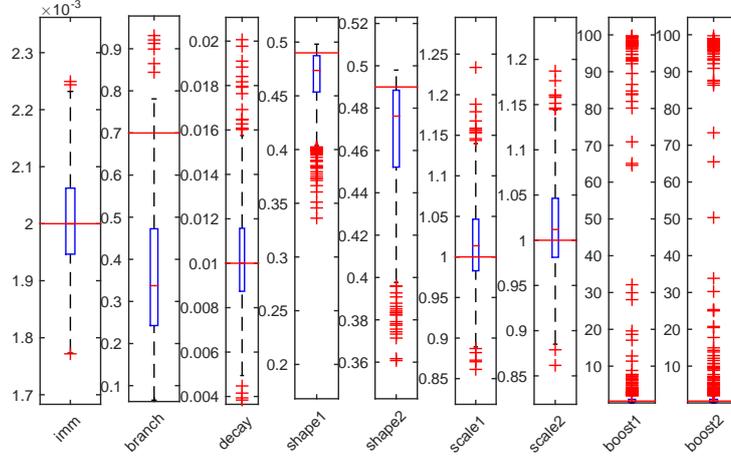


Figure 4.22: *Imposed upper bound.* Boxplot of parameter estimates for a Hawkes process, with linearly boosted, bivariate GPD marks $\text{GPD}(\zeta = 0.40, \delta = 1.00)$, with joint dependence modelled by a *Gaussian copula*, where $\rho_s = 0.50$. The sample size is $n = 1,000$, for 1,000 replicates. The moments for the boost normalization are estimated theoretically. The optimization has an upper bound of 0.4999 for the shape parameters. The solid red line represents the true value of each parameter.

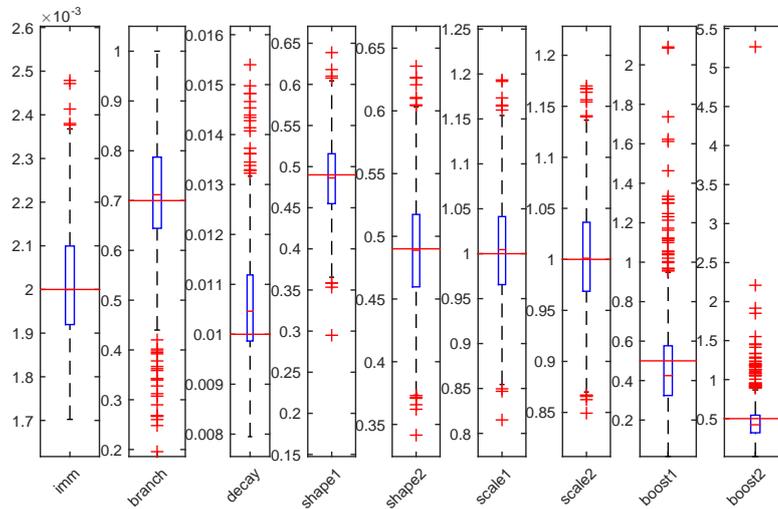


Figure 4.23: *Empirical moments.* Boxplot of parameter estimates for a Hawkes process with linearly boosted, bivariate GPD marks $\text{GPD}(\zeta = 0.40, \delta = 1.00)$, with joint dependence modelled by a *Gaussian copula*, where $\rho_s = 0.50$. The sample size is $n = 1,000$, for 1,000 replicates. The moments for the boost normalization are estimated empirically. The solid red line represents the true value of each parameter.

Joint dependence modelled via a Gumbel and Clayton copula

We complete the study of modelling a Hawkes process with jointly dependent marks, by presenting the simulation results for a bivariate Hawkes process, with a linear boost

function, marks $X_i \in \mathbb{R}^2$ with marginal distributions $X_i \sim \text{GPD}(\zeta = 0.05, \delta = 1.00)$ and jointly coupled by a copula model. The three copula models we consider are, Gaussian, Gumbel and Clayton models, where $\rho_s = 0.50$. As presented in Algorithm 8, the method for adjusting the normalization of the boost function requires the linear correlation of the marks. In the case of the Gaussian copula model, we have a closed form solution. For the Gumbel and Clayton copula model, this is not the case. We utilize a Monte-Carlo method to establish the long run linear correlation for the normalization, with the method outlined in Algorithm 8.

Table 4.7 presents the estimated parameter bias for the three models we consider. The models with a Gaussian copula model (Figure 4.24) and a Clayton copula model (Figure 4.25) present similar low bias in the estimated intensity and marks parameters. The Gumbel copula model captures dependence in the tails and we are modelling the marginal distribution of the marks with a heavy tailed model. As we noted previously, and can see in Table 4.7, the variations of the shape parameter estimates are much higher than the other estimated parameters, approximately 0.57 across all three models. This increase in variability of the ‘tail index’ may be influencing the ability of the Gumbel copula to accurately capture the dependence features in the tails. This is likely to be the cause of the observed small increase in the bias of the immigration intensity (3.19%) and higher bias in the boost functions parameter estimates, 16.85% and 11.12%.

Table 4.7: *Theoretical moments*. Parameter estimate bias (%) and variability ratio for three models, each being a bivariate Hawkes process model, with a linear boost function, with marks $X_i \in \mathbb{R}^2$ and marginal distributions $X_i \sim \text{GPD}(\zeta = 0.05, \delta = 1.00)$. The marks for each three models are jointly coupled via a Gaussian, Gumbel and Clayton model, respectively, where $\rho_s = 0.50$. The simulations each have 1,000 replicates, with a sample size of $n = 1,000$ each.

Estimated parameter bias (%)									
Copula	imm	branch	decay	parameter 1		parameter 2		boost 1	boost 2
				mark 1	mark 2	mark 1	mark 2		
Gaussian	2.21%	-0.22%	0.56%	-3.68%	-2.61%	0.23%	0.21%	12.54%	5.99%
Gumbel	3.19%	-0.26%	0.99%	-4.39%	-3.87%	0.24%	0.27%	16.82%	11.12%
Clayton	2.80%	-0.34%	1.42%	-6.89%	-4.72%	0.20%	0.31%	10.49%	5.65%
Variability ratio									
Copula	imm	branch	decay	parameter 1		parameter 2		boost 1	boost 2
				mark 1	mark 2	mark 1	mark 2		
Gaussian	0.0844	0.0505	0.0819	0.6224	0.6042	0.0438	0.0427	0.5582	0.4584
Gumbel	0.1067	0.0564	0.0874	0.5593	0.5581	0.0425	0.0416	0.6526	0.6229
Clayton	0.1011	0.0566	0.0851	0.6132	0.6804	0.0421	0.0444	0.4959	0.4649

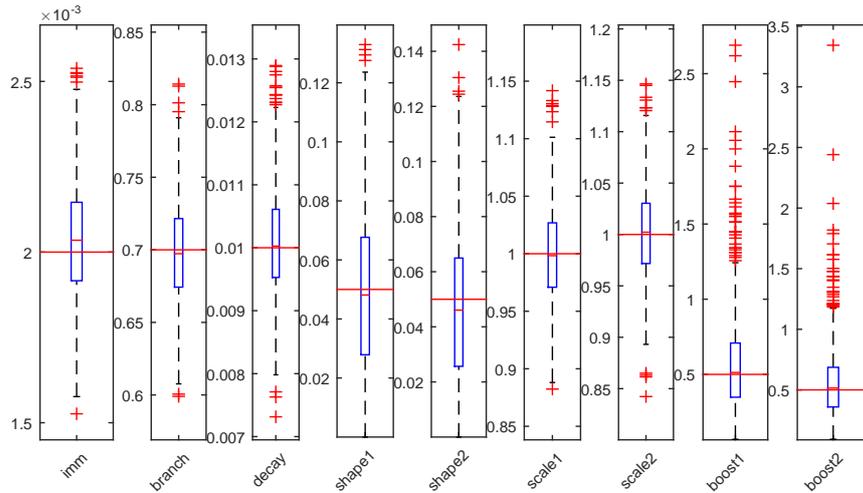


Figure 4.24: Boxplot of parameter estimates for a Hawkes process, with linearly boosted, bivariate GPD marks $\text{GPD}(\zeta = 0.05, \delta = 1.00)$, with joint dependence modelled by a *Gumbel copula*, where $\rho_s = 0.50$. The sample size is $n = 1,000$, for 1,000 replicates. The solid red line represents the true value of each parameter.

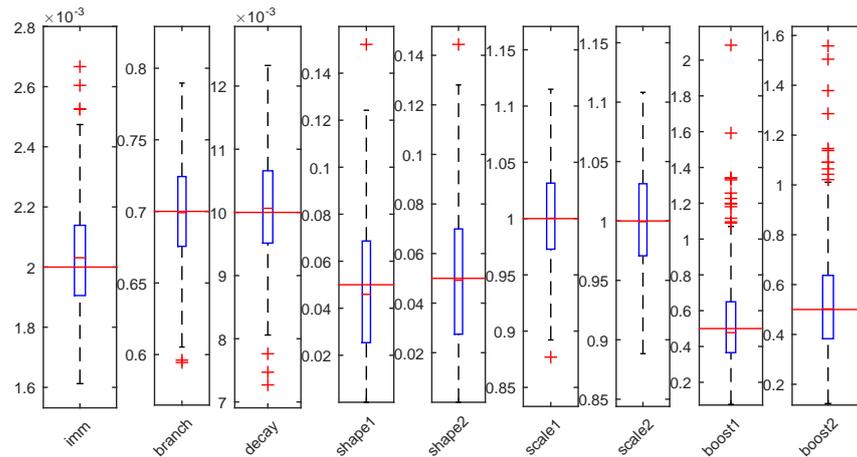


Figure 4.25: Boxplot of parameter estimates for a Hawkes process, with linearly boosted, bivariate GPD marks $\text{GPD}(\zeta = 0.05, \delta = 1.00)$, with joint dependence modelled by a *Clayton copula*, where $\rho_s = 0.50$. The sample size is $n = 1,000$, for 1,000 replicates. The solid red line represents the true value of each parameter.

In light of the increase in bias observed for the Hawkes process with marks that have joint dependence, and the success of reducing the bias in Section 4.5.5, by using the approximate likelihood method by decoupling marks parameters, we now present the bias and variation ratio for the three models using empirical rather than theoretical moments. Comparing the results in Table 4.7 with Table 4.8 below, we can see there is a significant reduction in the bias for the boost parameters and the immigration intensity parameters by a factor of approximately 0.7 across the three models. These findings suggest that in the case of marks with joint dependence, replacing empirical moments by theoretical moments, and using the approximate likelihood method, provides more robust estimates for the Hawkes process.

Table 4.8: *Empirical moments*. Parameter estimate bias (%) and variability ratio for three models, each being a bivariate Hawkes process model, with a linear boost function, with marks $X_i \in \mathbb{R}^2$ and marginal distributions $X_i \sim \text{GPD}(\zeta = 0.05, \delta = 1.00)$. The marks for each three models are jointly coupled via a Gaussian, Gumbel and Clayton model, respectively, where $\rho_s = 0.50$. The simulations each have 1,000 replicates with a sample size of $n = 1,000$ each.

Estimated parameter bias (%)									
Copula	imm	branch	decay	parameter 1		parameter 2		boost 1	boost 2
				mark 1	mark 2	mark 1	mark 2		
Gaussian	1.62%	-0.18%	1.50%	-3.01%	-6.32%	0.14%	0.34%	5.82%	7.79%
Gumbel	2.04%	-0.32%	1.08%	-4.71%	-4.22%	0.34%	0.31%	11.60%	9.86%
Clayton	1.42%	-0.31%	0.86%	-1.42%	-1.02%	-0.18%	0.00%	6.05%	7.39%
Variability ratio									
Copula	imm	branch	decay	parameter 1		parameter 2		boost 1	boost 2
				mark 1	mark 2	mark 1	mark 2		
Gaussian	0.0844	0.0505	0.0819	0.6224	0.6042	0.0438	0.0427	0.5582	0.4584
Gumbel	0.0846	0.0520	0.0867	0.5475	0.5516	0.0431	0.0432	0.5829	0.5380
Clayton	0.0881	0.0509	0.0840	0.6273	0.6269	0.0428	0.0423	0.4479	0.4715

4.6 Conclusion

Within this chapter, we have presented some key definitions and algorithms required for modelling, simulating and estimating the Hawkes process with multivariate marks. We presented procedures for both the estimation and simulation in the case of jointly dependent marks, proposing an adjustment to the boost function normalization and methods of estimating the copula adjustment, when a closed form solution is not possible. Recommendations were made regarding computational time, edge effects, initialization and burn-in times for the numerical algorithms.

Five simulation studies were presented to assess the fitting of the Hawkes process with increasingly complex models. Within this section we demonstrated the flexibility of the Hawkes process across a range of parametric distributions for marks, different functional forms for the boost function and extensions to higher dimensional marks and copula methods for jointly dependent marks.

The specification of the parameters in simulation, match closely to what one would expect when modelling real LOB data. These simulation studies highlighted various challenges when estimating the parameters of the via log-likelihood for a Hawkes process with multivariate marks, which to our knowledge have not been addressed in detail in current literature. These include:

- The estimated Hessian that the optimization procedure returns is inaccurate, leading to large and unreliable standard errors. A bootstrapping method for estimating standard errors is proposed, however this is computational expensive, due to the simulation and fitting of the Hawkes process many times.
- The shape parameter for the generalized Pareto distribution presents higher variability of the parameter estimates, compared with the other parameters within the Hawkes process. This is a well known challenge for the generalized Pareto distribution and a proposed two step estimation procedure produces reliable shape parameter

estimates when fitting the marginal generalized Pareto distribution to data. However, for this hybrid method, the properties would need to be studied in context of the log-likelihood method for the Hawkes process.

- The Hawkes process with coupled marks can be modelled via copula models, such as a Gaussian, Clayton or Gumbel copula model. For the case of marginally generalized Pareto distributed marks, this creates significant bias in the parameter estimates when the true value of the shape parameter is close to second moments not existing. This was addressed by replacing theoretically determined moments for normalizing the boost, by empirical estimates. However this approximation to the likelihood, which decouples the marks density parameters components of the likelihood from the Hawkes process component of the likelihood, requires further study, and which will be addressed in Chapter 7.

Sensitivity tests that varied the optimization initialization, showed that the parameters are well estimated with increased randomization of the initialization. However, given the complexity of the Hawkes process, which is further exacerbated by the introduction of marks and joint dependence, **the joint likelihood function may not be globally convex. Recall from Section 4.5.1 the impact of initialization for ultimate convergence of the optimization method used was assessed by using initialization, which had error of up to 50% from the true value, with negligible impact on the variation in final converged parameter estimates. Despite this insensitivity to the initialization of the optimization, we recommend varying initializations beyond what was implemented in this study, as alternative data sets may introduce further complexities and an introduced error of 50% may still be considered close to the true values. In addition, application of various optimization routines should be considered for implementation in practice, including use of regularization using a penalty function.**

An increase in the sample size of event times increased the reliability of the parameter estimates, as measured by the percentage bias of the parameter estimates and the variability ratio of the true values, for all parametric marks distributions considered within the Hawkes process. A sample size of $n > 600$ produced parameter estimates with a low bias.

The calculation of standard errors, by taking the inverse of the Hessian matrix for the evaluation of p-values, proved unreliable for even the most simplistic Hawkes process considered. The Hessian matrix had exceedingly large condition numbers, leading to numerically unstable results. Bootstrapping was proposed as an alternative method for estimating the standard errors. Despite the computational inefficiency of bootstrapping, the method produced reliable standard errors.

For a marked Hawkes process with different marginal distribution, exponential, generalized Pareto and Poisson, and linear boost functions, provided excellent parameter estimations for univariate, bivariate and higher dimensional marks. For marks of two or higher dimensions, combining the boost functions multiplicatively provided parameter estimates with a low bias and variability ratio, however, additive forms of combination presented a high bias for all parameter estimates.

The introduction of joint dependence between the marks presented additional challenges in the case of heavy tailed marks distribution. Second moments are required in the adjustment to the normalization of the boost function when extensions to the Hawkes process include joint dependence, modelled via copula methods. When the shape parameter approached second moments not existing, the parameter estimates were poor, even with a constraint placed on the optimization procedure. Replacing theoretical moments with empirical moments solved this issue.

For the copula models Gumbel and Clayton, a closed form solution for the theoretical moments does not exist. A proposed Monte-Carlo method for estimating the long run moments did not degrade the performance of the log-likelihood estimation of the parameters. Further to this, an increase in bias of the boost parameter estimates (up to 16%), for all Hawkes process with marks coupled via copula models, Gaussian, Gumbel and Clayton, was reduced to $\approx 5\%$, by using empirical moments, which was consistent with the bias observed in the study of Hawkes process with independent marks. It is recommended that if the marks are correlated, one should use empirical moments of the sample being modelled, rather than the theoretical moments to ensure a lower bias in the boost parameter estimates. This approximation to the likelihood, which decouples the marks density parameters components of the likelihood from the Hawkes process component of the likelihood, will be studied in more detail in the chapters that follow.

Chapter 5

Modelling the mark random vector

Chapter 3 explained how the Physical LOB was represented with an Observed LOB and the limitations this placed on the use of multivariate Hawkes processes for modelling the Observed LOB. Recall that, because of the one millisecond time granularity of observations, the Observed LOB has simultaneous events. The event process was defined on this time granularity, with an event e being any order type, market order (MO), limit order (LO) or cancellation (C) ($e \in \{MO, LO, C\}$) on any level of the LOB. To reduce the effect of the loss of information, due to the necessary amalgamation, information about the component events is captured using marks. Examples of such information that can be included as marks are: LOB depth volume; volume of events, such as MO, LO, C; counts of events; mid-price; spread; traded price, volatility measures; and various LOB imbalance measures.

The model that we consider appropriate for modelling the LOB intensity is the univariate Hawkes process, as defined in Section 4.1. On face value, one might consider using a multivariate Hawkes process, whereby components represent the levels or sides or event types of the LOB. However, the Hawkes process is a simple point process and this means that no two events can occur at the same event time t_i . As discussed in detail in Chapter 3, the Reported LOB contains data with a minimum event time granularity of one millisecond. If levels were treated as components, then events across components may occur at the same millisecond time-stamp in the Reported LOB, which breaches the requirement of the time of occurrence of the i th event of the process being, $\{t_i : t_i < t_{i+1}\} \in \mathbb{R}$.

One way of dealing with the limitations of the Reported LOB datasets when modelling the data via a Hawkes process, whilst maximizing the retention of information about the data, is to aggregate the data according to Algorithm 2 and 3. Within this process of aggregation, we collect useful attributes of the data sets that may help inform the future intensity of the process. In a general sense, this univariate Hawkes process with vector-valued marks can represent a collapsed version of the multivariate Hawkes process. What is assumed in this setting is the same decay function and immigration intensity, irrespective of the levels of the LOB considered. The marks can be modelled under different statistical assumptions, for instance, in the simplest settings with independence between marginals and no serial correlation, or with enhanced statistical features that account for different assumptions about dependence in the mark random vector sequences.

The marks can be endogenous or exogenous in nature. The marks that are derived from the Reported LOB, the same dataset that we derive the event process, are referred to as endogenous marks, whereas those that are derived from an external data source are exogenous marks. Exogenous marks are not to be confused with the immigration intensity η , that is often referred to as the exogenous intensity. Within this research we consider endogenous marks only, which are constructed from the available Reported LOB data sets. This is due to the readily available data and the vast amount of possible marks that can be constructed. However, for future research, exogenous marks should also be considered.

The simplest form of a marked Hawkes process is when the marks are marginally and jointly independent. However, as we will see in the studies that follow, rarely does this reflect the reality of the marks we consider. The choices that are made for the underlying model will depend not only on the assessment of the marginal distribution of the marks, but the extensions to the dependence structure of the marks, as well as the serial dependence between consecutive marks over time. This dependence structure could come in the form of serial dependence for individual marks, joint dependence between individual marks and as Daley and Vere-Jones (2007) note, in an extreme case, the marks could not only be dependent on the past evolution of the process, but be a direct function of it and therefore co-monotonic with the past intensity process.

In the spatial-temporal point processes literature, some formal tests for mark vector independence have been proposed by Guan (2006), Schlather et al. (2004) and Schoenberg (2004). This type of dependence has a very interpretable description in a geo-statistical setting. For example, Schlather et al. (2004) presents two situations where the assumption that the sampling points (locations) have been chosen independently of the spatial variables (mark vector) is violated. Prior scientific knowledge of a regionalized variable, and when there is dependence between the points and the variable is an intrinsic property of the data, i.e. positioning (location) of trees in a forest and growth rate (mark). These dependence studies address questions about a Hawkes process with a spatial dimension, and dependency between the mark vector and the point, whether it be spatial or temporal. Unlike these aforementioned studies, in this research we consider a temporal point process and the nature of the dependence within the mark vector.

5.1 A discussion of potential marks

In Section 1.6 we presented a review of the limited literature on the application of marked Hawkes process to financial data. Two approaches were proposed, the first used by Embrechts et al. (2011), Chavez-Demoulin and McGill (2012), Fauth and Tudor (2012) and Kirchner (2017b), include marks into the intensity process via a boost function, which is consistent with the method we propose in (4.14). The second approach considered by Rambaldi et al. (2017) use the marks to create a multivariate Hawkes process.

After reviewing the literature on the application of a marked Hawkes process to financial data in Section 1.6, the following summary and potential research extensions are noted:

- To date, the selection of marks in the literature is limited and not well motivated by the financial setting the marks are presented in. The limited number of marks considered include, the absolute value of excess returns, and LO and MO event volumes;
- Guidance of potential marks will therefore come from literature presented in Section 1.7 concerning, empirical features of the LOB, potential impacts to arrival rates of limit orders, or general LOB event and transactions, which we define as market orders (MO);
- Of the research that exists, there have been no formal studies on statistical model properties of the marks, rather they have just been assumed. We will progress this research by assessing the marginal distributional properties, and serial and joint dependence features of the mark vector;
- This work will underpin the research presented in Chapter 6, whereby we proposed a detection method for screening marks without fitting the full joint likelihood, but instead only requiring the fitting of the unmarked point process. This process has a significant practical advantage of this application domain, where the number of data points used in fitting is often very large and the number of potential marks one can consider is also high dimensional;
- The literature does not provide guidance on the appropriate boost functions, with only linear and power functions presented. The detailed study of the distributional and dependence properties will help guide the appropriate selection of boost functions for the inclusion of marks into the Hawkes process.

Given that the literature on marked Hawkes processes is insufficient to inform the selection of marks for a model of the LOB, we rely on the studies of the stylized features of the LOB. The discussion that follows, draws on the key points from Section 1.7, presenting some examples of potential marks that could be defined using the LOB datasets. The aim is to portray both the intuition from a market participants perspective, links to the literature on stylized features and the limitations in capturing the features of the LOB. A formal list of defined marks will be presented in Section 5.2, however we do provide example marks and links to the formal list throughout the discussion that follows. For a thorough summary of the key ideas that have emerged from studying and modelling the LOB, refer to Chakraborti et al. (2011), Gould et al. (2013) and Abergel et al. (2016).

The hypothesis proposed in numerous papers studying the volume relationship to price formation, is larger volumes, depicted as traders becoming more aggressive, result in an increase in MO submission. Research by, and not limited to Engle and Lunde (2003), Rinaldo (2004), Lo and Sapp (2010), Fauth and Tudor (2012), Rambaldi et al. (2017) provide support for this hypothesis. The findings by Fauth and Tudor (2012) and Rambaldi et al. (2017), via a genuine Hawkes point process with marked volumes is also consistent with this literature.

Numerous marks can be constructed from volumes in the LOB. If we first consider aggregate volume (depth) on the bid or ask (*Bid/Ask depth* in (5.1)), this mark aims to capture the total volume that is on the bid or ask side. If the total volume is increasing, this may lead to a boost in the clustering of events in the LOB. The limitation of this mark is that the total volume captured to represent the ‘depth’ is limited to the 5 to 10 levels that are reported in the Reported LOB, despite many more levels existing in the Physical LOB.

The volume of a MO on the bid and ask (*Bid/Ask vol MO* in (5.6)), is likely to send a stronger signal (more information) to the market than the volume of other event types such as a LO (*Bid/Ask vol LO* in (5.5)). If we consider a MO event, versus a LO submission within the depth of the LOB, and where both orders are equal size, the response of traders or trading algorithms is likely to be quite different, with a MO potentially triggering a greater clustering effect. Specification of the volume of the event is important, and capturing the additional detail of the event type is likely to improve the prediction of the future intensity of trade arrivals or limit order arrivals in the LOB.

Due to the requirement of unique event times in the Hawkes process discussed earlier, bid side and ask side of the LOB will be modelled as separate processes. Traders monitor the entire LOB when submitting MO, LO, C or amending orders, implying that the volume (depth) on the other side of the book (*Bid/Ask opp. side depth* (5.2)) may impact the arrival of new events in the LOB. If volumes are increasing on the bid and increasing on the ask simultaneously, this may impact the decision process of the market participant, compared with volumes increasing on the bid, but decreasing on the ask, which may be representative of momentum in the traded price. Two obvious limitations with this mark are: the representation of ‘depth’ via the maximum reported levels in the Reported LOB; and multiple changes on the ask side of the book between bid events that are not captured by the bid event process, but may hold some information about future events for the bid side.

Spread defined in (5.13), still plays a key role in the price discovery process, despite the significant increase in high frequency trading and the increase in liquidity of assets, thus the narrowing of spreads. For example, market-makers target an optimal operating spread to maximize profits and passive participants attempt to minimize price impact costs of which spread is a component. Numerous studies (Biais et al., 1995; Lo and Sapp, 2010; Toke and Pomponio, 2012; Engle and Russell, 1998; Rinaldo, 2004; Hall and Hautsch, 2006; Cao et al., 2008; Engle and Lunde, 2003; Ellul et al., 2003) have shown that wider spreads have been associated with an increase in the arrival rate of orders and an increase in limit orders arriving inside the spread. The Hawkes process captures the intensity of the order arrival, and incorporating spread as a potential mark will enable a deeper understanding of the linkage between spread and order flow.

The relative price is defined as the price away from the best bid (or ask) that a LO or C is submitted. Research by Gould et al. (2013), Biais et al. (1995), Bouchaud et al. (2002), Potters and Bouchaud (2003) and Zovko and Farmer (2002) demonstrates that the arrival rates of orders have been shown to be related to the relative price, rather than the actual

price and do not depend on features such as spread or mid-price.

Within the framework of a univariate Hawkes process, it is not possible to model the levels of the LOB as components. However, measures such as *Rel price LO* in (5.17) and *Rel price C* in (5.18) from the best bid (or ask) reflects LOB price level information in the Hawkes process. For example, a LO with a large relative price, residing deep in the LOB, is likely to have a smaller impact on the intensity function than a LO submitted at the best bid (or ask).

To our knowledge, studies on volume imbalance (Huang et al., 2014; Chordia et al., 2002; Ravi and Sha, 2014; Gould and Bonart, 2016) tend to focus on a relationship between imbalance and mid-price or traded price, with application to a specific trading strategy. Imbalance has not been studied directly with regards to LOB event arrivals. Gould and Bonart (2016) present queue imbalance in a LOB (*Imbalance* in (5.9)), as the normalized difference between the active buy and sell orders on the LOB. They use this queue imbalance to establish links between the states of the LOB and subsequent price changes within. The links were consistent with those in the research by Cao et al. (2009) who found that imbalances within the depth of the LOB were related to future short-term returns.

Extending on the generic imbalance measure, we consider the idea of relative depth profile presented by Gould et al. (2013). Relative depth profile can be described as the total size of all active orders with a relative price level. Utilizing the general ideas of relative prices and the relationship with arrival rates, we construct a novel relative imbalance measure (*Rel imbalance* in (5.10)), which is a price level weighted imbalance, with greater weight given to levels approaching the best bid or ask in the LOB.

From a trading perspective, imbalance in the LOB is often thought of as a measure of momentum, providing a signal for future price moves. For example, an increase in trading activity on the bid side in the form of LO submissions, will lead to increased volumes. If there is no or little change on the ask side, the combined effect will result in an ‘imbalance’ in the LOB. This will create pressure on the mid-price, causing it to potentially shift upwards. There are more buyers than sellers in the LOB, meaning the asset price is cheap. This causes upward pressure on the price until the asset becomes expensive and sellers enter the market to take advantage of the higher price.

The mid-price returns (*Mid-price returns* in (5.12)) reflect a shift up or down in the best bid and ask. Research by Ellul et al. (2003) found that on the NYSE, the rate of buy (or sell) LO arrivals was found to increase after periods of positive (negative) mid-price returns. However, Cao et al. (2008) found no evidence that mid-price returns had a significant impact on order arrivals on the ASX. Finally, empirical research by Rinaldo (2004), Ahn et al. (2001) and Engle and Russell (1998) suggests increased volatility being associated with higher periods of MO arrivals.

5.2 Defining the mark vector

The endogenous mark vector that we consider for this model is derived from the information that is present in the Reported LOB. In order to set notation, at each time t_i we obtain the following sequence of observations

$$(t_1, x_{1,1}, \dots, x_{1,d}), (t_2, x_{2,1}, \dots, x_{2,d}), \dots, (t_n, x_{n,1}, \dots, x_{n,d}),$$

where $x_i = (x_{i,1}, \dots, x_{i,d})^T$ is a random vector of d marks $\mathbf{X}_i \in \mathbb{X} \subset \mathbb{R}^d$, which we assume is always fully observable at each t_i . In the simplest form, the marks are independent of each other and the history of the process and have density $\mathbf{X} \sim f(\mathbf{x}; \phi)$.

Key to the formation of the random mark vector, is the underlying event process under consideration, refer to Section 3.2. In the context of this research, we can define the event process based on 3 key factors price level $l \in \{1, \dots, 10\}$, side $s \in \{B, A\}$ and event type $e \in \{LO, MO, C\}$.

We now proceed by detailing the full catalogued of candidate endogenous marks that we capture from the LOB data. This is not an exhaustive list, but it provides a detailed starting point, which is motivated by the literature on the stylized features of the LOB. The descriptions below provide a mathematical definition of the marks, which provides a clear description from which to develop algorithms for constructing the mark vector.

A summary of the shortened naming convention adopted for the marks for the remainder of this thesis is presented in Table 5.1, with reference to the formulation of the marks below.

Table 5.1: Shortened naming convention, with equation reference for the endogenous marks constructed from matched LOB data.

Depth based:	Centred depth/price based:	Price based:
Bid/Ask depth (5.1)	Imbalance (5.9)	Rel price LO (5.17)
Bid/Ask opp. side depth (5.2)	Rel imbalance (5.10)	Rel price C (5.18)
Volume based:	Mid-price:	Count based:
Bid/Ask vol MOLOC (5.3)	Mid-price (5.11)	Bid/Ask count MOLOC (5.19)
Bid/Ask vol MOLO (5.4)	Mid-price returns (5.12)	Bid/Ask count LO (5.20)
Bid/Ask vol MO (5.6)	Spread (5.13)	Bid/Ask count C (5.21)
Bid/Ask vol LO (5.5)	Traded MO price (5.14)	
Bid/Ask vol C (5.7)	Volatility mid-price (5.15)	
Inside vol MO (5.8)	Volatility mid-price ret. (5.16)	

5.2.1 Volume based marks

Bid/Ask depth. The volume depth of the LOB mark is the sum of the volume at t_i , given by

$$V_{t_i}^{(s,l)} = \sum_{k=1}^l \left(V_{t_i}^{(s,k)} \right), \quad (5.1)$$

where $V_{t_i}^{(s,k)}$ is the LOB volume in (3.2).

Bid/Ask opp. side depth. The volume depth of the opposite side of the book is the sum of the volume on the opposing side to what is being modelled at t_i . If we consider a bid side process, this mark reflects the depth volume on the ask side at t_i , and conversely for

the ask side process. This is denoted by

$$VO_{t_i}^{(s,l)} = \begin{cases} \sum_{k=1}^l \left(V_{t_i}^{(A,k)} \right), & \text{if } s = B; \\ \sum_{k=1}^l \left(V_{t_i}^{(B,k)} \right), & \text{if } s = A. \end{cases} \quad (5.2)$$

A number of marks are constructed from the volume associated with event types $e \in \{MO, LO, C\}$ and combinations thereof, across specified levels $l \in \{1, \dots, 10\}$, for each side $s \in \{B, A\}$. Before we define the volume of each event type, recall that we are not calculating the sum of volumes on the actual LOB, rather the volume associated with an event type. For this, we require the LOB price $P_{t_i}^{(s,l)}$ in (3.3), the trade direction D_{t_i} in (3.6), the MO (trade volume) $V_{t_i}^*$ in (3.4), and the MO price $P_{t_i}^*$ in (3.5), the LO volume $V_{t_i}^{LO,s,l}$ in (3.7) and the cancellation volume $V_{t_i}^{C,s,l}$ in (3.8). There are many ways one could define marks of event types. For this research, the volumes for each event type are aggregated across the price levels, which are specified when defining the event process. The specific event type combinations we consider are listed below.

1. *Bid/Ask vol MOLOC*. The volume of all events $e \in \{MO, LO, C\}$ at t_i , is defined as

$$V_{t_i}^{s,e} = \begin{cases} V_{t_i}^* \mathbb{I}[D_{t_i} = -1] + \sum_{k=1}^l \left(V_{t_i}^{(LO,s,k)} + V_{t_i}^{(C,s,k)} \right), & \text{if } s = B; \\ V_{t_i}^* \mathbb{I}[D_{t_i} = 1] + \sum_{k=1}^l \left(V_{t_i}^{(LO,s,k)} + V_{t_i}^{(C,s,k)} \right), & \text{if } s = A. \end{cases} \quad (5.3)$$

2. *Bid/Ask vol MOLO*. The volume of events $e \in \{MO, LO\}$ at t_i , is defined as

$$V_{t_i}^{s,e} = \begin{cases} V_{t_i}^* \mathbb{I}[D_{t_i} = -1] + \sum_{k=1}^l \left(V_{t_i}^{(LO,s,k)} \right), & \text{if } s = B; \\ V_{t_i}^* \mathbb{I}[D_{t_i} = 1] + \sum_{k=1}^l \left(V_{t_i}^{(LO,s,k)} \right), & \text{if } s = A. \end{cases} \quad (5.4)$$

3. *Bid/Ask vol LO*. The volume of event $e \in \{LO\}$ at t_i , is defined as

$$V_{t_i}^{s,e} = \sum_{k=1}^l \left(V_{t_i}^{(LO,s,k)} \right). \quad (5.5)$$

4. *Bid/Ask vol MO*. The volume of event $e \in \{MO\}$ at t_i , originating on the bid and ask, is defined as

$$V_{t_i}^{s,e} = \begin{cases} V_{t_i}^* \mathbb{I}[D_{t_i} = -1], & \text{if } s = B; \\ V_{t_i}^* \mathbb{I}[D_{t_i} = 1], & \text{if } s = A. \end{cases} \quad (5.6)$$

5. *Bid/Ask vol C*. The volume of event $e \in \{C\}$ at t_i , is defined as

$$V_{t_i}^{s,e} = \sum_{k=1}^l \left(V_{t_i}^{(C,s,k)} \right). \quad (5.7)$$

Inside vol MO. Volume that has transacted within the spread, not originating on either the bid or ask is defined as

$$VIS_{t_i}^* = V_{t_i}^* \left(\mathbb{I}[P_{t_i}^* < P_{t_i}^{(A,1)}] \cap \mathbb{I}[P_{t_i}^* > P_{t_i}^{(B,1)}] \right). \quad (5.8)$$

5.2.2 Volume imbalance marks

1. *Imbalance.* The first volume imbalance we consider is defined in Gould and Bonart (2016) and described as queue imbalance in a LOB, given by

$$I_{t_i} = \frac{\sum_{k=1}^l \left(V_{t_i}^{(B,k)} \right) - \sum_{k=1}^l \left(V_{t_i}^{(A,k)} \right)}{\sum_{k=1}^l \left(V_{t_i}^{(B,k)} \right) + \sum_{k=1}^l \left(V_{t_i}^{(A,k)} \right)}, \quad (5.9)$$

where I_{t_i} is the normalized difference between the active buy orders at t_i and the active sell orders at t_i . Due to the linear rescaling $I_{t_i} = [-1, 1]$.

2. *Relative imbalance.* The relative depth profile is defined as the set of ordered pairs $(l, V_{t_i}^{(s,l)})$ at t_i (Gould et al., 2013). We use the level l to determine the normalized weight in

$$w_k = \frac{(l - k)/l}{\sum_{k=1}^l (l - k)/l}, \text{ where } k \in \{1, \dots, l\}.$$

The relative imbalance is

$$RI_{t_i} = \frac{\sum_{k=1}^l \left(w_k V_{t_i}^{(B,k)} \right) - \sum_{k=1}^l \left(w_k V_{t_i}^{(A,k)} \right)}{\sum_{k=1}^l \left(w_k V_{t_i}^{(B,k)} \right) + \sum_{k=1}^l \left(w_k V_{t_i}^{(A,k)} \right)}. \quad (5.10)$$

5.2.3 Price based marks

1. *Mid-price.*

$$MP(t_i) = \frac{P_{t_i}^{(A,1)} - P_{t_i}^{(B,1)}}{2}. \quad (5.11)$$

2. *Mid-price returns.*

$$MPR(t_i) = \frac{MP(t_i) - MP(t_{i-1})}{MP(t_{i-1})}. \quad (5.12)$$

3. *Spread.*

$$S(t_i) = P_{t_i}^{(A,1)} - P_{t_i}^{(B,1)}. \quad (5.13)$$

4. *Traded MO price.*

$$P_{t_i}^*. \quad (5.14)$$

5. *Volatility mid-price.* The volatility of the mid-price across the previous 100 events is defined as

$$\sigma_{MP(t_i)} = \sqrt{\frac{1}{100} \sum_{j=i-100}^i \left(MP(t_j) - \frac{1}{100} \sum_{j=i-100}^i MP(t_j) \right)^2}. \quad (5.15)$$

6. *Volatility mid-price ret.* The volatility of the mid-price returns across the previous

100 events is defined as

$$\sigma_{MPR(t_i)} = \sqrt{\frac{1}{100} \sum_{j=i-100}^i \left(MPR(t_j) - \frac{1}{100} \sum_{j=i-100}^i MPR(t_j) \right)^2}. \quad (5.16)$$

To calculate the relative price of a LO and a C below, we use LOB data prior to aggregation (Section 3.1.3). We re-introduce the subscript j and let $t_{i,j} \in \mathbb{R}$ be a unique time-stamp of events $i \in \{1, \dots, n\}$, and within each unique time stamp there may be multiple events $j \in \{1, \dots, m_i\}$. The relative price of an event, refers to the price of the event on the bid or ask side relative to the best bid or ask, respectively. If there are say two LO events at level 3 and level 2 at event time $t_{i,j}$, where $j \in \{1, 2\}$, and for the purposes of constructing a mark, we need to summarize this information into a single relative price. To address this, we have taken the mean across the levels when more than one event occurs at an event time. We require the definition of net volume defined at all events, $V_{t_{i,j}}^{net,s,l}$ in (3.1), and the LOB price $P_{t_{i,j}}^{(s,l)}$ in Definition 5.

7. *Rel price LO*. The relative price of event $e \in \{LO\}$ is defined as

$$RP(t_i)^{s,e} = \frac{\sum_{k=1}^l \sum_{j=1}^{m_i} \left(P_{t_{i,j}}^{(s,l=1)} - P_{t_{i,j}}^{(s,l)} \right) \mathbb{I}[V_{t_{i,j}}^{net,s,k} > 0]}{\sum_{k=1}^l \sum_{j=1}^{m_i} \mathbb{I}[V_{t_{i,j}}^{net,s,k} > 0]}. \quad (5.17)$$

8. *Rel price C*. The relative price of event $e \in \{C\}$ is defined as

$$RP(t_i)^{s,e} = \frac{\sum_{k=1}^l \sum_{j=1}^{m_i} \left(P_{t_{i,j}}^{(s,l=1)} - P_{t_{i,j}}^{(s,l)} \right) \mathbb{I}[V_{t_{i,j}}^{net,s,k} < 0]}{\sum_{k=1}^l \sum_{j=1}^{m_i} \mathbb{I}[V_{t_{i,j}}^{net,s,k} < 0]}. \quad (5.18)$$

5.2.4 Count based marks

Similarly to the relative price calculations, to define the count based marks that follow, we use LOB data prior to aggregation. The formulation below, counts the number of events j that contribute to the unique time-stamped event at $t_{i,j}$. We require the definition of net volume $V_{t_{i,j}}^{net,s,l}$ in (3.1). Recall, $t_{i,j} \in \mathbb{R}$ is a unique time-stamp of events $i \in \{1, \dots, n\}$ and within each unique time stamp there may be multiple events $j \in \{1, \dots, m_i\}$.

- *Bid/Ask count MOLOC*. The count of all events $e \in \{MO, LO, C\}$ is the sum of events that make up a single aggregated unique event time across specified levels $l \in \{1, \dots, 10\}$, and is defined as

$$C_{t_i}^{(s,e)} = \begin{cases} \sum_{k=1}^l \sum_{j=1}^{m_i} \left(\mathbb{I}[V_{t_{i,j}}^{net,s,k} \neq 0] \right) + \sum_{j=1}^{m_i} \mathbb{I}[D_{t_{i,j}} = -1], & \text{if } s = B; \\ \sum_{k=1}^l \sum_{j=1}^{m_i} \left(\mathbb{I}[V_{t_{i,j}}^{net,s,k} \neq 0] \right) + \sum_{j=1}^{m_i} \mathbb{I}[D_{t_{i,j}} = 1], & \text{if } s = A. \end{cases} \quad (5.19)$$

- *Bid/Ask count LO*. The count of events $e \in \{LO\}$ is defined as

$$C_{t_i}^{(s,e)} = \sum_{k=1}^l \sum_{j=1}^{m_i} \left(\mathbb{I}[V_{t_i,j}^{net,s,k} > 0] \right). \quad (5.20)$$

- *Bid/Ask count C*. The count of events $e \in \{C\}$ is defined as

$$C_{t_i}^{(s,e)} = \sum_{k=1}^l \sum_{j=1}^{m_i} \left(\mathbb{I}[V_{t_i,j}^{net,s,k} < 0] \right). \quad (5.21)$$

5.3 Properties of the mark vector

In the simplest setting, we assume the marks are i.i.d and independent of the previous realized marks and event times. This section makes an assessment of the key assumptions related to the marks within a Hawkes point process framework. We first consider the empirical features of the marks. The methods of analysis used throughout this chapter for assessing the marks properties are described in advance. We then present three case studies that demonstrate the analysis in detail and highlight the extensive challenges of modelling marks. Within these case studies we test the i.i.d. assumption by assessing the marks for serial dependence. We apply transformations to the marks to ensure stationarity, before considering the distributional properties of the marks. The methods presented in the case studies are extended across time, so that we can assess the robustness of our findings. We complete this study by assessing the joint dependence of the mark vector. The findings from this section will have major implications on the appropriateness of each of the marks considered. In addition, extensions required for a Hawkes process to accommodate a relaxation of previously assumed strong statistical assumptions, which may not appropriate to the present application.

For the studies that follow, we will consider the futures asset SILVER for levels 1:5, across 10 trading days, 20-July-2015 to 31-July-2015. Within each day, we take time segments that represent 10,000 events per time segment across the day. This constitutes a total of 37 different time segments. Chapter 3 provides a detailed description of the LOB data that underpins the marks. In addition, Table 3.4 provides an asset description and the market hours, referred to as the *liquid market hours*, in local trading time of the exchange.

5.3.1 Statistical properties

For the definition of the marks, the underlying event process determines the side and levels that are incorporated in the construction of the marks. For the studies below, ‘bid’ marks utilizes event times related to bid events and likewise, ‘ask’ marks utilizes event times related to ask events. For those without a label of ‘bid’ or ‘ask, for example, imbalance and mid-price, the underlying event process is the bid side, with the construction of the mark utilizing the ask side information at the bid side event time.

As shown in Table 5.2, a number of marks have a high frequency of zeros. Marks that are associated with a specific event type $e \in \{LO, MO, C\}$ will exhibit a zero if, for example, an event occurred due to a LO only, then a mark that represents the volume of event types MO and C, will be zero at this event time. Whereas the mark, mid-price returns, has naturally occurring zeros that should not be removed. It is appropriate to remove the zeros from the marks in the analysis of the statistical properties when they are not naturally occurring. Section 5.4 outlines some suggested boost functions that are appropriate for incorporating these particular marks into the Hawkes process.

Table 5.2 demonstrates mean summary statistics across 10 trading days. For volume and count based marks, we can see evidence of heavy tails, with high kurtosis and right skewness. For the count based marks, the range between minimum and maximum possible values are indicative of a discrete distribution, whereas the volume based marks are suggestive of continuous distributions. A variety of distributions will need to be considered for the marks.

Table 5.2: Mean values of the summary statistics of the mark vector, related to both bid and ask side events, for levels 1:5, across a 10 trading days (July 2015), for SILVER.

Mark	Zeros	Mean	Med	Std	Skew	Kurt	Min	Max	Count
B depth (5.1)	0.00%	229.26	228.60	39.11	0.07	4.08	85.10	373.30	42846
A depth (5.1)	0.00%	223.66	226.40	37.09	-0.22	3.61	80.30	347.70	41754
B opp. depth (5.2)	0.00%	224.31	227.10	37.21	-0.19	3.62	80.90	347.20	42846
A opp. depth (5.2)	0.00%	229.71	229.10	39.16	0.07	4.10	86.00	373.00	41754
B vol MOLOC (5.3)	0.00%	5.19	2.00	8.48	5.05	56.78	1.00	181.80	42846
A vol MOLOC (5.3)	0.00%	4.94	2.00	7.91	4.66	52.27	1.00	169.30	41754
B vol MOLO (5.4)	25.66%	4.45	2.00	7.54	5.74	74.16	1.00	167.00	31963
A vol MOLO (5.4)	25.18%	4.22	2.00	7.00	5.46	79.87	1.00	162.90	31288
B vol MO (5.6)	91.74%	2.60	1.10	3.51	7.02	99.18	1.00	70.40	3551
A vol MO (5.6)	92.84%	2.69	1.20	4.69	12.59	341.18	1.00	137.60	3000
B vol LO (5.5)	28.36%	4.32	2.00	7.37	5.83	77.38	1.00	165.60	30813
A vol LO (5.5)	27.73%	4.10	2.00	6.78	4.78	42.74	1.00	116.30	30226
B vol C (5.7)	35.09%	2.90	2.00	3.14	3.57	32.63	1.00	65.10	27902
A vol C (5.7)	35.43%	2.75	1.90	2.91	2.99	17.84	1.00	45.70	27016
Inside vol MO (5.8)	95.68%	2.81	1.90	3.47	5.40	55.67	1.00	52.70	1848
Imbalance (5.9)	0.93%	0.01	0.01	0.10	0.20	2.87	-0.31	0.36	42459
Rel imbalance (5.10)	0.07%	0.01	0.01	0.10	0.10	3.07	-0.38	0.42	42817
Midprice (5.11)	0.00%	14.70	14.69	0.05	-0.15	2.62	14.58	14.82	42846
B rel price LO (5.17)	54.40%	0.01	0.01	0.01	1.27	8.05	0.00	0.12	19695
A rel price LO (5.17)	54.53%	0.01	0.01	0.01	1.03	4.65	0.00	0.08	19070
B rel price C (5.18)	59.35%	0.01	0.01	0.01	1.08	4.03	0.00	0.06	17565
A rel price C (5.18)	59.31%	0.01	0.01	0.01	1.06	3.81	0.00	0.05	17079
Mid-price ret (5.12)	87.40%	0.00	0.00	0.00	0.24	9.47	0.00	0.00	5418
Spread (5.13)	0.00%	0.01	0.01	0.00	1.89	6.60	0.00	0.03	42846
Trade MO P (5.14)	86.90%	14.70	14.70	0.05	-0.15	2.63	14.58	14.82	5631
Volat MP (5.15)	0.18%	0.00	0.00	0.00	2.73	30.88	0.00	0.01	42772
Volat MP ret (5.16)	0.96%	0.00	0.00	0.00	2.12	28.30	0.00	0.00	42441
B cnt MOLOC (5.19)	0.00%	2.68	2.00	2.53	2.89	16.75	1.00	40.30	42846
A cnt MOLOC (5.19)	0.00%	2.64	2.00	2.47	2.82	15.32	1.00	35.50	41754
B cnt LO (5.20)	28.36%	1.94	1.00	1.58	2.78	14.93	1.00	22.90	30814
A cnt LO (5.20)	27.73%	1.91	1.00	1.53	2.79	14.97	1.00	22.40	30227
B cnt C (5.21)	35.09%	1.72	1.00	1.27	2.89	16.47	1.00	18.60	27902
A cnt C (5.21)	35.43%	1.70	1.00	1.23	2.79	15.00	1.00	17.30	27016

5.3.2 Methods of analysis

To study the properties of the mark vector in more detail, we employ various statistical methods to assess whether there is an appropriate parametric or discrete distribution that can be used to model the marks. In addition, the nature of serial dependence in the marks and cross correlation between the marks. Much of this research will draw on statistical methods from Chapter 2, which laid the foundation for the volume based marks. Whilst the methods of analysis in Chapter 2 were used for investigating the properties of volume profiles, we will see that the features discovered are consistent across many of the marks we have defined.

Marginal distribution assessment

To assess whether the marks we consider can be sufficiently well modelled by a continuous distribution, we consider the marks on the bid and ask side for 10 trading days, segmented into intra-day time intervals containing 10,000 events each. We determine the range of the unique values of each mark for each segment.

The case studies in Section 5.3.4 present the time series plots, histograms and autocorrelation functions. The CDF $F_X(x)$ is evaluated using the estimated parameters and quantiles from the associated mark data. Then the random variable Y is defined as $Y_i = F_X(x_i)$, which has a uniform distribution. We present four probability integral transform (Angus, 1994) distributions as histograms to assess the ability of modelling the marks with a specified distribution.

The probability integral transformations form the basis for the Komogorov-Smirnov (KS) test. The statistic is computed by $\sup |F_n(x) - F_X(x)|$, where F_n for n i.i.d. observations x_i is the empirical distribution function

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}[x_i \leq x]. \quad (5.22)$$

It is well known that the KS-test is used for testing continuous distributions only (Massey, 1951) and it is more influence by deviations in the median than in the tails of the distribution. This is likely to be problematic given most of the marks present heavy tailed features. These issues will underpin the visual representation of the probability integral transform histograms. We provide the plots as another visual tool for attempting to better understand the statistical properties of the marks.

To further determine the appropriate distribution for the data, we also fit a parametric probability distribution to the data. We then assess these various fitted distributions using the Negative of the log-likelihood, BIC, AIC and AICc (AIC with correction for finite sample sizes). An ‘identity map’ is created, where the criteria are sorted and the most appropriate marginal distributional fit is selected. The list of continuous distributions considered are: beta; Birnbaum-Saunders; exponential; extreme value; gamma; generalized extreme value; generalized Pareto; inverse Gaussian; logistic; log-logistic; lognormal; Nakagami; normal; Rayleigh; Rician; Student-t location-scale; and Weibull. For the dis-

crete distributions we consider: binomial; negative binomial; and Poisson. For the discrete distribution assessment, it is worth noting that most marks are of integer form and where possible, for those that are not we transform the mark data to integer (Table 5.3) before assessing the appropriate discrete distribution.

The assessment described above to determine the appropriate parametric probability distribution, considers continuous and discrete distributions separately. We employ an additional assessment of the deviation of the empirical CDF compared with a selection of candidate distributions, generalized Pareto distribution, normal, negative binomial and Poisson, giving further guidance as to whether a discrete or continuous approximation is more suitable for each mark considered. We calculate the sum squared differences (SSD) of the empirical CDF F_n in (5.22), with the CDF $F_X(x)$ evaluated using the parameters estimated from the associated mark data and defined

$$SSD = \sum_{i=1}^n (F_n(x_i) - F_X(x_i))^2. \quad (5.23)$$

This test is very close in construction to the well known Cramer-von Mises criterion (Anderson, 1962), and it serves as a comparative estimate only of the deviation of the candidate CDF with the empirical CDF, rather than calculating a specific test statistic to assess the null hypothesis that a sample is drawn from a candidate distribution.

As we will demonstrate in the case studies in Section 5.3.4, the marks are not easily modelled by any single parametric distribution. The reason is that they contain statistical attributes which include heavy tailedness and serial dependence, and it is not clear in many cases whether a discrete or continuous distribution is most appropriate. The complexity of these data sets makes it difficult to use any one statistical method to determine the appropriate parametric distribution, necessitating the use of various methods described in this section.

Serial dependence

In the case studies in Section 5.3.4 we study the marginal serial dependence by first presenting ACF charts, which demonstrate serial dependence in all marks. For a selection of marks, differencing to the required order is performed to remove temporal trends. We investigate this more formally by assessing the serial dependence across the 10 trading days considered. The assessment is made on the transformed data, for example, the depth based mark vector will be differenced and then we assess the differenced mark in blocks of 10,000 events across a range of lags using the Ljung-Box test (Ljung and Box, 1978) of serial dependence. The test statistic is defined as

$$Q = n(n+2) \sum_{k=1}^h \frac{\hat{\rho}_k^2}{n-k},$$

where n is the sample size, $\hat{\rho}_k$ is the sample autocorrelation at lag k , and h is the number of lags being tested, where $h \in \{3, 5, 20, 50\}$. Under the null hypothesis, if there is no serial dependence the test statistic follows a $\chi_{(h)}^2$.

Long memory of the mark random variables

In Section 2.2.2 we introduced the Hurst exponent to test for long memory in the volume of the LOB, by implementing a de-trended fluctuation analysis (Kantelhardt et al., 2001; Hu et al., 2001) to estimate the Hurst exponent. This provides an *index of long-range dependence*, giving a quantitative measure of the relative tendency of a time series to regress strongly to the mean, or to cluster.

To extend the tests of serial dependence in the marks, and the potential for presence of long memory, we calculate the Hurst exponent for each time segment and day considered in this analysis for the asset SILVER. As a guide, values for the index in the range $0.5 < H < 1$ indicate a time series with long-term positive autocorrelation (Simonsen and Hansen, 1998). This indicates momentum in the mark, whereby a high value in the series is likely to be succeeded by another high value period. Values of the Hurst exponent between $0 < H < 0.5$ indicates a time series with long-term switching between high and low values.

Cross correlation

As we have presented in Section 4.1.1, it is possible to accommodate dependence between marks by using a copula to couple the marginals into a joint distribution. The techniques described above motivate the choice of marginal distributions. There are many measures of dependence available to assess the dependence between the marks. In this research we consider only pairwise measures of dependence to adequately capture the relationship between marks. This provides a natural starting point for the measurement of the dependency of the mark random vector.

Two such measures of pairwise dependence used in this research are the Pearson's linear correlation and the non-parametric measure of dependence, Spearman's rank correlation. The two classes of copula model we consider in this work, Gaussian and Archimedean (Appendix A.1), can use these measures to evaluate the model copula parameter, however the rank correlations such as Spearman's rank correlation does not depend on the marginal distribution, unlike the Pearson's correlation.

5.3.3 Simulating a Hawkes process with serial dependences

We propose a method which is an extension to the simulation algorithm for a Hawkes process with multivariate marks in Algorithm 8, to enable the creation of a simple time series model, which captures the key observed features of the marks including serial dependence in the marginal distribution. The aim of simulating the data with serial dependence is to produce quantile-quantile plots to visually assess the appropriateness of the chosen marginal distributions once serial dependence is accounted for. This work also provides the foundation for power studies in Chapter 6 and model simulations in Chapter 7.

To simulate a Hawkes process with serially dependent marks and non-Gaussian marginal distributions, we specify an unobserved (latent) parameter evaluation of an appropriate parametric distribution. To generate the time series of marks, we let δ_i be a

latent stationary time series. Conditional on δ_i , the marks are independent with a specified parametric density $f(x_i|\theta_i)$, where for a selected component of θ_i , we use δ_i and for the other components of θ_i , we use fixed values.

For the single parameter distributions that we consider, they simply become $\text{Exp}(\lambda_i)$ and $\text{Pois}(\mu_i)$, where $\lambda_i = \delta_i$ and $\mu_i = \delta_i$, respectively. In the case of two parameter distributions, and for example in the case of heavy tailed replicates, the conditional generalized Pareto distribution has a shape parameter ζ and a scale parameter δ_i . For the negative binomial distribution, serial dependence is introduced into the process via success rate $r_i = \delta_i$.

Throughout this thesis we use a latent process of the form

$$\delta_i = a_0 + \sum_{k=1}^p a_k \delta_{i-k} + \epsilon_i, \quad (5.24)$$

where ϵ_i is Gaussian white noise.

To date, methods for fitting models with the observed characteristics present in the marks are not established and are beyond the scope of this thesis. We take a pragmatic approach of trial-and-error, choosing parameters that best emulate the features of the data, with the view to simulate the methodology and test how it performs under more complex assumptions.

For a conditional generalized Pareto distribution, the scale parameter becomes $\delta_i = a_0 + \sum_{k=1}^p a_k \delta_{i-k} + \epsilon_i$, and we set the lag $p = 1$, $a_0 = 0$, and the coefficients term a_1 will vary.

The marks enter the Hawkes process via a normalized boost function. For i.i.d. marks, this boost function is normalized via the estimated theoretical moments of the marginal distribution. However, in the case of serially dependent marks, it is not obvious how to proceed with specifying the normalization for a conditional generalized Pareto distribution, for example. No theory has been developed with respect to the normalization of a boost function in the presence of marks with conditional distributions. The normalization for every time point, for example, an estimate of δ_i from an observed mark, is a non-trivial filtering problem.

There are many ways we could define a data generating mechanism (DGM). Two such ways are:

1. Normalize the boost function locally using δ_i within the theoretical moment expression $\mathbb{E}[X] = \frac{\delta_i}{1-\zeta}$;
2. Generate a very long series of serially dependent marks $X \sim GPD(\zeta, \delta_i)$. Calculate the empirical mean of this series and use the global estimate to normalize the boost function.

We will proceed with the second DGM proposed. The algorithm below outlines the steps that need to be replaced in Algorithm 8 to simulate a series of random events according to a Hawkes process. The steps presented, go beyond what is required in

this chapter, however all enhancements will be utilized in the chapters that follow. For completeness, we present the enhancements in their entirety here.

Algorithm 11 (Enhancements to Algorithm 8). *Replacement of the ‘Monte-Carlo method’ in Algorithm 8.*

1. *Specify the autoregressive structure, which will be used to invoke serial dependence. In this setting, a reasonable choice is*

$$\delta_i = 0.9\delta_{i-1} + \epsilon_i, \quad (5.25)$$

where ϵ_i is Gaussian white noise. The support of the parameter needs to be considered, for example, the generalized Pareto distribution requires $\delta > 0$, therefore we take the absolute value of the autoregressive structure.

2. *Specify the marginal distributions $F_{i,j}$ of the mark and where the appropriate, the parameter of the distribution is specified by the autoregressive structure defined above. Some sensible choices for δ_i in (5.25), are*

$$Y_i \sim \text{Exp}(\lambda_i) \text{ where } \lambda_i = \delta_i;$$

$$Y_i \sim \text{Pois}(\mu_i), \text{ where } \mu_i = \delta_i;$$

$$Y_i \sim \text{GPD}(\delta_i, \zeta);$$

$$Y_i \sim \text{NB}(r_i, p), \text{ where } r_i = \delta_i.$$

3. *In the case that the simulated marks are independent, sample a very long time series (i.e. $n = 20,000$) of random variables from $Y_{i,j} \sim F_{i,j}$, where $j \in \{1, \dots, d\}$ and where the appropriate parameter of the distribution is specified by the autoregressive structure defined above.*
4. *In the case of $d \geq 2$ jointly dependent marks with independence copula W and marginal mark distributions $F_{i,1}, \dots, F_{i,d}$, let $U_{i,1}, \dots, U_{i,d}$ have joint distribution C . For a very long time series $i = 1 \dots n$:*
 - (a) *Simulate a random variable $u_{i,1}$ from Uniform(0, 1);*
 - (b) *Simulate a random variable $u_{i,2}$ from $C_{i,2}(\cdot|u_{i,1})$. Continue simulating a random variable $u_{i,d}$ from $C_{i,d}(\cdot|u_{i,d-1})$;*
 - (c) *Sample $(Y_{i,1}, \dots, Y_{i,d})$ from $F^{-1}(u_{i,1}), \dots, F^{-1}(u_{i,d})$.*
5. *Define $y_{i,j} := Y_{i,j}$ and calculate the long run empirical moments and correlation in the event of joint dependence which we will consider below, and as required for the specification of boost function, for example, $\mathbb{E}[Y_j]$, $\text{Var}[Y_j]$ and $\rho(Y_1, \dots, Y_d)$.*

Replacement of ‘step 4’ in Algorithm 8.

Using step 1 and step 2 above, with the specified marginal distribution $F_{i,j}$ and autoregressive structure for an appropriately chosen parameter of the marginal distribution, we proceed with the following.

- In the case of univariate marks or multi-dimensional marks that are independent, sample a random variable from $X_{i,j} \sim F_{i,j}$, where $j \in \{1, \dots, d\}$.
- In the case of $d \geq 2$ jointly dependent marks with independence copula W and marginal mark distributions $F_{i,1}, \dots, F_{i,d}$, let $U_{i,1}, \dots, U_{i,d}$ have joint distribution C .
 - Simulate a random variable $u_{i,1}$ from $Uniform(0, 1)$.
 - Simulate a random variable $u_{i,2}$ from $C_{i,2}(\cdot | u_{i,1})$. Continue simulating a random variable $u_{i,d}$ from $C_{i,d}(\cdot | u_{i,d-1})$.
 - Sample $(X_{i,1}, \dots, X_{i,d})$ from $F^{-1}(u_{i,1}), \dots, F^{-1}(u_{i,d})$.
- Define $x_{i,j} := X_{i,j}$

Replacement of ‘step 6’ in Algorithm 8.

Calculate the boost function $g(\mathbf{x}; \phi, \psi)$. However, for the normalization of the boost function we do not use the theoretical moments based on the marginal distribution, we instead make use of the long run empirical moments from step 5 above, $\mathbb{E}[Y_j]$, $Var[Y_j]$.

In the case of $d \geq 2$ jointly dependent marks, recall that we have to make an adjustment to the normalization of the boost function, as described in Section 4.2.1. For the evaluation of $\mathbb{E}[X_1 X_2] \neq 0$, we require the estimation of the linear correlation. Rather than calculating the linear correlation (explicitly or otherwise depending on the copula model specified), we utilize the long run empirical linear correlation from step 5 above, $\rho(Y_1, \dots, Y_d)$.

5.3.4 Case studies and categorization

This section aims to identify the non-trivial challenges when attempting to fit a parametric distribution to the marks, and to provide recommendations when attempting to fit a full joint likelihood function for the Hawkes process with multivariate marks. As we will present in Chapter 6, this won’t impact the assessment of the mark on the intensity function via a score test, as the score test utilizes empirical moments.

The analysis of the statistical properties of the marks has uncovered some common themes, which has enabled the broad categorization of the marks. Within each category, the marks exhibit common distributional features and challenges for model fitting. For brevity, within each category we present a single case study for a single mark, which represents the analysis undertaken across all marks. The categories and the single mark example within each are:

1. *Marks which are difficult to model.* Case study: Bid/Ask depth in (5.1);
2. *Marks best modelled by a continuous distribution.* Case study: Bid/Ask vol MOLOC in (5.3);
3. *Marks best modelled by a discrete distribution.* Case study: Bid/Ask vol C in (5.7).

For each mark we assess whether an appropriate transformation is required to obtain the time series x'_t , which is closer to stationarity and has a more easily modelled distribution. Table 5.3 presents the final choices of transformation across all marks, including

first order differencing, removal of zeros in the cases that a boost function can account for event types and the suggested discrete transformation.

Table 5.3: A summary of the transformations applied to the mark vector. This is related to both bid and ask side (where appropriate), for levels 1:5, across a 10 trading days (July 2015), for SILVER

Mark	$ x_t - x_{t-1} $	$x_t \setminus \{0\}$	Range of unique values	Discrete transformation
Bid/Ask depth (5.3)	✓		[49,125]	integer
Bid/Ask opp. side depth (5.2)	✓		[51,117]	integer
Bid/Ask vol MOLOC (5.3)			[58,99]	integer
Bid/Ask vol MOLO (5.4)		✓	[46,84]	integer
Bid/Ask vol MO (5.6)		✓	[16,29]	integer
Bid/Ask vol LO (5.5)		✓	[42,82]	integer
Bid/Ask vol C (5.7)		✓	[22,44]	integer
Inside vol MO (5.8)		✓	[13,24]	integer
Imbalance (5.9)	✓		(7960,8915)	N/A
Rel imbalance (5.10)	✓		(9360,9913)	N/A
Mid-price (5.11)	✓		[11,25]	$z_t = \lfloor x_t - x_{t-1} \times 1000 \rfloor / 25$
Rel price LO (5.17)		✓	(164,245)	N/A
Rel price C (5.18)		✓	(146,242)	N/A
Mid-price returns (5.12)	$ x_t $		(64,225)	N/A
Spread (5.13)			[4,12]	$z_t = \lfloor x_t \times 1000 \rfloor / 5$
Traded MO price (5.14)	✓	✓	[37,135]	$z_t = \lfloor x_t - x_{t-1} \times 1000 \rfloor$
Volatility mid-price (5.15)	✓		(6639,8467)	N/A
Volatility mid-price ret. (5.16)	✓		(1884,3418)	N/A
Bid/Ask count MOLOC (5.19)			[21,44]	integer
Bid/Ask count LO (5.20)		✓	[13,25]	integer
Bid/Ask count C (5.21)		✓	[10,23]	integer

Marks which are difficult to model

The features we present for the case study that follows are consistent across the selection of marks that are outlined below:

1. *Volume based:* Bid/Ask depth (5.1), Bid/Ask opp. side depth (5.2), Imbalance (5.9) and Rel imbalance (5.10);
2. *Price based:* Mid-price (5.11), Mid-price returns (5.12), Spread (5.13), Traded MO price (5.14), Volatility mid-price (5.15) and Volatility mid-price ret (5.16).

The marks in this category appear to exhibit long memory, however after first level differencing only limited serial dependence remains. It is worth noting that Mid-price returns are not differenced, but exhibit similar properties to the differenced marks.

Section 5.4 presents different boost function structures that may be appropriate for marks that can take a negative value. We assume a linear boost for the marks throughout this discussion, therefore we study the absolute value of the differenced mark. It should be noted that within this formulation, the magnitude of the boosts impact on the intensity function can vary according to the sign the mark takes prior to transformation, the impact is not required to be symmetric. However, a power law or exponential boost would also be appropriate. Investigation of the broad range of boost functions for the Hawkes process is beyond the scope of this thesis.

Case study 1: Bid depth

Figure 5.1 presents the time series plots, histograms and ACF charts for Bid depth. The histograms show that after the transformations are applied, we can see the mark appears to have heavy tailed features and we might consider a distribution such as a generalized Pareto distribution or a negative binomial distribution.

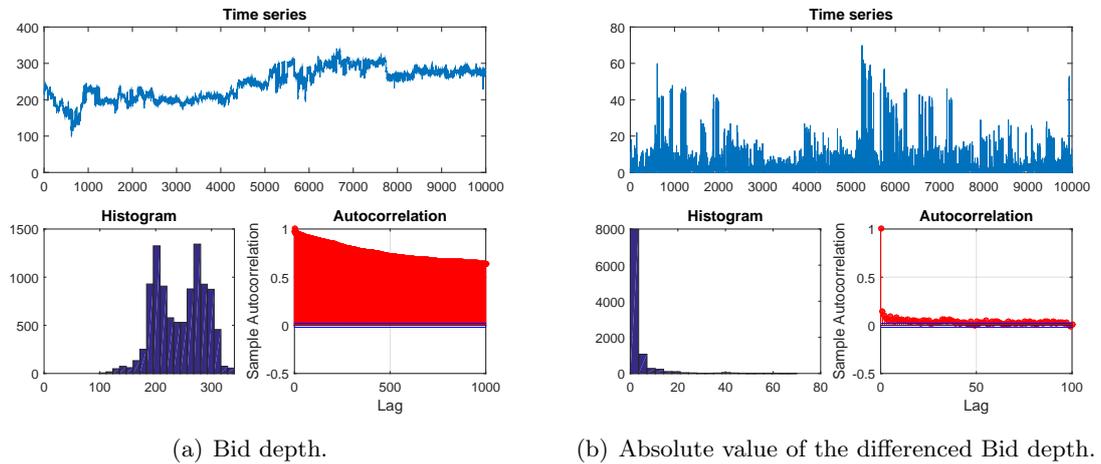


Figure 5.1: Time series plots, ACF and histograms of a sample of 10,000 events, for Bid depth, with levels 1:5, for SILVER, on trading date 31-July-2015.

Section 5.3.2 introduced the probability integral transform (PIT) histograms, which are used as a visual tool to explore the fit of various marginal distributions. The selection of marginal distributions used in the PIT histograms comes from an extensive assessment of various continuous and discrete distributions outlined in Section 5.3.2. For this particular case study, we found that the generalized Pareto distribution was the most appropriate continuous distribution for Bid depth and the negative binomial distribution was the most appropriate choice for a discrete distribution. In Figure 5.2, we've included the PIT plots for the normal and Poisson distributions for comparative purposes only.

The PIT histogram will be uniform if the distribution is appropriate for the data. From Figure 5.2, this is clearly not the case for all distributions considered. However, the heavy tailed distributions, generalized Pareto distribution and the negative binomial distribution, do a marginally better job than the normal distribution as one would expect based on the statistical properties presented in Table 5.2.

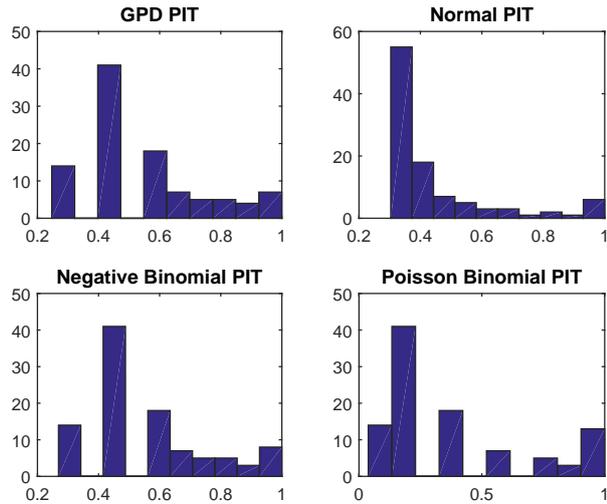


Figure 5.2: Probability integral transform (PIT) histogram, for a sample of 10,000 events, for Bid depth, with levels 1:5, for SILVER, on trading date 31-July-2015.

The assessment of the appropriate marginal distribution and the analysis of the PIT histogram presented in Figure 5.2, are both limited by the presence of serial dependence in the mark data. To address this, we begin by fitting the chosen marginal distribution to the mark. The parameter estimates are used as a guide for the simulation method outlined in Section 5.3.3. We simulate a mark with a conditional generalized Pareto distribution and conditional negative binomial distribution with serial dependence.

Table 5.4 presents the summary statistics for Bid depth and a replicate from both simulations. We see for the generalized Pareto distribution replicate, the kurtosis is more than twice that of the real data, whereas the negative binomial distribution fails to mimic sufficient kurtosis and skewness.

Table 5.4: Summary statistics for Bid depth, for levels 1:5, across a 10 trading days (July 2015), for SILVER and two simulates from a conditional GPD and a conditional negative binomial distribution, with serial dependence.

Mark	Mean	Med	Std	Skew	Kur	Min	Max
Bid depth	4.28	2	6.37	4.93	32.84	1	71
GPD replicate with serial dependence	2.19	2	2.26	5.99	74.64	1	49
NB replicate with serial dependence	6.39	4	6.63	1.96	8.47	1	65

Figures 5.3 and 5.4 top panels present the quantile-quantile plot for the generalized Pareto distribution and the conditional generalization Pareto distribution (serial dependence), and the negative binomial distribution and the conditional negative binomial distribution (serial dependence), respectively. The aim is to assess how closely the replicate matches the real data. The second panel in these charts presents the histograms for the real data and the replicate with serial dependence. Finally, the bottom panel shows the ACF plots for the real data and the replicate with serial dependence. What is apparent from these plots is that the replicate with a conditional generalized Pareto distribution (Figures 5.3) with serial dependence is providing the best fit, but it is still underestimating

the centre of the distribution and the tails. Translated Bid depth is difficult to model and replicate even with the introduction of heavy tailed models and serial dependence.

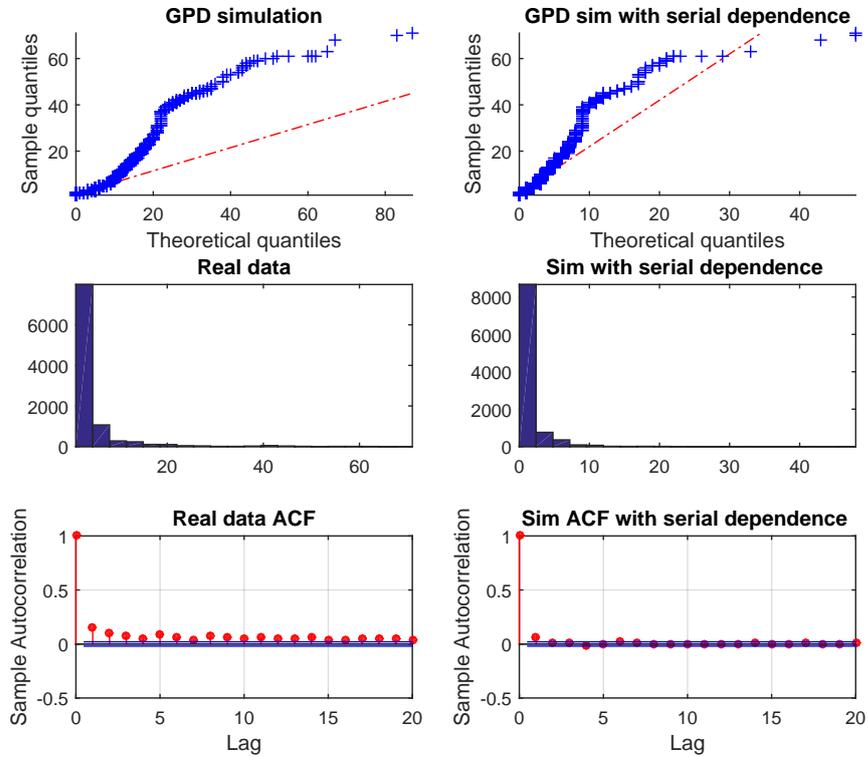


Figure 5.3: QQ-plots, histograms and ACF plots. *Top Left*: GPD simulated data without serial dependence; *Right Column*: simulated data with serial dependence; and *Mid and Bottom Left*: a sample of 10,000 events for Bid depth ('real data') associated with levels 1:5, for SILVER, on trading date 31-July-2015. Theoretical quantiles correspond to a simulated sample from the model.

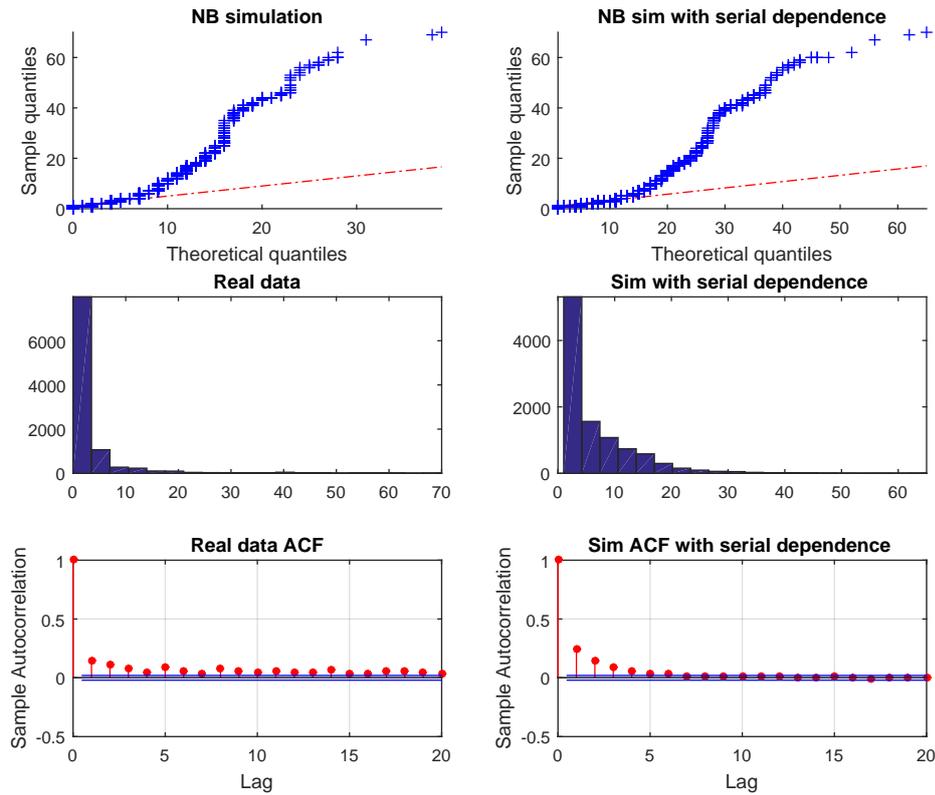


Figure 5.4: QQ-plots, histograms and ACF plots. *Top Left*: negative binomial distributed simulated data without serial dependence; *Right Column*: simulated data with serial dependence; and *Mid and Bottom Left*: a sample of 10,000 events for Bid depth ('real data') associated with levels 1:5, for SILVER, on trading date 31-July-2015. Theoretical quantiles correspond to a simulated sample from the model.

Marks best modelled by a continuous distribution

The marks considered in this section also exhibit serial dependence, however when these features are accounted for, promising results for a reasonable distributional fit to the data are observed. The marks that exhibit similar features are outlined below:

1. *Volume event based*: Bid/Ask vol MOLOC ((5.3)), Bid/Ask vol MOLO (5.4), Bid/Ask vol MO (5.6) and Bid/Ask vol LO (5.5);
2. *Price event based*: Rel price LO (5.17) and Rel price C (5.18);
3. *Count based*: Bid/Ask count MOLOC (5.19) and Bid/Ask count LO (5.20).

Case study 2: Bid vol MOLOC

Figure 5.5 shows the time series plots, histograms and ACF chart for Bid vol MOLOC. Consistent with our observations in the summary statistics Table 5.2, where we noted high skewness and kurtosis, heavy tailed features and serial dependence are observed.

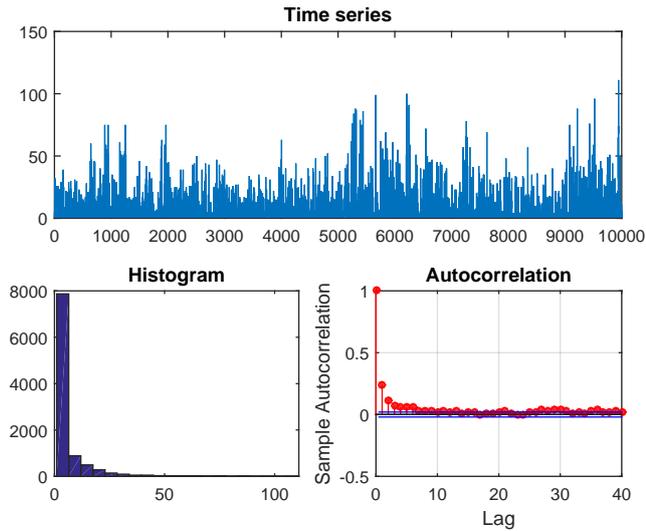


Figure 5.5: Time series plots, ACF and histograms of a sample of 10,000 events for Bid vol MOLOC, with levels 1:5, for SILVER, on trading date 31-July-2015.

For the assessment of the appropriate marginal distribution, we consider both continuous and discrete distributions, as it is not clear from the empirical features of the mark which is more appropriate. Upon initial assessment of various distributional fits to the data, the generalized Pareto distribution appears to be the most appropriate continuous marginal fit and the negative binomial distribution is the most appropriate discrete marginal fit. The PIT histograms again provide no guidance as to the most appropriate distribution from the reduced selection of distributions. It is worth noting that the volume data is integer and no transformations are required when modelling the mark with a discrete distribution.

In the previous cases studies, we utilized the marginal parameter estimates for the specification of the simulation parameters that are not specified as a deterministic function for incorporation of serial dependence. For the generalized Pareto distribution, the estimation of the marginal parameters assuming i.i.d., when in fact the mark has serial dependence, results in an over inflated shape parameter estimate, due to the correlated events in the tails of the distribution. To date, methods for fitting models with these observed characteristics present in the marks, have not been established, therefore we take a pragmatic approach of trial-and-error when choosing the parameters to best emulate the features of the mark. A reduction in the shape parameter to $\zeta = 0.3$, and specifying the scale parameter as δ_i in (5.24), with autoregressive coefficients, $a_0 = 0$ and $a_1 = 0.9$ with lag 1, best mimics this mark.

Table 5.5 presents the summary statistics for Bid vol MOLOC and a replicate from the simulation with serial dependence, with a conditional generalized Pareto distribution and a conditional negative binomial distribution. The skewness and kurtosis still appears too high, but the mean and maximum volume simulated is within close proximity of what is observed in the marks data.

Table 5.5: Summary statistics for Bid vol MOLOC, for levels 1:5, across a 10 trading days (July 2015), for SILVER, and two simulates from a conditional GPD and a conditional negative binomial distribution, respectively.

Mark	Mean	Med	Std	Skew	Kur	Min	Max
Bid vol MOLOC	5.67	2	9.25	4.04	25.22	1	111
GPD replicate with serial dependence	3.56	2	5.32	7.29	93.47	1	122
NB replicate with serial dependence	9.24	6	9.86	1.91	7.91	1	85

Figures 5.6 and 5.7 presents the quantile-quantile plots, histogram and ACF plots for the generalized Pareto distribution and the conditional generalization Pareto distribution (serial dependence), and the negative binomial distribution and the conditional negative binomial distribution (serial dependence), respectively. In Figure 5.7, the negative binomial distribution reflects the serial dependence well, but compared with the sample data, the theoretical quantiles are too low and the tails of the distribution are not long enough. From this assessment, we conclude that the generalized Pareto distribution with serial dependence (Figure 5.6) best reflects the features of the Bid vol MOLOC mark.

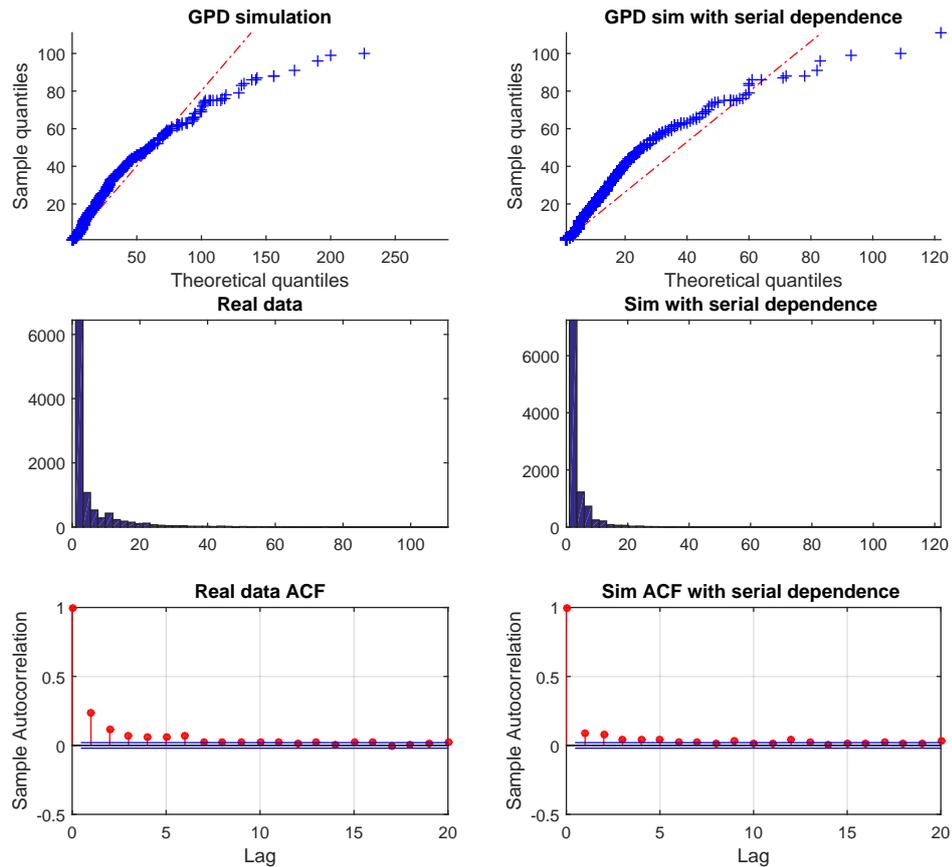


Figure 5.6: QQ-plots, histograms and ACF plots. *Top Left*: GPD simulated data without serial dependence; *Right Column*: simulated data with serial dependence; and *Mid and Bottom Left*: a sample of 10,000 events for Bid vol MOLOC (‘real data’) associated with levels 1:5, for SILVER, on trading date 31-July-2015. Theoretical quantiles correspond to a simulated sample from the model.

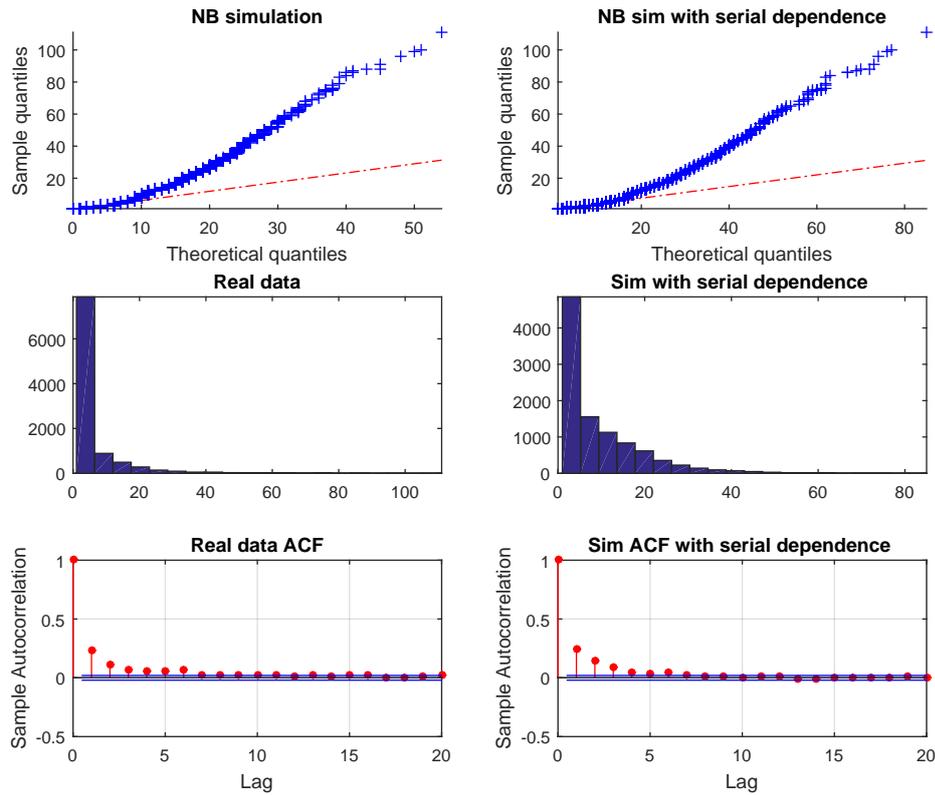


Figure 5.7: QQ-plots, histograms and ACF plots. *Top Left*: negative binomial distributed simulated data without serial dependence; *Right Column*: simulated data with serial dependence; and *Mid and Bottom Left*: a sample of 10,000 events for Bid vol MOLOC (‘real data’) associated with levels 1:5 for SILVER, on trading date 31-July-2015. Theoretical quantiles correspond to a simulated sample from the model.

Marks best modelled by a discrete distribution

For a small selection of marks that exhibit very low counts, the appropriate distributional form will be discrete. Assessment of the marginal distribution shows that a negative binomial distribution is most appropriate for Bid/Ask vol C and Inside vol MO, however a Poisson distribution is most appropriate for the Bid/Ask Count C, with the mean and variance being close to equality for the 10 trading days considered.

1. *Negative binomial distributed*: Bid/Ask vol C (5.7) and Inside volume MO (5.8).
2. *Poisson distributed*: Bid/Ask count C (5.21).

Case study 3: Bid Vol C

The final case study investigates the statistical features of the mark Bid vol C. Figure 5.8 shows serial dependence in the mark, even after the transformation that removes the zero event data.

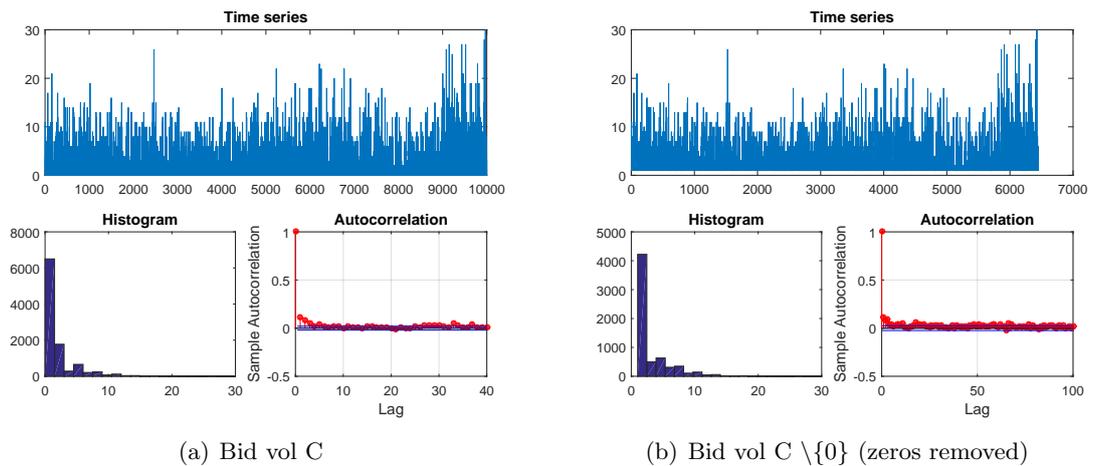


Figure 5.8: Time series plots, ACF and histograms of a sample of 10,000 events for Bid vol C, with levels 1:5, for SILVER, on trading date 31-July-2015.

Simulating a mark with a conditional negative binomial distribution with serial dependence, produces similar summary statistics to the mark.

Table 5.6: Summary statistics for Bid vol C, for levels 1:5, across a 10 trading days (July 2015), for SILVER, and a simulate from a conditional negative binomial distribution with serial dependence.

Mark	Mean	Med	Std	Skew	Kur	Min	Max
Bid vol C	3.02	2	3.17	2.51	11.77	1	30
NB replicate with serial dependence	2.91	2	2.57	1.97	8.12	1	22

From Figure 5.9, the quantile-quantile plots show a slightly improved replication of the real data with the introduction of serial dependence, however the autocorrelation is low, so the improvement of the fit is minimal. As we have seen in previous cases of count based marks, but to a lesser extent here, the negative binomial distribution slightly underestimates the tails of the distribution.

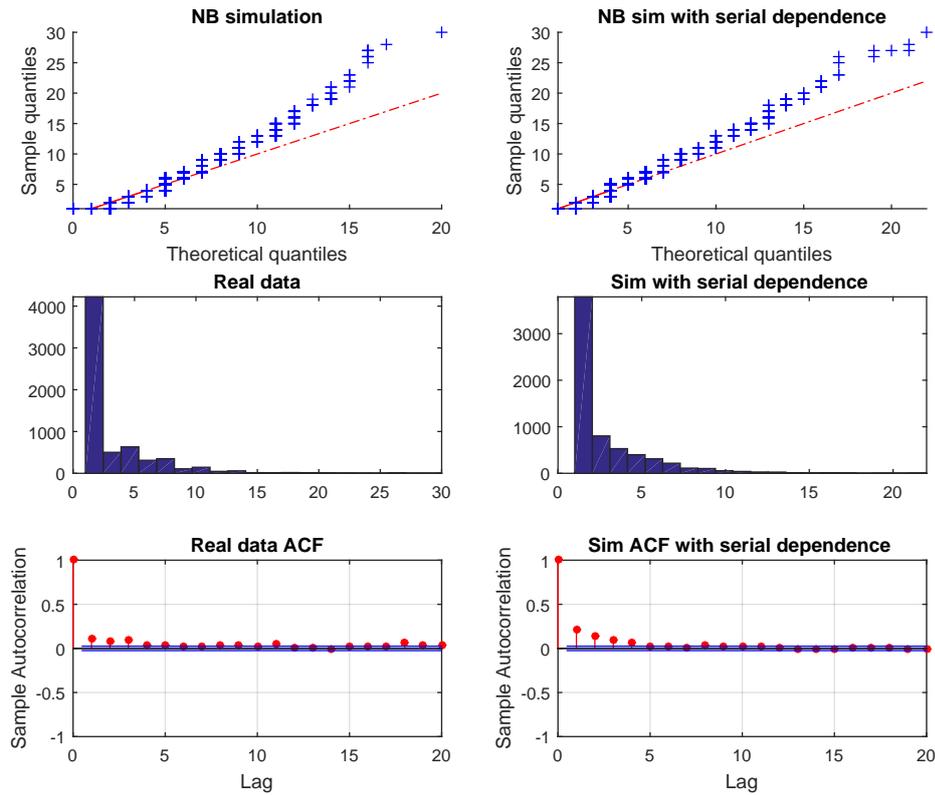


Figure 5.9: QQ-plots, histograms and ACF plots. *Top Left*: negative binomial distributed simulated data without serial dependence; *Right Column*: simulated data with serial dependence; and *Mid and Bottom Left*: a sample of 10,000 events for Bid vol C (‘real data’) associated with levels 1:5, for SILVER, on trading date 31-July-2015. Theoretical quantiles correspond to a simulated sample from the model.

Summary

As discussed in Section 5.3.2, the underlying LOB data that the marks are constructed from is discrete. However, in some cases the construction of the marks results in continuous data, for example imbalance measures, where a continuous distribution may provide a better approximation of the marginal distribution. For some marks that are low count, transformations to integer are necessary to make an assessment of the appropriate discrete marginal distribution.

Whilst there are many possible marks that can be derived from the LOB data, the case studies have shown that the marks pose significant challenges when trying to fit a marginal distribution to the data. Even after transformations (Table 5.3) and considerations of serial dependence, it is clear from each of the case studies that the complexity of the data sets does not allow for a simple application of standard methods to find the correct parametric distribution. Whilst the identification of some marks properties and distributional form has been made, there is substantial research still required to better understand the marks statistical properties and methods of modelling that allow a relaxation of assumptions of

independence. In support of our findings, we broadly categorize the marks in Table 5.7.

Table 5.7: Categorization of the mark vector

Difficult to model	Continuous distribution approx.	Discrete distribution
<i>Depth based:</i>	<i>Volume based:</i>	<i>Volume based:</i>
Bid/Ask depth (5.1)	Bid/Ask vol MOLOC (5.3)	Bid/Ask vol C (5.7)
Bid/Ask opp. side depth (5.2)	Bid/Ask vol MOLO (5.4)	Inside vol MO (5.8)
<i>Centred depth/price based:</i>	Bid/Ask vol MO (5.6)	<i>Count based:</i>
Imbalance (5.9)	Bid/Ask vol LO (5.5)	Bid/Ask count C (5.21)
Rel imbalance (5.10)	<i>Price based:</i>	
Mid-price (5.11)	Rel price LO (5.17)	
Mid-price returns (5.12)	Rel price C (5.18)	
Spread (5.13)	<i>Count based:</i>	
Traded MO price (5.14)	Bid/Ask count MOLOC (5.19)	
Volatility mid-price (5.15)	Bid/Ask count LO (5.20)	
Volatility mid-price ret. (5.16)		

5.3.5 Extended studies across time

The section that follows, extends upon the analysis discussed in the case studies in Section 5.3.4, by making assessment of the various features across time segments for 10 trading days. Whilst we have identified the complexity and challenges in identifying the distributional properties of the marks, this extended study attempts to provide recommendations required for modelling the marks in a Hawkes process, despite the acknowledged limitations.

Marginal distributions for the mark vector

Using the information garnered in Section 5.3.4, we now formally assess the appropriate marginal distributional fits for the marks, with the required transforms outlined in Table 5.3. It was observed in Table 5.2, the mean summary statistics across 10 trading days, that the marks have heavy tails and right skewness. This is further supported by the initial assessment of appropriate marginal distributions, the generalized Pareto distribution and the negative binomial distribution in the discrete case.

Table 5.8 presents the appropriate marginal distributions. For the discrete marks, the equality of the mean and variance is used to inform the suitability of a Poisson distribution. We estimate the shape parameter for the generalized Pareto distribution and note the range of the estimates. This will assist in informing the limitations of the appropriate boost function, where first moments are required ($\zeta < 1$) for a linear boost and higher moments are required for higher degree boost functions. For the majority of marks, first moments exist and in most cases, second moments exist.

Table 5.8: Proposed continuous and discrete marginal distribution for each mark, for levels 1:5, across 10 trading days (July 2015), for SILVER. Shape range refers to the GPD shape parameter estimates, across the 10 trading days.

Mark	Continuous	Discrete	Zeros	Mean	Var.	Shape range
Bid depth (5.1)	GPD	Neg Bin	0.0%	4.03	41.00	$\zeta \in [-0.07, 0.29]$
Ask depth (5.1)	GPD	Neg Bin	0.0%	2.85	29.02	$\zeta \in [-0.07, 0.32]$
Bid opp. side depth (5.2)	GPD	Neg Bin	0.0%	3.37	27.20	$\zeta \in [0.02, 0.34]$
Ask opp. side depth (5.2)	GPD	Neg Bin	0.0%	2.63	38.88	$\zeta \in [0.02, 0.36]$
Bid vol MOLOC (5.3)	GPD	Neg Bin	0.0%	5.41	80.32	$\zeta \in [0.31, 0.45]$
Ask vol MOLOC (5.3)	GPD	Neg Bin	0.0%	5.09	67.37	$\zeta \in [0.31, 0.50]$
Bid vol MOLO (5.4)	GPD	Neg Bin	25.7%	4.70	66.11	$\zeta \in [0.23, 0.44]$
Ask vol MOLO (5.4)	GPD	Neg Bin	25.2%	4.40	53.30	$\zeta \in [0.23, 0.51]$
Bid vol MO (5.6)	GPD	Neg Bin	91.7%	2.64	13.37	$\zeta \in [0.02, 0.19]$
Ask vol MO (5.6)	GPD	Neg Bin	92.8%	2.75	28.31	$\zeta \in [0.02, 0.30]$
Bid vol LO (5.5)	GPD	Neg Bin	28.4%	4.55	63.55	$\zeta \in [0.23, 0.45]$
Ask vol LO (5.5)	GPD	Neg Bin	27.7%	4.27	49.62	$\zeta \in [0.23, 0.51]$
Bid vol C (5.7)	N/A	Neg Bin	35.1%	2.95	10.24	
Ask vol C (5.7)	N/A	Neg Bin	35.4%	2.80	8.95	
Inside vol MO (5.8)	GPD	Neg Bin	95.7%	2.87	12.77	$\zeta \in [-0.14, 0.22]$
Imbalance (5.9)	GPD	N/A	0.0%	1.01	0.00	$\zeta \in [-2.27, -1.06]$
Rel imbalance (5.10)	GPD/GEV	N/A	0.0%	1.02	0.00	$\zeta \in [-1.90, -1.12]$
Mid-price (5.11)	GPD	Neg Bin	0.0%	1.00	0.00	$\zeta \in [-2.61, -0.99]$
Bid rel price LO (5.17)	GPD	Neg Bin	54.4%	0.01	0.00	$\zeta \in [-0.48, -0.05]$
Ask rel price LO (5.17)	GPD	Neg Bin	54.5%	0.01	0.00	$\zeta \in [-0.48, -0.08]$
Bid rel price C (5.18)	GPD	Neg Bin	59.4%	0.01	0.00	$\zeta \in [-0.53, -0.11]$
Ask rel price C (5.18)	GPD	Neg Bin	59.3%	0.01	0.00	$\zeta \in [-0.53, -0.21]$
Mid-price returns (5.12)	GPD	Pois	87.4%	1.00	0.00	$\zeta \in [-1.88, -1.00]$
Spread (5.13)	N/A	Pois	0.0%	0.01	0.00	
Traded MO price (5.14)	GPD	Neg Bin	86.9%	1.00	0.00	$\zeta \in [-2.56, -0.97]$
Volatility mid-price (5.15)	GPD/GEV	N/A	0.0%	1.00	0.00	$\zeta \in [-2.06, -1.00]$
Vol. mid-price ret. (5.16)	GPD/GEV	N/A	0.0%	1.00	0.00	$\zeta \in [-1.79, -1.00]$
Bid count MOLOC (5.19)	GPD	Neg Bin	0.0%	2.77	6.91	$\zeta \in [-0.08, 0.04]$
Ask count MOLOC (5.19)	GPD	Neg Bin	0.0%	2.73	6.60	$\zeta \in [-0.08, 0.04]$
Bid count LO (5.20)	GPD	Neg Bin	28.4%	2.02	2.70	$\zeta \in [-0.14, -0.04]$
Ask count LO (5.20)	GPD	Neg Bin	27.7%	1.98	2.54	$\zeta \in [-0.14, -0.06]$
Bid count C (5.21)	N/A	Pois	35.1%	1.77	1.73	
Ask count C (5.21)	N/A	Pois	35.4%	1.75	1.62	

As described in Section 5.3.2, we consider an additional assessment to help determine whether a discrete distribution, or an approximation by a continuous distribution is more appropriate for the marks. Figure 5.10 provides an example of Bid vol MOLOC. In this figure we present the histogram, empirical CDF plots and the theoretical CDFs for the generalized Pareto distribution, negative binomial distribution and Poisson distributions. Visually, the generalized Pareto distribution is the closest match to the empirical distribution. This is supported in Table 5.9, which shows the generalized Pareto distribution provides the smallest deviation across all time intervals and trading days considered.

For each distribution, we calculate the mean SSD in (5.23), which is a measure of deviation of the empirical CDF from the theoretical CDF, across all the time segments. The SSD presented in Table 5.9 is associated with the selected distribution which has the lowest difference presented in the final column. For the case studies in Section 5.3.4, the selected distribution is in agreement with the findings shown in Table 5.9.

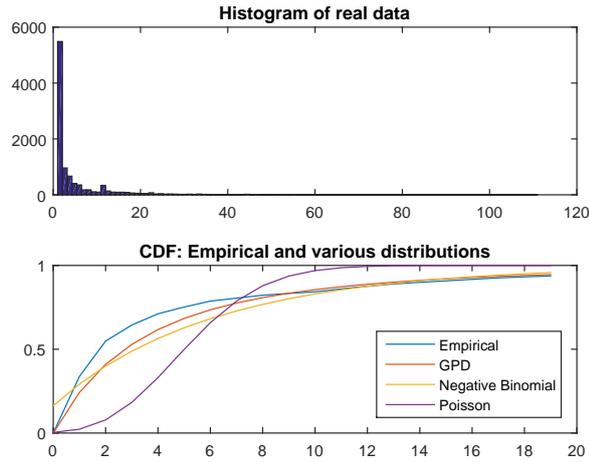


Figure 5.10: Histogram, empirical CDF plots and the theoretical CDFs for the generalized Pareto distribution, negative binomial distribution and Poisson distributions, for the mark, Bid vol MOLOC associated with levels 1:5, for SILVER on 31st July 2015.

Table 5.9: Mean values of the summary statistics of the mark vector, related to both bid and ask side events, for levels 1:5, across a 10 trading days (July 2015), for SILVER.

Mark	Range unique values	Mean unique values	Mean SSD	Lowest SSD Distribution
Bid depth (5.1)	[49, 125]	85.57	0.11	GPD
Ask depth (5.1)	[52, 127]	80.35	0.11	GPD
Bid opp. side depth (5.2)	[51, 117]	79.97	0.07	GPD
Ask opp. side depth (5.2)	[49, 127]	85.73	0.07	GPD
Bid vol MOLOC (5.3)	[58, 99]	76.32	0.06	GPD
Ask vol MOLOC (5.3)	[53, 93]	70.22	0.07	GPD
Bid vol MOLO (5.4)	[46, 84]	63.32	0.06	GPD
Ask vol MOLO (5.4)	[40, 86]	59.22	0.07	GPD
Bid vol MO (5.6)	[16, 29]	21.59	0.07	GPD
Ask vol MO (5.6)	[16, 36]	21.64	0.07	GPD
Bid vol LO (5.5)	[42, 82]	60.97	0.07	GPD
Ask vol LO (5.5)	[39, 79]	56.70	0.07	GPD
Bid vol C (5.7)	[22, 44]	28.81	0.06	Neg Bin
Ask vol C (5.7)	[21, 34]	26.0	0.06	Neg Bin
Inside vol MO (5.8)	[13, 24]	18.41	0.06	GPD
Imbalance (5.9)	[7960, 8915]	8499.70	N/A	
Rel imbalance (5.10)	[9360, 9913]	9839.50	N/A	
Mid-price (5.11)	[11, 25]	16.89	0.14	GPD
Bid rel price LO (5.17)	[164, 245]	204.05	0.36	GPD
Ask rel price LO (5.17)	[173, 247]	206.51	0.35	GPD
Bid rel price C (5.18)	[146, 242]	188.67	0.28	GPD
Ask rel price C (5.18)	[139, 228]	189.22	0.27	GPD
Mid-price returns (5.12)	[64, 225]	132.97	1.68	Pois
Spread (5.13)	[4, 12]	6.14	0.15	Pois
Traded MO price (5.14)	[37, 135]	75.59	1.53	Neg Bin
Volatility mid-price (5.15)	[6639, 8467]	7842.80	N/A	
Volat. mid-price ret. (5.16)	[1884, 3418]	2696	N/A	
Bid count MOLOC (5.19)	[21, 44]	25.29	0.04	Neg Bin
Ask count MOLOC (5.3)	[20, 31]	25.14	0.04	Neg Bin
Bid count LO (5.5)	[13, 25]	16	0.05	Neg Bin
Ask count LO (5.5)	[13, 19]	15.59	0.05	Neg Bin
Bid count C (5.7)	[10, 23]	13	0.05	Pois
Ask count C (5.7)	[9, 15]	12.27	0.05	Pois

Ljung-Box test for serial dependence

To investigate the serial dependence reported in the case studies in Section 5.3.4, we use the Ljung-Box test of serial dependence to assess each mark across the 10 trading days considered. Assessment is made on the transformed data in blocks of 10,000 events across a range of lags $h \in \{3, 5, 20, 50\}$.

Table 5.10 presents the percentage of times that the null hypothesis is rejected in favour of the data exhibiting serial dependence for each mark. Consistent with our observations above, serial dependence is present in almost all marks with exception to Bid/Ask vol MO and Inside vol MO. For the case of MO volumes, the frequency of these events relative to all events considered is very low, and agents executing trades would operate independently of one another with exogenous drivers behind the decision to execute an order.

Table 5.10: Ljung Box test, with percentage of times that the null hypothesis is rejected in favour of the data exhibiting serial dependence, for each mark and time segments of size 10,000, across 10 trading days.

Marks	Lag =3	Lag =5	Lag =20	Lag =50	Marks	Lag =3	Lag =5	Lag =20	Lag =50
Bid depth	100%	100%	100%	100%	mid-price	100%	100%	100%	100%
Ask depth	100%	100%	100%	100%	Bid rel price LO	100%	100%	100%	100%
Bid opp. depth	100%	100%	100%	100%	Ask rel price LO	100%	100%	100%	100%
Ask opp. depth	100%	100%	100%	100%	Bid rel price C	100%	100%	100%	100%
Bid vol MOLOC	100%	100%	100%	100%	Ask rel price C	100%	100%	100%	100%
Ask vol MOLOC	100%	100%	100%	100%	Mid-price returns	100%	100%	100%	100%
Bid vol MOLO	100%	100%	100%	100%	Spread	100%	100%	100%	100%
Ask vol MOLO	100%	100%	100%	100%	Traded MO price	84%	81%	65%	58%
Bid vol MO	29%	29%	36%	29%	Vol. mid-price	100%	100%	100%	100%
Ask vol MO	26%	29%	19%	19%	Vol. mid-price ret.	68%	68%	71%	77%
Bid vol LO	100%	100%	100%	100%	Bid cnt MOLOC	100%	100%	100%	100%
Ask vol LO	100%	100%	100%	100%	Ask cnt MOLOC	100%	100%	100%	100%
Bid vol C	100%	100%	100%	100%	Bid cnt LO	100%	100%	100%	100%
Ask vol C	100%	100%	100%	100%	Ask cnt LO	100%	100%	100%	100%
Inside vol MO	16%	19%	19%	13%	Bid cnt C	100%	100%	100%	97%
Imbalance	100%	100%	100%	100%	Ask cnt C	100%	100%	100%	100%
Rel imbalance	100%	100%	100%	100%					

Hurst exponent (long memory) for the mark vector

We extend the studies of serial dependence further by utilizing the Hurst exponent to investigate the presence of long memory in the marks. We implement the Hurst exponent estimation across the marks on both the bid and ask side for 10,000 event time segments, across 10 trading days in July 2015. Figures 5.11 and 5.12 show the Hurst exponents in a sequence of box plots comprised of estimates for the 10 trading days.

Recall from Section 5.3.2, values for the index in the range $0.5 < H < 1$ indicate a time series with long-term positive autocorrelation. This shows momentum in the mark, that is, high values in the series are likely to be succeeded by high values. The results show that the majority of the marks are on average well above 0.5, indicating long-term positive autocorrelation. Consistent with our findings from the Ljung-Box test for serial dependence in Table 5.10, the long term serial dependence is lower for Bid/Ask vol MO and Inside vol MO, and to a lesser degree, Traded MO price and Volatility mid-price ret.

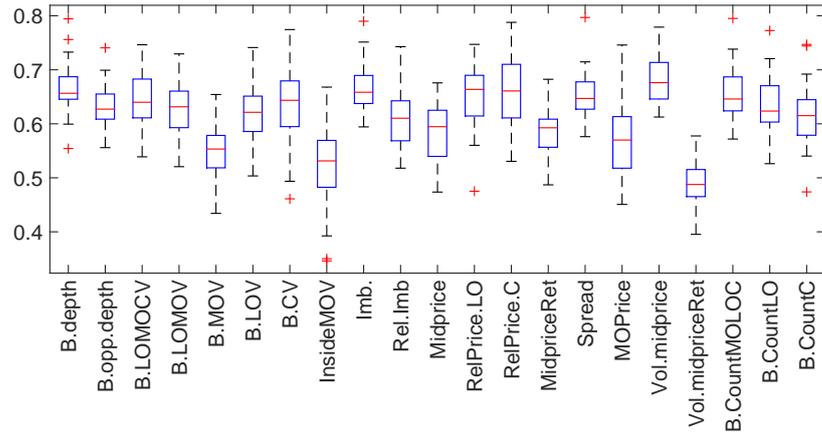


Figure 5.11: Boxplots of intra-day Hurst exponent across 10 trading days in July 2015, for bid side marks, for SILVER

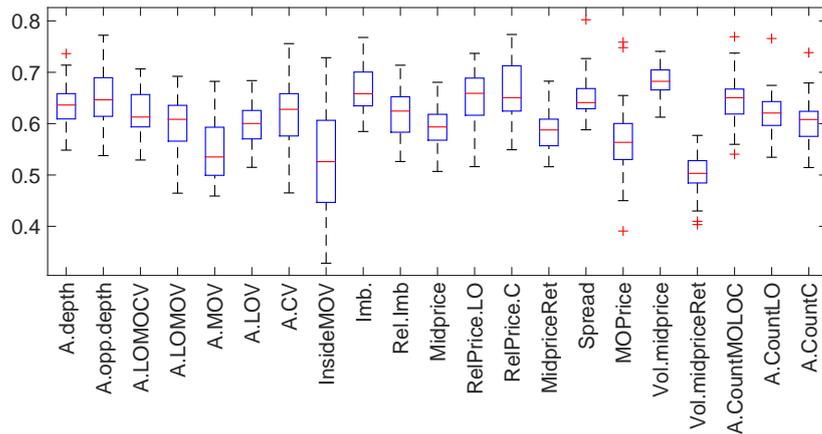


Figure 5.12: Boxplots of intra-day Hurst exponent across 10 trading days in July 2015, for ask side marks, for SILVER

Cross-correlation between mark vector

Figures 5.13 and 5.14 present the correlation between the marks for the bid side for SILVER on a single trading date, 31-July-2015. Both Pearson's correlation and Spearman's rank correlation are considered (Section 5.3.2) and are predominately in agreement within this study. The transformations to the data stated in Table 5.3 apply here, however it is not possible to assess the joint correlation with removal of zeros from the mark vector. We expect there to be strong correlation between the marks that are subsets of one another, for example, Bid/Ask vol MOLOC and Bid/Ask vol MOLO. For those cases, we will not comment further on any apparent correlations. Intuitively, we expect that Bid/Ask depth will be correlated to Imbalance, and that is in fact the case, with correlations on the bid side of $\rho^B(V_{t_i}^{(B)}, I_{t_i}) = 0.59$ and the ask side of $\rho^A(V_{t_i}^{(A)}, I_{t_i}) = 0.63$.

On both the bid and ask side of the LOB, the count based marks are correlated with the

volume based marks, as one would expect. However, the exception to this is the diminished correlation between the count based marks and Bid/Ask vol MO. Bid/Ask count MOLOC are most highly correlated with Bid/Ask vol MOLOC, $\rho^B(C_{t_i}^{(B,e)}, V_{t_i}^{(B,e)}) = 0.87$ and $\rho^A(C_{t_i}^{(A,e)}, V_{t_i}^{(A,e)}) = 0.88$. As the number of events increase, the total volume increases, implying some regularity of event size, rather than single large events entering the LOB.

With less effect, both counts and volumes that contain one of, or both $e \in \{LO, C\}$ are correlated with Rel price LO and Rel price C, within the range of $[0.38, 0.68]$.

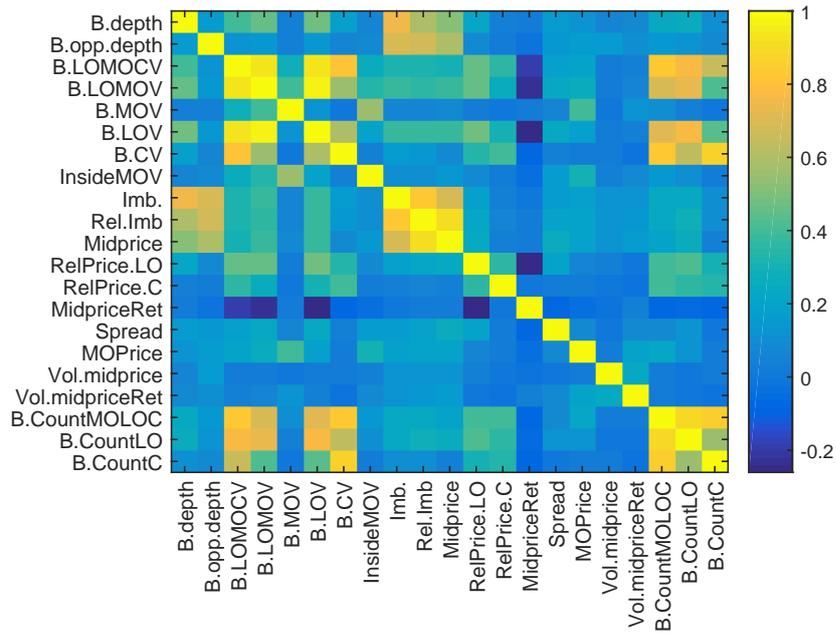


Figure 5.13: Pearson's linear correlation of the mark vector for events associated with levels 1:5, for SILVER, bid side, for the trading date 31-July-2015.

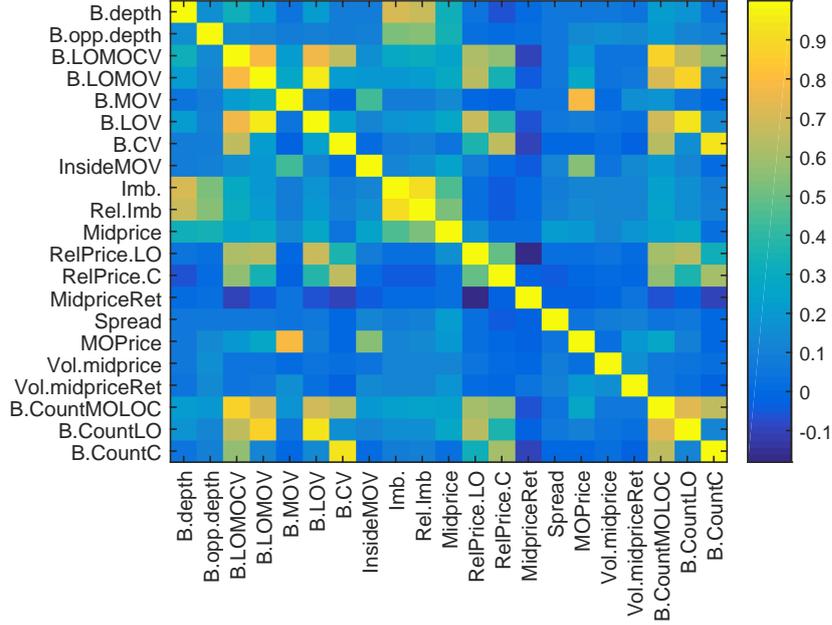


Figure 5.14: Spearman rank correlation of the mark vector for events associated with levels 1:5, for SILVER, bid side, for the trading date 31-July-2015.

5.4 Extensions to boost functions for partially observed marks and other variants

For a number of marks considered, a mark may not be recorded at every event time. For example, Bid/Ask vol MO will contain zeros at the event times that are related to LO and C only. Table 5.2 presents the frequency of zeros for each mark. If we consider the mark Bid/Ask vol MO, approximately 92% of the events times are zero, i.e. 92% of events are a LO and/or C. However, the mark Mid-price return has naturally occurring zeros that contain information about the mark and should not be excluded.

Marks such as Bid/Ask vol MO, should only boost the intensity function when they are non-zero. For any unique time t_i , there are six possible combinations of event types that can occur $q \in \{MO, MOLO, MOLOC, LO, LOC, C\}$. For volumes of these event type combinations, a possible linear boost function is

$$g(\mathbf{V}; \psi) = \frac{1 + \sum_{q=1}^6 \psi_q \mathbb{I}[q] V^q}{1 + \psi_q \mathbb{E}[V^q | q] P(q)}.$$

The advantage of this form of boost function is that the contribution of each event type combination can vary by the boost parameter ψ_q . For example, it is expected that a MO would have a much greater impact on boosting the intensity process, compared with a C, then this will be reflected by $\psi_{MO} > \psi_C$.

For positive random variables, our formulation allows the local intensity to be boosted or decreased depending on whether $h(\mathbf{X}; \psi) > \mathbb{E}_\phi[h(\mathbf{X}; \psi)]$, or not in (4.4). However,

the boost is always a positive number, so cannot lead to negative intensities. For random variables which can take negative values, it is possible that the boost itself will be negative. This has the potential to result in a negative intensity value, which may not have the interpretation as an intensity of a Poisson model.

For the purpose of this research, we adopt the standard approach of using a Hawkes process to model the increase in the intensity, with marks that contribute to boosting the intensity. For mark random variables that can take negative values, such as differenced marks, for example Bid/Ask depth and return series, such as Mid-price returns, there are various boost functions that may be appropriate.

For differenced marks and return based marks, the boost is a function of the size of the *change* in the mark, not the magnitude of the mark as in the case of volume of an event. A power law in (4.27) or an exponential boost in (4.28) would be appropriate boost functions that will give a positive boost. In a linear setting we could simply take the absolute value of the mark. This however assumes that an increase or decrease in the mark will have a symmetric boosting of the intensity process

$$g(V'; \psi) = \frac{1 + \psi|V'|}{1 + \psi\mathbb{E}[|V'|]},$$

where $V'_i = V_{t_i} - V_{t_{i-1}}$, being the differenced mark.

Still within the linear formulation, if the impact of the change in the mark on the intensity process is different for an increase versus a decrease, which is expected in the case of Bid/Ask depth and Imbalance measures, it is possible to adopt a boost function such as

$$g(\mathbf{V}; \psi) = \frac{1 + \psi_1\mathbb{I}[V' < 0]|V'| + \psi_2\mathbb{I}[V' > 0]V'}{1 + \psi_1\mathbb{E}[|V'| | V' < 0]P(V' < 0) + \psi_2\mathbb{E}[V' | V' > 0]P(V' > 0)}.$$

All other marks and transformed marks that are considered in Table 5.3, can be modelled by the standard boost functions presented in Section 4.2.1 and do not required special mention up-front. A study of a large variety of boost functions is beyond the scope of this thesis, but should be considered carefully when selecting the appropriate marks and their impact to the boosted intensity process.

5.5 Conclusion

This chapter considered the formal definition of a mark vector within the context of a marked Hawkes process. We considered the underlying assumptions required for the inclusion of a mark vector and what assumptions may need to be relaxed to achieve a model that better reflects what will be observed in practice.

Our search for appropriate marks was largely guided by research on the stylized features of the LOB, with very little literature on marked Hawkes process in financial application. We defined a large list of potential endogenous marks derived from the Reported LOB data. With many hundreds of potential variants of the marks, we limited our study of the properties of the marks, to those derived from the LOB of SILVER across 10 trading days

in July 2015. Upon assessment of the distributional properties and serial dependence of the marks, we observed numerous marks, such as Bid/Ask depth and Mid-price, required differencing to achieve stationarity.

Using multiple information criteria, we assessed the appropriate marginal distribution for the marks. Due to the presence of serial dependence, which was studied via ACF plots, Ljung-Box tests and the Hurst exponent, we extended the study further by simulating the mark with the chosen distribution and incorporating serial dependence into a simulation, in attempt to mimic the features of the mark. What became clear from the case studies is that fitting a parametric distribution to these complex data sets is non-trivial. In many cases (as summarized in Table 5.7) the marks cannot be sufficiently modelled by a parametric distribution. There are a selection of marks, after the incorporation of serial dependence, that are well modelled by a conditional generalized Pareto distribution, conditional negative binomial, or conditional Poisson distribution.

An assessment of pairwise correlation within the mark vector was made using both Pearson's correlation and Spearman's rank correlation. For a marked Hawkes process, we can model this joint dependence via a copula. We are careful not to draw conclusions about marks that contain subset marks, for example Bid/Ask MOLOC. We observe a consistency across the bid and ask side. Some notable pairwise correlations exist between Bid/Ask count MOLOC and Bid/Ask vol C. We also see that counts and volumes (which are both inherently correlated) are both correlated with the Rel price LO and Rel price C. Inside vol MO demonstrated some correlation with the current Mid-price.

In conclusion, the following attributes need to be considered when incorporating a LOB based mark vector into a Hawkes process,

1. A diversity of distributions needs to be considered for the marks and in many cases it is difficult to encapsulate the mark within a standard distribution, as shown in Table 5.7.
2. Most marks exhibit heavy tailed features and an appropriate marginal distribution is the generalized Pareto distribution. For the discrete valued marks, a negative binomial distribution is appropriate, but for Bid/Ask count C and Spread, the low counts and equality of the mean and variance suggests a Poisson distribution as the more appropriate distribution.
3. For non-stationary marks, first order differencing is required. Despite this differencing, we still observe persistent serial dependence across the majority of the marks. In addition, joint correlation exists within the mark vector.
4. The complexity of the mark vector and potential for the marks to be significant in dimension, highlights the necessity for not only a robust selection method of mark inclusion, without specification of the distribution of the mark, but also one which does not require the fitting under the joint model.

Chapter 6

The score test for detecting marks

As discussed in Section 1.6, if we exclude marks from the Hawkes process, we are failing to accurately capture the dynamics of the LOB in all applications, for example: the link between order flows and price formation; the endogenous nature of price fluctuations; predicting buying and selling intensities for construction of optimal liquidation strategies; measuring market reflexivity; and capturing volatility clustering. In Chapter 4 we demonstrated the incorporation of marks into the Hawkes process through simulation and various complex model formations. However, as discovered in Chapter 5, there are many possible marks that can be derived from the LOB data and the complex nature of the mark vector creates further challenges for incorporating marks into the Hawkes process.

The purpose of this chapter is to develop a novel formulation of the score (or Lagrange multiplier) test of the null hypothesis that the marks do not impact the intensity of the point process. We do this by formulating the impact of the marks through a boost function, which multiplicatively modifies the decay function specifying the intensity process as in Embrechts et al. (2011) and Liniger (2009), for example. Under this product form, we show that the score statistic has a particularly appealing form for practical application to complex model formulations, required for modelling the arrival of events in the LOB for high frequency financial data. The test statistic is easily constructed using parametric estimation of the non-marked intensity process that is estimated under the null hypothesis that marks do not impact the intensity, with separate estimation of certain moments of the marks parametrically or non-parametrically.

We refer the to Section 4.1 for a description of a univariate stationary marked Hawkes process, along with the likelihood and quasi likelihood for the non stationary process in \mathbb{R}_+ . Useful for the discussions below and to recap, a univariate marked Hawkes process, $N_g(t) \in \mathbb{N}$ is observed over the interval $t \in [0, T]$ and takes the value 0 at $t = 0$. The event times are denoted $0 = t_0 < t_1 < t_2 < \dots < t_n \leq T$ and a vector of d marks $\mathbf{X}_i \in \mathbb{X} \subset \mathbb{R}^d$ is associated with the i th event. The observed points of this process are $\{(t_i, \mathbf{x}_i), i = 1, \dots, n\}$.

The marked Hawkes process intensity is given in (4.14) and it will be assumed that the non-negative decay function w satisfies (4.2). Recall that the immigration rate is η , the branching coefficient ϑ and the decay function parameter is α . The marks \mathbf{X} have density $f(\mathbf{x}; \phi)$ and impact the intensity through the scalar valued boost function $g(\mathbf{X}; \phi, \psi)$, which

is parametrized with a vector ψ of length r specifying the way in which marks enter the boost function. As well, to obtain a stationary solution to (4.14), the boost function is normalized so that $\mathbb{E}_\phi[g(\mathbf{X}; \phi, \psi)] = 1$ for all ϕ and ψ .

The null hypothesis that marks do not boost the intensity is equivalent to $g(\mathbf{x}; \phi, \psi^*) \equiv 1$ for some suitable ψ^* . In this chapter we work with a general class of normalized boost functions, for which $H_0 : \psi^* = 0$ corresponds to the null hypothesis $g(\mathbf{X}; \phi, \psi^*) \equiv 1$. Under H_0 , the observed event times are those of an unmarked Hawkes process N with intensity denoted by

$$\lambda(t; \theta) = \eta + \vartheta \int_{[0,t)} w(t-s; \alpha) N(ds). \quad (6.1)$$

The score or Lagrange multiplier test is well known in statistics and econometrics and relies on the maximum likelihood estimation of the parameters of the model under the null hypothesis only. For example Solo and Pasha (2012) develop the score test for independence between a point process and a continuously observed analogue signal. Their process is not a marked process and because of this, their test statistic is not appropriate for testing the impact of marks as is considered in this research.

A major advantage of the score test compared with the likelihood ratio and Wald tests, is that it requires fitting of the point process only under the null hypothesis that marks do not impact the intensity in (6.1). The likelihood ratio and Wald tests both require maximum likelihood fitting of the marked point process using the intensity in (4.14), jointly with the marks density. Particularly when the form of boost function is moderately complicated or the dimension of the marks vector d is moderate or large, maximum likelihood fitting of the marked point process can be computationally challenging, as we discovered in Chapter 4 and more so than is presented in the literature. Whereas maximum likelihood fitting of the unmarked process is a much lower dimensional optimization. A major appeal of the procedure we propose, particularly in the context of modelling the events in a limit order book, is that many marks can be screened together or individually using a single fit of the unmarked intensity process, obviating the need to refit the intensity process for each choice of marks under consideration. From a theoretical perspective there are also advantages. Ergodic properties of unmarked processes are well established (Bremaud and Massoulié, 1996). Consistency and asymptotic normality for the likelihood estimates for an unmarked Hawkes process are also well established (Clinet and Yoshida, 2017; Ogata, 1978) and are relied on by us for the large sample distribution theory of our proposed score test, at least in the exponential decay case. However, as far as we are aware, there are no corresponding results for the substantially more difficult theory required for the marked self-exciting point process.

Within this chapter, the score is derived with respect to the boost parameters ψ and the corresponding block of the information matrix. We define the score test and detail the implementation of the score statistics. The asymptotic distribution is derived as the usual chi-squared distribution, with degrees of freedom equal to the number of parameters required to specify the boost function. Finally, a comprehensive set of simulation studies are presented under various combinations of boost functions and dependence features to show the finite sample size and power properties of the proposed test statistic.

6.1 Log-likelihood, score and information

Section 4.1.2 introduced the log-likelihood in the case where marks are conditionally independent given the event times. To proceed with the definition of the score test, let $\theta = (\eta, \vartheta, \alpha) \in \Theta$, $\phi \in \Phi$ and $\psi \in \Psi$ for some parameter spaces Θ , Φ and Ψ to be precisely defined below. Let $\nu = (\theta, \phi, \psi) \in \Theta \times \Phi \times \Psi$ be the collection of all parameters for the marked process with intensity function (4.14). We denote the true value of the parameters as $(\theta^*, \phi^*, \psi^*)$. When the null hypothesis holds, the true parameter vector is denoted $\nu^* = (\theta^*, \phi^*, 0)$.

Since the score test will be developed under the null hypothesis, we only require details of Θ in the derivation and theory to follow. Specifically, we let Θ be a finite dimensional relatively compact open subset of \mathbb{R}^K , $K > 1$. For example, for the exponential decay function $w(s; \alpha) = \alpha \exp(-\alpha s)$, $K = 3$ and we assume that $0 < \underline{\eta} \leq \eta \leq \bar{\eta} < \infty$, $0 < \underline{\vartheta} \leq \vartheta \leq \bar{\vartheta} < \infty$, $0 < \underline{\alpha} \leq \alpha \leq \bar{\alpha} < \infty$. The parameter space Φ for the marks density will typically be the natural space of parameters for the specified density, and the boost parameter space Ψ will be defined as required in the example below.

The log-likelihood was defined in (4.15) and is restated here for reference.

$$l_g(\nu) = \int_{[0, T] \times \mathbb{X}} \ln \lambda_g(t; \nu) N_g(dt \times d\mathbf{x}) - \Lambda_g(T; \nu) + \int_{[0, T] \times \mathbb{X}} \ln f(\mathbf{x}; \phi) N_g(dt \times d\mathbf{x}), \quad (6.2)$$

in which the compensator at T is

$$\Lambda_g(T; \nu) = \int_{[0, T]} \lambda_g(t; \nu) dt. \quad (6.3)$$

The likelihood was derived assuming the mark vectors for each event time are conditionally independent and identically distributed given the event times. We will maintain that assumption for the derivations to follow. Later we generalize this to include serially dependent marks. However, this generalization does not result in a different definition of the score test statistic defined in Section 6.2.

When $\psi = 0$, the boost is the identity, marks do not impact the intensity, and the marks process and the event process are independent. Hence, the log-likelihood in (6.2) separates as the sum of a contribution from terms involving the intensity without marks and the log-likelihood for the marks distribution as

$$l(\theta, \phi) = l(\theta) + \sum_{i=1}^n \ln f(\mathbf{x}_i; \phi), \quad (6.4)$$

where the log-likelihood for the unmarked process $N(t)$ is

$$l(\theta) = \int_{[0, T]} \ln \lambda(s; \theta) N(ds) - \Lambda(T; \theta), \quad (6.5)$$

with corresponding compensator

$$\Lambda(T; \theta) = \int_{[0, T]} \lambda(s; \theta) ds. \quad (6.6)$$

Hence, under H_0 , the parameters of the unmarked Hawkes process are decoupled from those of the marks distribution, so that these can be separately estimated.

For the definition of the score statistic the score with respect to ψ evaluated under H_0 is needed. In general

$$\partial_\psi l_g(\nu) = \int_{[0, t] \times \mathbb{X}} \lambda_g(t; \theta)^{-1} \partial_\psi \lambda_g(t; \nu) N_g(ds \times d\mathbf{x}) - \int_{[0, T]} \partial_\psi \lambda_g(t; \nu) dt, \quad (6.7)$$

where

$$\partial_\psi \lambda_g(t; \nu) = \vartheta \int_{[0, t] \times \mathbb{X}} w(t - s; \alpha) \partial_\psi g(\mathbf{x}; \phi, \psi) N_g(ds \times d\mathbf{x}). \quad (6.8)$$

We use the notation N_g^0 for the point process generated under H_0 . This point process has an event intensity identical to N defined in (6.1). Marks are observed at the event times of this process, but do not impact the intensity of it. Using this notation, the derivative of (6.2) with respect to ψ at any values of θ , ϕ and evaluated under H_0 , is expressible as

$$\partial_\psi l_g(\theta, \phi, 0) = \int_{[0, T]} \lambda(t; \theta)^{-1} \partial_\psi \lambda_g(t; \theta, \phi, 0) N(dt) - \int_{[0, T]} \partial_\psi \lambda_g(t; \theta, \phi, 0) dt, \quad (6.9)$$

in which

$$\partial_\psi \lambda_g(t; \theta, \phi, 0) = \vartheta \int_{[0, t] \times \mathbb{X}} w(t - s; \alpha) \partial_\psi g(t; \theta, \phi, 0) N_g^0(ds \times d\mathbf{x}). \quad (6.10)$$

When evaluated at the true parameter vector, $\nu^* = (\theta^*, \phi^*, 0)$ under H_0 , the score (6.9) can be written as

$$\partial_\psi l_g(\nu^*) = \int_{[0, T]} \lambda(t; \theta^*)^{-1} \partial_\psi \lambda_g(t; \nu^*) \tilde{N}(dt), \quad (6.11)$$

where $\tilde{N}(dt) = N(dt) - \lambda(t; \theta^*) dt$. The information matrix for all parameters ν is

$$\mathcal{I}(\nu) = -\mathbb{E}_\nu[\partial_{\nu\nu}^2 l_g(\nu)] = \mathbb{E}_\nu[\partial_\nu l_g(\nu) \partial_\nu l_g(\nu)^\top]. \quad (6.12)$$

6.2 The score test

This section outlines the properties we require for the boost function, which was previously introduced in Section 4.1. These properties allow the general score and information matrix introduced in the Appendix B.1, to be simplified, leading to a definition of the general form of the proposed score test. Section 6.2.3 gives a representation of the score test in terms the event times and observed mark values at these event times, which is used for computational purposes.

6.2.1 Boost function details

In Section 4.1 we presented the general normalized boost function, and expressions for the additive and multiplicative functions of the marks. This is restated for flow of the discussions to follow. Quite general normalized boost functions can be constructed by starting with a function $h(\mathbf{X}, \psi)$ and

$$g(\mathbf{X}; \psi, \phi) = \frac{h(\mathbf{X}; \psi)}{\mathbb{E}_\phi[h(\mathbf{X}; \psi)]}. \quad (6.13)$$

Boost functions which are additive in functions of the marks are specified as

$$h(\mathbf{X}; \psi) = 1 + \sum_{j=1}^r \psi_j h_j(\mathbf{X}), \quad (6.14)$$

or boosts which are multiplicative in the functions of the marks are specified as

$$h(\mathbf{X}; \psi) = \prod_{j=1}^r (1 + \psi_j h_j(\mathbf{X})). \quad (6.15)$$

For the derivation of the score vector in (6.7), we require the vector of derivatives of g with respect to ψ denoted

$$\partial_\psi g(\mathbf{X}; \phi, \psi) = \frac{1}{\mathbb{E}_\phi[h(\mathbf{X}; \psi)]} [\partial_\psi h(\mathbf{X}; \psi) - g(\mathbf{X}; \phi, \psi) \mathbb{E}_\phi[\partial_\psi h(\mathbf{X}; \psi)]]. \quad (6.16)$$

Condition 1 (Conditions on boost function specification). *Throughout we assume h and its first and second derivatives with respect to ψ , denoted $\partial_\psi h$ and $\partial_{\psi\psi}^2 h$, exist and satisfy the following properties:*

- (i) $h(\mathbf{X}; 0) \equiv 1$;
- (ii) $H(\mathbf{X}) := \partial_\psi h(\mathbf{X}; 0)$ and $H'(\mathbf{X}) := \partial_{\psi\psi}^2 h(\mathbf{X}; 0)$ are assumed to be functions only of \mathbf{X} ;
- (iii) $\mathbb{E}_\phi[h(\mathbf{X}; \psi)]$ and $\mathbb{E}_\phi[\partial_\psi h(\mathbf{X}; \psi)]$ exist for all $\psi \in \Psi$, $\phi \in \Phi$;
- (iv) $\partial_\phi \mathbb{E}_\phi(h(\mathbf{X}; \psi))|_{\nu^*} = 0$;
- (v) $\partial_\phi \mathbb{E}_\phi[\partial_\psi h(\mathbf{X}; \psi)]|_{\nu^*}$ exists for all $\phi \in \Phi$.

(vi) $\text{cov}_\phi(H(\mathbf{X})) = \Omega_G(\phi)$ where $\Omega_G(\phi)$ is a finite positive definite matrix for any $\phi \in \Phi$.

Within this research we consider boost functions that are additive in functions of the marks, specified in (6.14) and multiplicative in functions of marks, specified in (6.15). For the additive specifications in (6.14), where $H_j(0) = 1$ for $j = 1, \dots, r$, the parameters ψ enter linearly, the Condition 1 parts (i) to (v) are satisfied. For the linear and multiplicative specifications (6.14) and (6.15) we get $H(\mathbf{X}) = [H_1(\mathbf{X}), \dots, H_r(\mathbf{X})]^\top$. These can also be included in the additive and multiplicative forms. Condition 1 part (i) is without loss of generality, since the boost function is normalized. Part (ii) to (v) are also satisfied for other non-linear specifications of h applied to scalar marks, such as the exponential, $h(X; \psi) = \exp(\psi H(X))$, which includes the power boost function, $g(X) = X^\psi$ by putting $H(X) = \log(X)$.

Recall, that based on the properties required of h , $\mathbb{E}_\phi[g(\mathbf{X}; \phi, \psi)] = 1$ for all $\psi \in \Psi$, $g(\mathbf{X}; \phi, 0) \equiv 1$, and letting $G(\mathbf{X}, \phi) = \partial_\psi g(\mathbf{X}; \phi, 0)$, $\mathbb{E}_\phi[G(\mathbf{X}; \phi)] = 0$. With the above specification, the null hypothesis of marks not impacting the intensity is achieved by setting $\psi = 0$. Note that $G(\mathbf{X}; \phi) = H(\mathbf{X}) - \mathbb{E}_\phi[H(\mathbf{X})]$ is a vector of dimension r , comprised of functions of the components of the vector mark centred at their expectations. The requirements that $\mathbb{E}_\phi[h(\mathbf{X}; \psi)]$, $\mathbb{E}_\phi[\partial_\psi h(\mathbf{X}; \psi)]$ and $\partial_\phi \mathbb{E}_\phi[\partial_\psi h(\mathbf{X}; \psi)]|_{\nu^*}$ exist imposes obvious conditions on the marginal distribution of \mathbf{X} . Recall that if $h(\mathbf{X}; \psi)$ is a polynomial of degree p in \mathbf{X} then $\mathbb{E}_\phi[\mathbf{X}^p]$ needs to exist. Condition 1 parts (iv) and (v) are required in order that the information matrix for all parameters in the full model likelihood is block diagonal, allowing simplification of the score statistic defined below. For the derivation of the large sample distribution of the score statistic, we require additional moment conditions on the functions $G(\mathbf{X}; \phi)$.

6.2.2 Definition of the score test

Let $\hat{\nu}_T = (\hat{\theta}_T, \hat{\phi}_T, 0)$ where $\hat{\theta}_T$ is the quasi asymptotic maximum likelihood estimates, as in Clinet and Yoshida (2017, page 1804), based on the likelihood (6.5) under H_0 of the intensity process parameters, and $\hat{\phi}_T$ is the MLE for the parameters of the marks density. Denote the derivatives of the log-likelihood for process $N_g(t)$ with respect to ν as $\partial_\nu l_g(\nu)$ at the parameter value ν , so that $\partial_\theta l_g(\nu^*)$ and $\partial_\nu l_g(\hat{\nu}_T)$ are evaluated at ν^* and $\hat{\nu}_T$, respectively. The score (or Lagrange multiplier) test statistic (Breusch and Pagan, 1980) is defined as

$$\hat{Q}_T = \partial_\nu l_g(\hat{\nu}_T)^\top \mathcal{I}(\hat{\nu}_T)^{-1} \partial_\nu l_g(\hat{\nu}_T), \quad (6.17)$$

where the information matrix for all parameters ν is

$$\mathcal{I}(\nu) = -\mathbb{E}_\nu[\partial_{\nu\nu}^2 l_g(\nu)] = \mathbb{E}_\nu[\partial_\nu l_g(\nu) \partial_\nu l_g(\nu)^\top], \quad (6.18)$$

and $\mathcal{I}(\nu^*)$ and $\mathcal{I}(\hat{\nu}_T)$ evaluates this at the true parameters ν^* , and those estimated under H_0 , $\hat{\nu}_T$. Also the information matrix can be replaced by any matrix with the same limit in probability (Breusch and Pagan, 1980), for example the negative of the matrix of second derivatives of the log-likelihood, and the large sample properties of the score statistic will be the same.

Under $H_0 : \psi = 0$ it can be shown that (6.18) is block diagonal with diagonal blocks corresponding to parameters θ , ϕ and ψ , respectively and evaluated using $\psi = 0$. In particular, to show that the off diagonal block corresponding to $\mathbb{E}_\nu[\partial_{\phi\psi}^2 l_g(\nu)]$ is 0 evaluated under H_0 , requires Condition 1: (iv) & (v). Details are in Appendix B.1. Noting that $\partial_\nu l_g(\hat{\nu}_T) = (0, 0, \partial_\psi l_g(\hat{\nu}_T))^\top$, the general formula for the score test (6.17) simplifies to

$$\hat{Q}_T = \partial_\psi l_g(\hat{\nu}_T)^\top \hat{\mathcal{I}}_{\psi,\psi}(\hat{\nu}_T)^{-1} \partial_\psi l_g(\hat{\nu}_T). \quad (6.19)$$

Under the above assumptions on the way in which the marks impact the intensity, the form of boost functions g defined above and their required properties, we have

$$\begin{aligned} \mathcal{I}_{\psi,\psi}(\nu^*) &= E_{\nu^*}[\partial_\psi l_g(\nu^*) \partial_\nu l_g(\nu^*)^\top] \\ &= \mathbb{E}_{\nu^*} \left[\left(\int_{[0,T]} \lambda(t; \theta^*)^{-1} \partial_\psi \lambda_g(t; \nu^*) \tilde{N}(dt) \right)^{\otimes 2} \right] \\ &= \mathbb{E}_{\nu^*} \left[\int_{[0,T]} \lambda(t; \theta^*)^{-1} \partial_\psi \lambda_g(t; \nu^*)^{\otimes 2} dt \right], \end{aligned} \quad (6.20)$$

or, alternatively,

$$\begin{aligned} \mathcal{I}_{\psi,\psi}(\nu^*) &= -\mathbb{E}_{\nu^*}[\partial_{\psi\psi}^2 l_g(\nu^*)] \\ &= -\mathbb{E}_{\nu^*} \left[\int_{[0,T]} \lambda(t; \theta^*)^{-1} (\partial_{\psi\psi}^2 \lambda_g(t; \nu^*)) \tilde{N}(dt) \right] \\ &\quad + \mathbb{E}_{\nu^*} \left[\int_{[0,T]} \lambda(t; \theta^*)^{-2} (\partial_\psi \lambda_g(t; \nu^*))^{\otimes 2} N(dt) \right]. \end{aligned} \quad (6.21)$$

Now

$$\partial_{\psi\psi}^2 g(X; \phi, \psi)|_\nu^* = H'(X) - \mathbb{E}[H'(X)] - 2\mathbb{E}[H(X)](H(X) - \mathbb{E}_{\phi^*}[H(X)]),$$

which has expectation zero, hence

$$\mathbb{E}[\partial_{\psi\psi}^2 \lambda_g(t; \nu^* | \mathcal{F}_{t-})] = 0,$$

so that the first term in (6.21) is zero. Hence

$$\mathcal{I}_{\psi,\psi}(\nu^*) = -\mathbb{E}_{\nu^*}[\partial_{\psi\psi}^2 l_g(\nu^*)] = \mathbb{E}_{\nu^*} \left[\int_{[0,T]} \lambda(t; \theta^*)^{-2} (\partial_\psi \lambda_g(t; \nu^*))^{\otimes 2} N(dt) \right], \quad (6.22)$$

which is equal to (6.20) by using $\tilde{N}(dt) = N(dt) - \lambda(t; \theta^*)dt$ in (6.22).

When the marks are independent and identically distributed, the expectation in (6.22) can be simplified by first calculating the expected value of $(\partial_\psi \lambda_g(t; \nu^*))^{\otimes 2}$ with respect to

the marks conditional on the point process events to give

$$\mathcal{I}_{\psi,\psi}(\nu^*) = \mathbb{E}_{\nu^*} \left[\int_{[0,T]} \lambda(t; \theta^*)^{-2} \left\{ \vartheta^{*2} \int_{[0,t]} w(t-s; \alpha^*)^2 N(ds) \right\} N(dt) \right] \times \Omega_G^*, \quad (6.23)$$

where $\Omega_G^* = \mathbb{E}_{\phi^*} [G(\mathbf{X}; \phi^*) G(\mathbf{X}; \phi^*)^\top]$.

To implement the score test, $I_{\psi,\psi}(\hat{\nu}_T)$ needs to be evaluated. Since the expectation required is not computable in closed form, we suggest empirical evaluation based on (6.22) using $\hat{\mathcal{I}}_{\psi,\psi}(\hat{\nu})$, where some numerical procedure is required. Obviously a Monte Carlo evaluation based on repeated samples from the point process could be used. However this requires a model for the marks distribution to be specified. Later we implement the score test using a non-parametric approach, which can avoid modelling the marks distribution. Since in our application T is large we suggest

$$\hat{\mathcal{I}}_{\psi,\psi}(\hat{\nu}) = \int_{[0,T]} \lambda(t; \hat{\theta})^{-2} (\partial_\psi \lambda_g(t; \hat{\nu}))^{\otimes 2} N(dt). \quad (6.24)$$

Under the assumption that the marks are independent and identically distributed, we could also use the following empirical estimates of (6.23) as

$$\hat{\mathcal{I}}_{\psi,\psi}(\hat{\nu}) = \int_{[0,T]} \lambda(t; \hat{\theta})^{-2} \left\{ \hat{\vartheta}^2 \int_{[0,t]} w(t-s; \hat{\alpha})^2 N(ds) \right\} N(dt) \times \hat{\Omega}_G, \quad (6.25)$$

and $\hat{\Omega}_G$ is any consistent estimator of the Ω_G . Note the proposed empirical estimates simply use an empirical time average in place of the expectation. Ergodicity of the marks ensures that (6.24) and (6.22), when normalized by T^{-1} , have the same limit as $T \rightarrow \infty$. We also present an alternate, asymptotically equivalent form of these empirical estimates of $\mathcal{I}_{\psi,\psi}(\hat{\nu})$ in the next section.

6.2.3 Implementation of the score statistic

To recap notation that is useful for the discussions below, N_g^0 is the point process generated under H_0 , observed over the interval $t \in [0, T]$ and a vector of d marks $\mathbf{X}_i \in \mathbb{X} \subset \mathbb{R}^d$ is associated with the i th event. By interchanging the order of integration, (6.11) can be equivalently expressed as

$$\partial_\psi l_g(\nu^*) = \vartheta^* \int_{[0,T] \times \mathbb{X}} A_T(t; \theta^*) G(\mathbf{x}; \phi^*) N_g^0(dt \times d\mathbf{x}), \quad (6.26)$$

where

$$\begin{aligned} A_T(t; \theta^*) &= \int_{(t,T]} \lambda(s; \theta^*)^{-1} w(s-t; \alpha^*) \tilde{N}(ds) \\ &= \int_{(t,T]} \lambda(s; \theta^*)^{-1} w(s-t; \alpha^*) N(ds) - W(T-t; \alpha^*), \end{aligned} \quad (6.27)$$

where $\tilde{N}(ds) = N(ds) - \lambda(s; \theta^*)ds$ and $W(s) = \int_0^s w(u)du$ is the cumulative decay function which satisfies conditions in (4.2). A common example is the cumulative exponential decay function $W(s) = 1 - e^{-\alpha s}$.

To simplify computations as in Ozaki (1979), we assume $T = t_n$ when calculating the log-likelihood in (6.26), (6.27). This approximation has negligible impact for large values of T . Correspondingly, the alternative form of the score vector (6.26), can be evaluated using the observed event times $0 < t_1 < t_2 < \dots < t_n = T$, and the functions $G(\mathbf{X}_m; \phi^*)$ of the observed marks as a summation

$$\partial_\psi l_g(\nu^*) = \vartheta^* \sum_{m=1}^{n-1} A_{n,m}(\theta^*) G(\mathbf{X}_m; \phi^*), \quad (6.28)$$

where

$$A_{n,m}(\theta^*) := A_{t_n}(t_m; \theta^*) = \sum_{i=m+1}^n \frac{1}{\lambda(t_i; \theta^*)} w(t_i - t_m; \alpha^*) - W(t_n - t_m; \alpha^*). \quad (6.29)$$

Note that the summation in (6.28) has upper limit $n - 1$ because $A_{n,n}(\theta^*) = 0$. Also, if $T > t_n$ was used, there would be the negligible additional term in (6.28), given by $\vartheta^* W(T - t_n) G(\mathbf{X}_n)$. Note that $G(\mathbf{X}_m; \phi^*)$ and $A_{n,m}(\theta^*)$ have no parameters in common.

The representation of the score vector using (6.28) and (6.29), can also be derived by first writing the score vector (6.11) using summations over event times as follows (once more making the asymptotically negligible simplification that $t_n = T$)

$$\partial_\psi l_g(\nu^*) = \sum_{i=1}^n \lambda(t_i, \theta^*)^{-1} \vartheta^* \left\{ \sum_{j=1}^{i-1} w(t_i - t_j; \alpha^*) - W(t_n - t_i; \alpha^*) \right\} G(\mathbf{X}_i; \phi^*), \quad (6.30)$$

and then interchanging the order of summation.

Let the normalized score vector evaluated under H_0 be

$$S_n(\nu^*) = \frac{1}{\sqrt{n}\vartheta^*} \partial_\psi l(\nu^*) = \frac{1}{\sqrt{n}} \sum_{m=1}^{n-1} A_{n,m}(\theta^*) G(\mathbf{X}_m; \phi^*). \quad (6.31)$$

Since $\mathbb{E}_{\theta^*}[A_{n,m}(\theta^*)] = 0$ and $\mathbb{E}_{\phi^*}[G(\mathbf{X}_m, \phi^*)] = 0$, we have $\mathbb{E}_{\phi^*}[S_n(\theta, \phi^*)] = 0$ and the covariance matrix of S_n is given when assuming the i.i.d. properties of the marks

$$\Omega_n = \mathbb{E}_{\theta^*} \left[\frac{1}{n} \sum_{l=1}^{n-1} A_{n,l}(\theta^*)^2 \right] \Omega_G^*. \quad (6.32)$$

This is easily estimated using empirical estimates for the various quantities. When quantities are evaluated at the quasi-likelihood estimates of the parameters $\hat{\theta}_T$, we use the abbreviated notation, $\hat{A}_{n,m}$ to denote $A_{n,m}(\hat{\theta}_T)$. We let \hat{G}_m represent $G(\mathbf{X}_m, \hat{\phi}_T)$, $\hat{\Omega}_G$ for $\Omega_G(\hat{\phi}_T)$, \hat{S}_n for the normalized score vector $S_n(\hat{\theta}_T, \hat{\phi}_T)$ so that

$$\hat{S}_n = \frac{1}{\sqrt{n}} \sum_{m=1}^{n-1} \hat{A}_{n,m} \hat{G}_m, \quad (6.33)$$

and

$$\hat{\Omega}_n = \frac{1}{n} \sum_{m=1}^{n-1} \hat{A}_{n,m}^2 \hat{\Omega}_G, \quad (6.34)$$

so that the score statistic (6.19) is computed with

$$\hat{Q}_n = \hat{S}_n^\top \hat{\Omega}_n^{-1} \hat{S}_n. \quad (6.35)$$

An alternative to using theoretical moments evaluated at $\hat{\phi}_T$ for calculating $G(\mathbf{X}, \phi)$ and $\Omega_G(\phi)$, is to use sample moments. In this case we continue to use the same notation putting $\hat{G}_m = H(\mathbf{X}_m) - \bar{H}$, where \bar{H} is the vector of empirical means of the components of H and $\hat{\Sigma}_{G,n}$ is evaluated using the observed covariance matrix of the samples \hat{G}_m . Using empirical moments has the advantage that a density for the marks is not required to be specified for implementation of the score test. Both approaches to obtaining \hat{G}_m and $\hat{\Sigma}_G$ yields the same asymptotic chi-squared distribution for the score test, provided the moments of required orders are consistently estimated. Using simulations, we verify that the empirical power of the test is unaffected and for use of the score test in practice, there seems to be little advantage in fitting a distribution to the marks for the purpose of screening.

6.3 Asymptotic distribution of the score statistic

The proof that the score statistic \hat{Q}_T has a large sample chi-squared distribution under the null hypothesis relies on an extension to the large sample results presented in Clinet and Yoshida (2017) for the unmarked process. Specifically we assume the following conditions.

Condition 2 (Conditions presented in Clinet and Yoshida (2017)). *Conditions on the family $(\lambda(t, \theta))_{t \in \mathbb{R}_+, \theta \in \Theta}$:*

[A1] *The mapping $\lambda : \Omega \times \mathbb{R}_+ \times \Theta \rightarrow \mathcal{F} \otimes \mathbf{B}(\mathbb{R}_+) \otimes \mathbf{B}(\Theta)$ -measurable. Moreover, almost surely,*

i for any $\theta \in \Theta$, $s \rightarrow \lambda(s, \theta)$ is left continuous.

ii for any $s \in \mathbb{R}_+$, $\theta \rightarrow \lambda(s, \theta)$ is in $C^3(\Theta)$, and admits a continuous extension to $\bar{\Theta}$.

[A2] *The intensity processes and their first derivatives satisfy*

i for any $p > 1$, $\sup_{t \in \mathbb{R}_+} \sum_{i=0}^3 \|\sup_{\theta \in \Theta} |\partial_\theta^i \lambda(t, \theta)|\|_p < +\infty$.

ii for any $p > 1$, for any $\alpha \in \mathbf{I}$, $\sup_{t \in \mathbb{R}_+} \|\sup_{\theta \in \Theta} |\lambda^\alpha(t, \theta)^{-1}| 1_{\{\lambda^\alpha(t, \theta) \neq 0\}}\|_p < +\infty$.

iii For any $\theta \in \Theta$, for any $\alpha \in \mathbf{I}$, $\lambda^\alpha(t, \theta) = 0$ if and only if $\lambda^\alpha(t, \theta^) = 0$.*

[A3] *For any $\alpha \in \mathbf{I}, \theta \in \Theta$, the triplet $(\lambda^\alpha(\cdot, \theta^*), \lambda^\alpha(\cdot, \theta), \partial_\theta \lambda^\alpha(\cdot, \theta))$ is ergodic in the sense of Clinet and Yoshida (2017, Definition 3.1). In other words, there exists a mapping*

$\pi_\alpha : C_b(E, \mathbb{R}) \times \Theta \rightarrow \mathbb{R}$ such that for any $(\psi, \theta) \in C_b(E, \mathbb{R}) \times \Theta$, the following convergence holds:

$$\frac{1}{T} \int_0^T \psi(\lambda^\alpha(s, \theta^*), \lambda^\alpha(s, \theta), \partial_\theta \lambda^\alpha(s, \theta)) ds \xrightarrow{P} \pi_\alpha(\psi, \theta).$$

[A4] For any $\theta \in \Theta - \{\theta^*\}$, $\mathbb{Y}(\theta) \neq 0$.

These generalize Conditions A, B and C of Ogata (1978) applied to the intensity process defined in (6.1) for the unmarked process. Ogata (1978) provided the first consistency and asymptotic normality results for the unmarked Hawkes process and verified that his conditions apply to the exponential decay function $w(t; \alpha)$. Clinet and Yoshida (2017) also verify their conditions for the exponential decay function case. As far as we are aware, there has been no published verification of the conditions of Ogata (1978) or Clinet and Yoshida (2017) for the power law decay function. Under Condition 2, Clinet and Yoshida (2017) show (Theorem 3.9) that any asymptotic QMLE $\hat{\theta}_T$ is consistent, $\hat{\theta}_T \xrightarrow{P} \theta^*$, and (Theorem 3.11) asymptotically normal $\sqrt{T}(\hat{\theta}_T - \theta^*) \xrightarrow{d} \Gamma^{-\frac{1}{2}} \zeta$ where ζ has a standard multivariate normal distribution and Γ is the asymptotic information matrix, assumed to be positive definite.

In order to establish the asymptotic distribution of the score vector with respect to ψ , an extension to Condition 2 [A2] (i) is required: The following condition relates the existence of moments of the $G(\mathbf{X})$ (evaluated at ϕ^* , the true value under H_0) to the uniformity of derivatives of the the decay function.

Condition 3. For $p = \dim(\Theta) + 1$, consider the univariate marked Hawkes process N_g , with intensity process λ_g , where the intensity process satisfies

$$\sup_{t \in \mathbb{R}_+} \sum_{i=0}^2 \left\| \sup_{\theta \in \Theta} |\partial_\theta^i (\partial_\psi \lambda_g(t; \nu))|_{(\theta, \phi, 0)} \right\|_p < \infty. \quad (6.36)$$

Using extensions to the proof of lemma [A5] of Clinet and Yoshida (2017), which shows their Condition 2 [A2] is satisfied for the exponential decay unmarked Hawkes process, it is shown in Dunsmuir et al. (2018) (in preparation), that this holds for the exponential decay function model with $p = 4$ and marks which satisfy the statement of Proposition 1, in which case fourth moments of $G(\mathbf{X})$ are required to exist.

We now state the main result.

Proposition 1. Assume that the marks \mathbf{X}_n are observations on a stationary ergodic process with $\mathbb{E}[|G(\mathbf{X})|^4] < \infty$ and Conditions 1, 2 and 3 are satisfied. Under H_0 , the score statistic defined in (6.19) with information matrix $\mathcal{I}_{\psi, \psi}(\hat{\nu}_T)$ estimated by (6.24) satisfies

$$\hat{Q}_T \xrightarrow{d} \chi_{(r)} \quad \text{as } T \rightarrow \infty, \quad r = \dim(\psi). \quad (6.37)$$

An outline of the proof follows which will appear in Dunsmuir et al. (2018) (in preparation). The approach follows somewhat closely to that of Clinet and Yoshida (2017). We first consider the normalized process corresponding to (6.11) and for any non zero vector

of constants $c \in \mathbb{R}^r$ define the process

$$S_u^T = \frac{1}{\sqrt{T}} \int_{[0, uT]} \lambda(t; \theta^*)^{-1} c^\top \partial_\psi \lambda_g(t; \nu^*) \tilde{N}(dt), \quad (6.38)$$

where $u \in [0, 1]$. Note that $S_1^T = \frac{1}{\sqrt{T}} c^\top \partial_\psi l_g(\nu^*)$. Similarly to Clinet and Yoshida (2017), we establish a functional CLT when $T \rightarrow \infty$

$$(S_u^T)_{u \in [0, 1]} \xrightarrow{d} \Omega^{1/2} (W_u)_{u \in [0, 1]}, \quad (6.39)$$

where W is standard Brownian motion (and convergence is in the Skorokhod space $\mathbf{D}([0, 1])$) and

$$\frac{1}{T} \mathcal{I}_\psi(\nu^*) \xrightarrow{P} \Omega, \quad (6.40)$$

where Ω is a symmetric positive definite matrix. As in Clinet and Yoshida (2017), Jacod and Shiryaev (2013, 3.24 Chapter VIII) is used to establish (6.39). To complete the proof we show that the estimated information matrix $\hat{\mathcal{I}}_\psi^{(2)}$, defined in (6.24) satisfies

$$\frac{1}{T} \{\hat{\mathcal{I}}_\psi(\hat{\psi}) - \mathcal{I}_\psi(\nu^*)\} \xrightarrow{P} 0, \quad (6.41)$$

and hence $\frac{1}{T} \hat{\mathcal{I}}_\psi(\hat{\psi}) \xrightarrow{P} \Omega$. The proof of this uses a Taylor series expansion with respect to the estimates $\hat{\theta}$ around θ^* and ergodicity in a similar way to Clinet and Yoshida (2017, Lemma 3.12).

$$\frac{1}{\sqrt{T}} (\partial_\psi l_g(\hat{\nu}) - \partial_\psi l_g(\nu^*)) \xrightarrow{P} 0. \quad (6.42)$$

The result (6.42) also uses a Taylor series expansion with respect to θ , consisting of estimates of $\mathbb{E}[G(\mathbf{X})]$ and ergodicity again, similarly to Clinet and Yoshida (2017). Proposition 1 also applies to the implemented version (6.35), which is equivalent to \hat{Q}_T up to some asymptotically negligible terms as discussed above.

6.4 Stationary serially dependent marks

In our empirical analysis in Chapter 5, we observed that the time series of marks $\{\mathbf{X}_m\}$, where \mathbf{X}_m is the mark indexed by m at event time t_i , is serially dependent. This impacts application of the score test when (6.32) is used to compute the covariance of the score vector. In this section we give an informal discussion of the score test when marks are not i.i.d., but have serial dependence. Our treatment is informal because, as far as we can determine, there is no theoretical treatment of this case in the available literature for marked Hawkes processes. Indeed issues such as how to rigorously define the product measure $N_g(dt \times d\mathbf{x})$ and the likelihood for the marked point process will require further research. Additionally, there are subtleties about the way in which independence between the marks and the event times, predictability and so forth should be formulated.

Recall that for derivation of the log-likelihood $l_g(\nu)$ we assumed, as did Embrechts et al. (2011), that the marks be unpredictable as defined in Daley and Vere-Jones (2007, Defini-

tion 6.4.III(b)). While unpredictability holds for marks which are conditionally i.i.d given the past of the process, it does not hold for serially dependent marks. In our informal treatment to follow, we conceive of the joint distribution of the marks $\mathbf{X}_1, \dots, \mathbf{X}_n$ conditional on the event times t_1, \dots, t_n as being that of a stationary discrete or continuous time process with finite variance at least.

The score vector with respect to the boost parameters ψ , remains the same as above, but its covariance matrix is more complicated than given by (6.32), which is derived assuming the marks are i.i.d. As a result, falsely assuming the marks are i.i.d. in the computation of the score statistic, will lead to incorrect inferences using the asymptotic $\chi_{(r)}^2$ distribution for the score test statistic defined in (6.35). To see this, consider the situation where the covariance function for the marks process is $\mathbb{E}[G_k G_l^T] = \gamma(t_k - t_l; \phi)$, which is a function of the time between events. Then

$$\Omega_n = \mathbb{E} \left[\frac{1}{n} \sum_{l=1}^{n-1} A_{n,l}^2 \right] \Omega_G^* + \frac{1}{n} \sum_{m=1}^{n-1} \sum_{l=1}^{n-1} \mathbb{I}(l \neq m) \mathbb{E} [A_{n,k} A_{n,l}] \{ \gamma_G(t_l - t_m; \phi) + \gamma_G(t_m - t_l; \phi) \}. \quad (6.43)$$

A simpler alternative, is to assume that the marks form a stationary time series indexed by m with the lag h autocovariance of $G(\mathbf{X}_m, \phi)$, denoted by $\gamma_G(h; \phi)$. Then the variance of the score vector calculated under the i.i.d. assumption and given in (6.32) should be replaced by

$$\Omega_n = \mathbb{E} \left[\frac{1}{n} \sum_{l=1}^{n-1} A_{n,l}^2 \right] \Omega_G + \sum_{h=1}^{n-1} \mathbb{E} \left[\frac{1}{n} \sum_{l=1}^{n-h} A_{n,l} A_{n,l+h} \right] \{ \gamma_G(h; \phi)^T + \gamma_G(h; \phi) \}. \quad (6.44)$$

Obviously, if there is serial dependence in the marks and this is ignored in calculating the variance of the normalized score vector using (6.32), the resulting score statistic will not converge to the chi-squared distribution, but will converge to a constant times the chi-squared. If there is positive serial dependence, the constant will be greater than 1, in which case the (incorrect) test statistic will be over-sized under the null hypothesis, which will lead to falsely inflated power. This is illustrated in a simulation experiment below. If the series dependence is predominantly negative then the constant will be less than 1 and the size will be falsely low. Negative correlation can arise, but it is less common than positive dependence in our application. In either case, it is critical to correct for serial dependence by using a consistent estimate of Ω_n in (6.44).

There are several options available for estimating the covariances $\gamma_G(h; \phi)$. Firstly, one could estimate the parameters ϕ needed to specify the full joint conditional density $f(x_1, \dots, x_n | t_1, \dots, t_n; \phi)$ and use these in theoretical formulae for the needed covariances. For many of the series we have encountered, derivation of the theoretical autocovariances may not be a straightforward task and so this method may not be easily implemented. An alternative is to use non-parametric estimation of the required covariances. This is in principle possible without requiring a model and theoretical autocovariances to be derived. However, it will not always be possible to consistently estimate all the $n - 1$ terms required for estimation of (6.44), the main reason being the lack of data to obtain

consistent estimation of higher lag autocovariances of the $\{G(\mathbf{X}_m)\}$ process. There are two sets of serial dependence to take into account in forming (6.44), serial dependence associated with the $A_{n,m}$ series and serial dependence associated with the $G(\mathbf{X}_m)$ series. For the exponential decay case, the autocorrelation in the $A_{n,l}$ appears, at least empirically, to decay geometrically fast to zero as h in increase. In that case, the rate of decay of the $\gamma_G(l)$ is not as important for an accurate truncation method to give accurate calculation. Typically, when quantities such as (6.44) need to be calculated from a single realization of a time series, various tapering methods are employed. A simple form of tapering is hard truncation at a suitably large lag $K \ll n$. In the simulations this simple practical procedure is used with $K = \sqrt{n}$ to get

$$\Omega_n \approx \mathbb{E} \left[\frac{1}{n} \sum_{l=1}^{n-1} A_{n,l}^2 \right] \gamma_G(0) + \sum_{k=1}^K \mathbb{E} \left[\frac{1}{n} \sum_{l=1}^{n-k} A_{n,l} A_{n,l+k} \right] \{ \gamma_G(k)^T + \gamma_G(m)(k) \}.$$

We finally replace autocovariances with their sample estimates calculated using \hat{G}_m series as well as discarding the expectations over the Hawkes process event times to get

$$\hat{\Omega}_{K,n} \approx \left\{ \frac{1}{n} \sum_{l=1}^{n-1} \hat{A}_{n,l}^2 \right\} \hat{\gamma}_G(0) + \sum_{k=1}^K \left[\frac{1}{n} \sum_{l=1}^{n-k} \hat{A}_{n,l} \hat{A}_{n,l+k} \right] [\hat{\gamma}_G(k)^T + \hat{\gamma}_G(k)]. \quad (6.45)$$

We then calculate the adjusted score statistic, correcting for serial dependence, as

$$\hat{Q}_{H,n} = \hat{S}_n^T \hat{\Omega}_{H,n}^{-1} \hat{S}_n. \quad (6.46)$$

Simulations suggest that this simple approximate estimation of Ω_n is effective in obtaining a score tests statistic $\hat{Q}_{H,n}$, which has the correct size using the $\chi_{(r)}^2$ distribution of Proposition 1.

For continuous time version (6.43), non-parametric estimation of the $\gamma_G(t_l - t_m; \phi)$ is considerably more challenging. One could use variogram based estimates as discussed in Diggle et al. (2002), from which non-parametric estimates can be obtained. Typically these methods require very large amounts of data to be effective and the lack of consistency of the estimated covariances for high order time separation remains an issue, so some form of tapering would be required in this case also.

Another option which avoids all of the above issues, is to use an empirical version of $-\mathcal{I}_{\psi,\psi}(\nu^*)/(n\nu^2)$ from (6.22) to give

$$\hat{\Omega}_n = \frac{1}{n} \sum_{m=1}^n \lambda(t_i; \hat{\theta})^{-2} \left(\sum_{t_j < t_m} w(t_m - t_j; \hat{\alpha}) \hat{G}_m \right)^{\otimes 2}. \quad (6.47)$$

This method only requires likelihood estimation of the parameters of the unmarked process parameters θ , α , and parametric or non-parametric estimation of the expected values of $H(\mathbf{X})$, similarly to above. Simulations that follow, use (6.45) exclusively and show that this gives accurate adjustment for serial dependence in obtaining the correct χ^2 distribution for determining the size of the test. Future work could compare the use of (6.47).

6.5 Simulation methodology

The power of the score test gives the likelihood of rejecting the null hypothesis when the null is false, with an increase in power implying a decrease in the type II error. The power is calculated for a specific value under the alternative hypothesis. There are many choices for selecting the alternative hypothesis, with our choice being motivated by the ability to examine the power properties for a range of complex model choices and to allow for a simple presentation of the power properties.

The null hypothesis of the score test is $H_0 : \psi = 0$, where ψ is the parameter of the boost function. As we have demonstrated in Section 4.2.1, the boost function can take many forms. Within the simulation experiments to follow, we consider both a linear boost

$$h(\mathbf{X}; \psi) = 1 + \psi \mathbf{X} = 1 + \sum_{j=1} \psi_j X_j, \quad (6.48)$$

and a quadratic boost for a scalar mark X

$$h(X; \psi_1, \psi_2) = 1 + \psi_1 X + \psi_2 X^2. \quad (6.49)$$

When the dimension r of ψ is greater than 1, simulation of power against all combinations of the values ψ_1, \dots, ψ_r becomes computationally burdensome and graphical or tabular display of the results challenging. In the simulations presented here, we display the power of the test against alternatives on the 45° line $\psi_1 = \psi_2 = \psi$ with ψ increasing away from zero. While the choice of alternatives along a 45° line is acknowledged to be somewhat arbitrary, it treats both parameters equally in the examination of power. Of course other restrictions expressing ψ_1, \dots, ψ_2 in terms of a single ψ could be made and the methods we present applied.

To calculate the size and power properties of the score test, we simulate 1,000 replicates for each value of $\psi = [0, 0.01, 0.02, \dots, 1]$ of a given boost and mark distribution choice using the method described in Section 6.2.3. As ψ increases, the impact the mark has on the intensity is greater, and therefore we expect the power of the score test to approach 1. This will be indicative of good power properties of the score test. In the cases of a quadratic boost in (6.49), the increase in ψ_1 leads to an equivalent increase in ψ_2 . This facilitates the graphical representation of the power curves as we increase both boost function parameters. It is worth noting that the power curves will have some small degree of simulation variability, due to the fact that we estimated a new set of 1,000 replicates at each ψ . It is not possible to fix the replicates as we increase ψ , because when ψ changes, the intensity is boosted differently and so the number of events and their occurrence times will change. Hence it is not possible to reuse the same random numbers each time ψ changes and so simulation error cannot be controlled in this way. The variation due to simulation error is minimal and smoothers are not applied to the power curves.

The empirical tail probabilities are calculated by estimating the proportion of simulated score statistics that exceed the nominal 1%, 5%, 10% upper tail probabilities from the $\chi^2_{(r)}$ distribution with r degrees of freedom. The size of the score test is evaluated as

the probability of falsely rejecting the null hypothesis, and this should conform to the asymptotic $\chi^2(r)$ distribution. The 95% ranges expected around the nominal values are calculated as $p \pm 1.96\sqrt{p(1-p)}/1,000$, where p is the empirical exceedence probability. If we consider a single mark that is linearly boosted and set $\psi = 0$, the proportion of the 1,000 score statistic stimulants that exceed the $\alpha = 5\%$ upper tail probability for the $\chi^2_{(r=1)}$ distribution, should be $\hat{\alpha} \approx 5\%$, implying that the empirical type I error matches the asymptotic chi-squared distribution. If we observe a deviation, this would warrant further investigation and be suggestive of poor size properties of the score test.

Recall that for the Hawkes process, N_g is observed over the interval $t \in [0, T]$. The observed points of this process are $\{(t_i, \mathbf{X}_i), i = 1, \dots, n\}$ $N_g(t) \in \mathbb{N}$ with event times $0 = t_0 < t_1 < t_2 < \dots < t_n \leq T$ and a vector of d marks $\mathbf{X}_i \in \mathbb{X} \subset \mathbb{R}^d$. For the simulation experiments in Section 6.6 we consider the time interval $[0, T]$, which produces approximately n observations: $T = 130$ ($n \approx 500$), $T = 300$ ($n \approx 1,000$) and $T = 1,000$ ($n \approx 4,000$). Section 6.7 considers $T = 1,000$ ($n \approx 4,000$) and Section 6.8 considers a sample size of $n = 1,000$. If we consider one power curve, with $\psi = [0, 0.01, 0.02, \dots, 1]$, for each ψ we simulate 1,000 replicates of sample size n . In the studies where two parameters vary (Section 6.8), for example, $\psi = [0, 0.01, 0.02, \dots, 1]$ and serial dependence, $a_1 = [0.1, 0.2, \dots, 0.9]$, each combination of the two varying parameters represents a 1,000 replicates of sample size $n = 1,000$.

The steps involved in evaluating the power and size properties of the score test are outlined below.

Algorithm 12 (Size and power characteristics of the score test). *The steps that follow, evaluate the size and power characteristics of the score test.*

1. *Specify the sample size as either an end time T or count of events n .*
2. *Specify the dimension of the marks, form of the boost function, the marks distribution and associated parameters. In the event of serial dependence within the marks, specify the structure of serial dependence. In the event of joint dependence between the marks, specify the copula model and parameters.*
3. *Simulate the marked Hawkes process as per Algorithm 8 and with adjustments in Algorithm 11 for the serially dependent case, under the alternative hypothesis. 1,000 replicates are simulated for each ψ , starting with the null, $\psi = 0$. Under the null case, the boost function will be equal to 1.*
4. *For each replicate, estimate the decay parameter α under the null. Calculate the score statistic as outlined in Section 6.2.3.*
5. *For each $\psi = [0, 0.01, 0.02, \dots, 1]$, we will have 1,000 score statistics. For each ψ count the proportion of the score statistics that exceed the nominal 1%, 5%, 10% upper tail probabilities from the $\chi^2_{(r)}$ distribution with r degrees of freedom.*
6. *To create the power curves, plot the proportions against the ψ . The size properties are assessed at $\psi = 0$ under the null and the power properties are presented in the power curve.*

6.6 Simulation Experiments with i.i.d. Marks

In these simulation experiments we investigate if the score test, for a reasonably simple combinations of boost functions, of one or more independent marks conforms to those predicted by the asymptotic theory of Proposition 1. The intensity function is defined in terms of an exponential decay $w(t; \alpha)$ in (4.14) under various combinations of length of observation interval, boost function g and marks density $f(\mathbf{x}; \phi)$. To evaluate $G(\mathbf{X}; \phi)$, we can calculate the empirical or estimated theoretical moments.

The selection of appropriate parameters for the purposes of simulation studies of the marked Hawkes process, were chosen to represent reasonable parameter values that give a balance between the immigration intensity and branching ratio. For the initial studies in this Section 6.6) and in Section 6.7 we specify the intensity parameters as immigration rate $\eta = 0.80$, branching ratio $\vartheta = 0.80$ and decay rate $\alpha = 2.10$, that is $\theta = (0.80, 0.80, 2.10)$. For later studies, we will investigate the performance of the score test under conditions more closely mimicking those seen in the LOB. For the simulation experiments with i.i.d. marks that follow, we consider $T = 130$ ($n \approx 500$), $T = 300$ ($n \approx 1,000$) and $T = 1,000$ ($n \approx 4,000$).

6.6.1 Size and power of the score test

Objective

To study the size and power of the score test, we consider different combinations of marks dimensions $d \in \{1, 2, 4\}$, boost function and marks density. This will provide an assessment of the accuracy of the asymptotic chi-squared distribution for determining size of the score test under the null hypothesis that $\psi = 0$, for a variety of data generating mechanisms (DGM).

DGM

We consider both a linear boost $1 + \sum_1^r \psi_j X_j$ where $r = d$ in (6.48), and quadratic boost $1 + \psi_1 X^2 + \psi_2 X^2$ where $d = 1, r = 2$ in (6.49). We consider different combinations of marks dimensions $d \in \{1, 2, 4\}$, which are combined multiplicatively, and estimated theoretical moments for the evaluation of $G(\mathbf{X}; \phi)$.

Concerning the existence of moments of the generalized Pareto distribution, for the score test against a linear boost function we use $\zeta = 0.25$ which guarantees that $\mathbb{E}[X^r] < \infty$ for $r = 1/\zeta < 4$ but the fourth moment is not finite. For a linear boost, the existence of the second moment of the marks distribution is required for the score test statistic to be correctly defined and for Proposition 1 to hold. For quadratically boosted generalized Pareto distributed marks, the fourth moment is required to exist for Proposition 1 to be valid. For this first study we use $\zeta = 0.24$, which is slightly smaller than the required fourth moment to exist.

The marks distributions considered and associated simulation parameters are:

- Poisson distribution, denoted $\text{Pois}(\mu = 1.00)$ with μ average counts;

- Exponential distribution with density, $f(x; \lambda)$ and denoted $\text{Exp}(\lambda = 1.00)$;
- Generalized Pareto distribution with location parameter 0, $f(x; \zeta, \delta)$ with shape parameter ζ and scale parameter δ , denoted $\text{GPD}(\zeta = 0.25, \delta = 1.00)$ for a linear boost and $\text{GPD}(\zeta = 0.24, \delta = 1.00)$ for a quadratic boost.

We also consider a pair of linearly boosted dependent marks, $X_1 \sim \text{GPD}(\zeta = 0.24, \delta = 1.00)$ and $X_2 \sim \text{GPD}(\zeta = 0.24, \delta = 1.00)$, with dependence between them modelled using a Gaussian copula with Spearman's correlation, $\rho_s = 0.8$. The final simulation study considers four independent, linearly boosted marks, with $X_j \sim \text{GPD}(\zeta_j = 0.24, \delta_j = 1.00)$ for $j \in \{1, \dots, 4\}$. All three sample sizes are considered in this study, $T = 130$, $T = 300$ and $T = 1,000$.

Results

The combinations selected are shown in Figure 6.1, along with the empirical tail probabilities obtained using the proportion of simulated score statistics exceeding the nominal 1%, 5%, 10% upper tail probabilities from the $\chi^2_{(r)}$ distribution with r degrees of freedom. The degrees of freedom depend on the combination of options selected, $r \in \{1, 2, 4\}$ and are listed in Figure 6.1. Almost all simulated probabilities are within the 95% ranges expected around the nominal values. There is a slight bias above the nominal values, but no obvious pattern to the individual variations with combination of sample size, represented by $\log(T)$ on the horizontal axis and DGM.

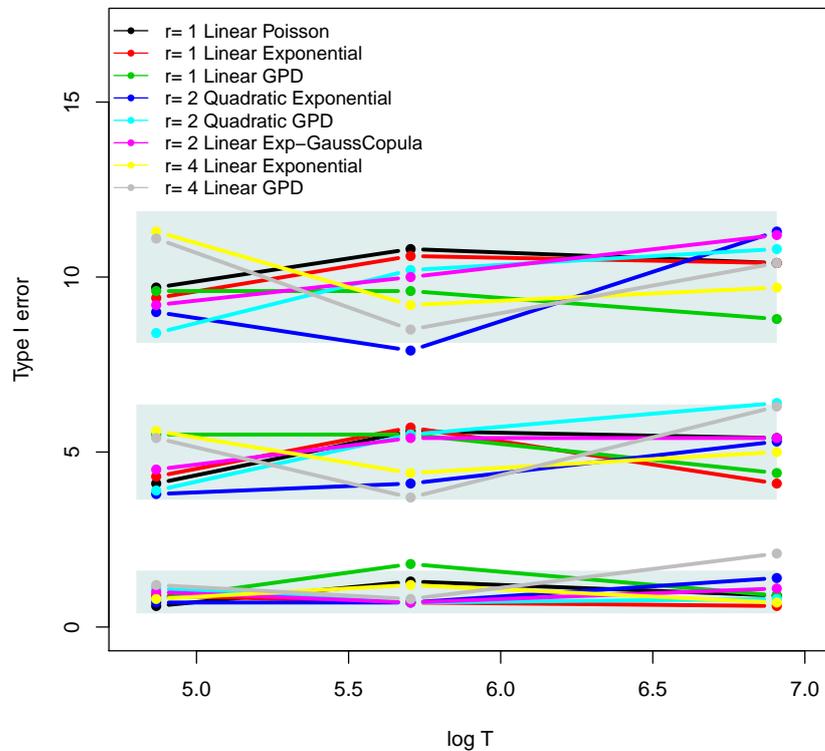


Figure 6.1: Simulated upper tail probabilities of the score test under a range of data generating mechanisms. The light blue bands reflect 95% ranges around the nominal chi-squared upper 1%, 5%, 10% type I errors.

Figure 6.2 shows the quantile comparison between simulated sampling distributions and the theoretical $\chi^2_{(r)}$ distribution for the smallest sample size considered, $T = 130$. We consider two distributions, the exponential distribution and the generalized Pareto distribution, with combinations of linear and quadratic boost and marks dimension $d \in \{1, 4\}$. The agreement between the empirical and theoretical distributions demonstrates that the theoretical chi-squared distribution is a sufficiently accurate approximation for practical use.

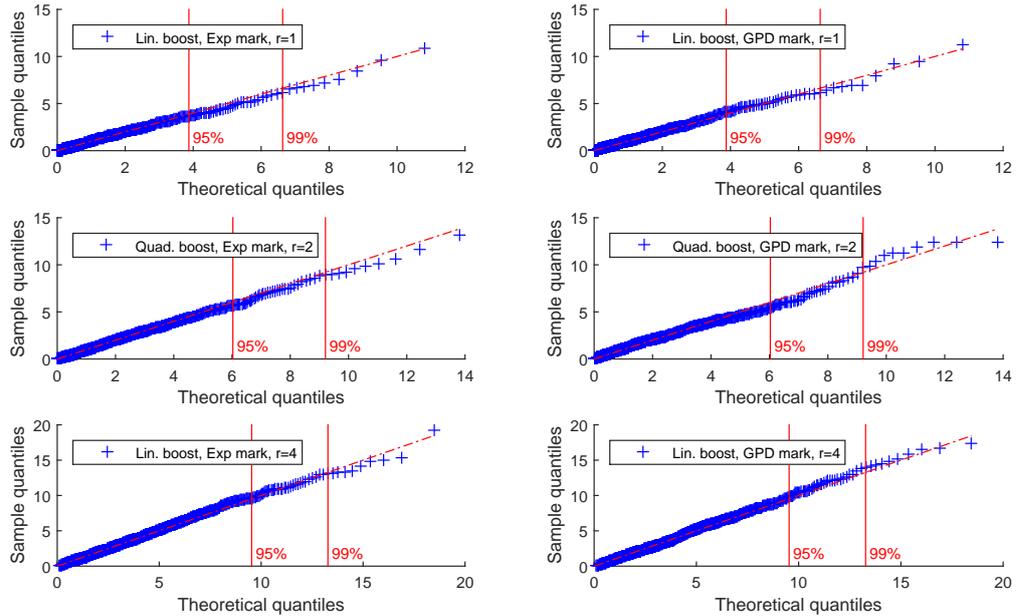


Figure 6.2: Comparison of sampling distribution of the score statistic against the asymptotic chi-squared distribution for several cases, which have varying degrees of freedom, $r \in \{1, 2, 4\}$: $X_i \sim \text{Exp}(\lambda = 1.00)$, with linear and quadratic boost; $X_i \sim \text{GPD}(\zeta = 0.25, \delta = 1.00)$, with linear boost; and $X_i \sim \text{GPD}(\zeta = 0.24, \delta = 1.00)$, with quadratic boost. We consider a small sample size of $T = 130$.

A selection of power curves for two different sample sizes corresponding to Figure 6.1, are presented in Figure 6.3. Whilst there is some degradation in power for the smaller sample size, the power properties for the score test perform well for most cases. The light tailed marks distributions present very similar power properties, whereas the marks with the heavy tailed generalized Pareto distribution have stronger power properties. We see a reduction in power performance for the quadratic boost function, compared with the linear boost. Marks that exhibit joint dependence and are modelled via a copula model, show stronger power properties for the score test.

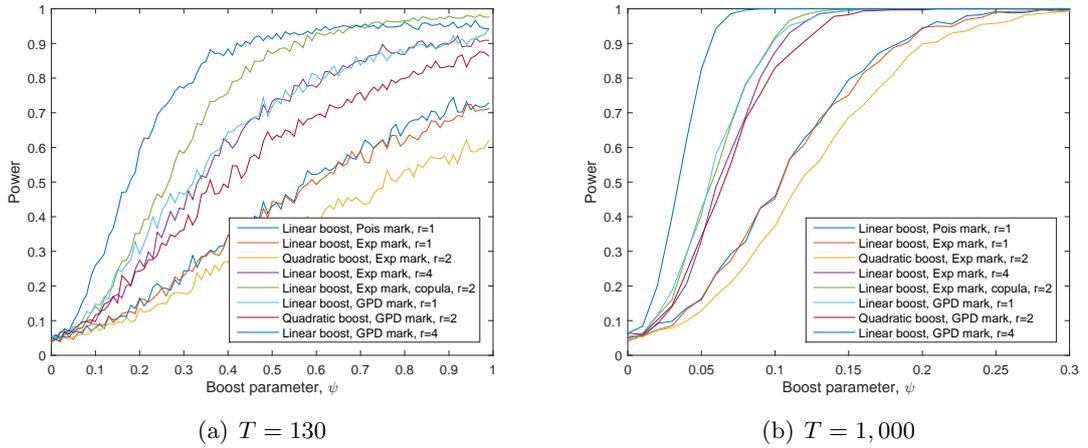


Figure 6.3: Power curves for two sample sizes of $T = 130$ and $T = 1,000$ for a selection of cases: $X_i \sim \text{Pois}(\mu = 1.00)$, with linear boost; $X_i \sim \text{Exp}(\lambda = 1.00)$, with linear boost for i.i.d. marks and jointly dependent marks (Gaussian copula model, $\rho_s = 0.8$) and quadratic boost; $X_i \sim \text{GPD}(\zeta = 0.25, \delta = 1.00)$, with linear boost; and $X_i \sim \text{GPD}(\zeta = 0.24, \delta = 1.00)$, with quadratic boost.

To explore the case of linear versus quadratic boost function, we present a single replicate each for the exponential distributed marks, with a linear and quadratic boost function and the generalized Pareto distributed marks, with a linear and quadratic boost function, where $\psi_1 = \psi_2 = 0.5$ (Figure 6.4). For both quadratic boost functions, the impact on the boost on the intensity function will be less than in the case of a linear boost. This is reflected in the shift down in the power properties in Figure 6.3. The case studies that follow, will consider these power curves and compare the power performance across a range of data generating mechanisms in more detail.

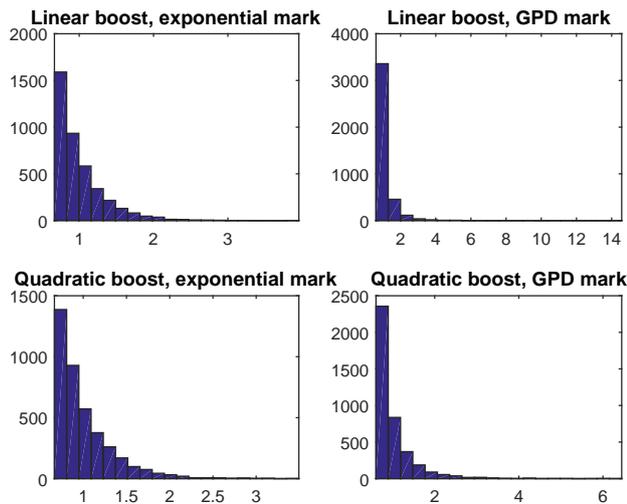


Figure 6.4: Histograms of linear and quadratic boost functions for a single replicate of sample size $T = 1,000$ for: $X_i \sim \text{Exp}(\lambda = 1)$, with linear and quadratic boosts; $X_i \sim \text{GPD}(\zeta = 0.25, \delta = 1.00)$, with linear boost; and $X_i \sim \text{GPD}(\zeta = 0.24, \delta = 1.00)$, with quadratic boost, where $\psi_1 = \psi_2 = 0.5$.

Conclusion

Overall we conclude that the score test statistic, over a range of situations and samples sizes, is achieving empirical type I errors that match (within simulation variation) those obtained from the asymptotic chi-squared distribution of Proposition 1. This provides some assurance that the chi-squared distribution can be used for obtaining sufficiently accurate null hypothesis quantiles for practical applications of the proposed score test of impact of marks. Whilst we do see an improvement in the power properties of the score test as the sample size increases, the score test continues to perform well for smaller sample sizes.

6.6.2 Comparison of empirical and estimated theoretical moments of the marks distribution

Objective

We explore whether or not using empirical moments or estimated theoretical moments for the marks distribution has an effect on the performance of the score statistic. Recall from Section 6.2.3, we denote \hat{G}_m for $G(\mathbf{X}_m, \hat{\phi}_T)$ when $G(\mathbf{X}_m, \hat{\phi}_T)$ is evaluated at the quasi-likelihood estimates of the parameters. For implementing the score statistic, use of empirical moments of the marks in defining \hat{G}_m , and in the estimated covariance matrix, is simple in practice and avoids the need to formulate and estimate a parametric model for the marks distribution. With this method, the assumption of the existence and well defined moments is required. Additionally, we also investigate the impact on the performance of the score test using ‘estimated parameters’ for the intensity process and marks distribution, and using the ‘true parameters’ for the intensity process and marks distribution. The latter method is not implementable in practice, however it useful for benchmarking performance.

DGM

We consider a linear boost (6.48), with an exponential distributed mark, $X \sim \text{Exp}(\lambda = 1.00)$. Recall that $\nu = (\theta, \phi, \psi)$ is the collection of all parameters for the marked process with intensity function in (4.14).

The six situations we consider correspond to combinations of the two groups presented below.

1. Two options for the intensity function $\lambda(t; \nu)$, calculated under the null are:
 - true parameters ν ;
 - estimated parameters $\hat{\nu}$ fitted under the null, treated as a compound test.
2. Three options for the estimation of moments used in calculating $G(\mathbf{X}, \phi)$ and estimating the covariance matrix Ω_n :
 - empirical moments;
 - theoretical moments evaluated using the true values of ϕ ;

- theoretical moments evaluated using the maximum likelihood estimates $\hat{\phi}$.

We consider sample sizes of $T = 130$ and $T = 1,000$ within this study. We proceed with the simulation method described in Section 6.5, however for this study it is possible to apply the six methods of estimating the score statistic to identical 1,000 replications. This allows for a direct comparison and control of simulation variability.

Results

Figure 6.5 compares the simulation estimates for the power curves for the six combinations described above, with linearly boosted, exponentially distributed marks with parameter $\lambda = 1.00$. This DGM matches the second choice in Figure 6.1. Two sample sizes $T = 130$ and $T = 1,000$ are compared. For the larger sample size the power curves were almost identical, with no discernible difference between them.

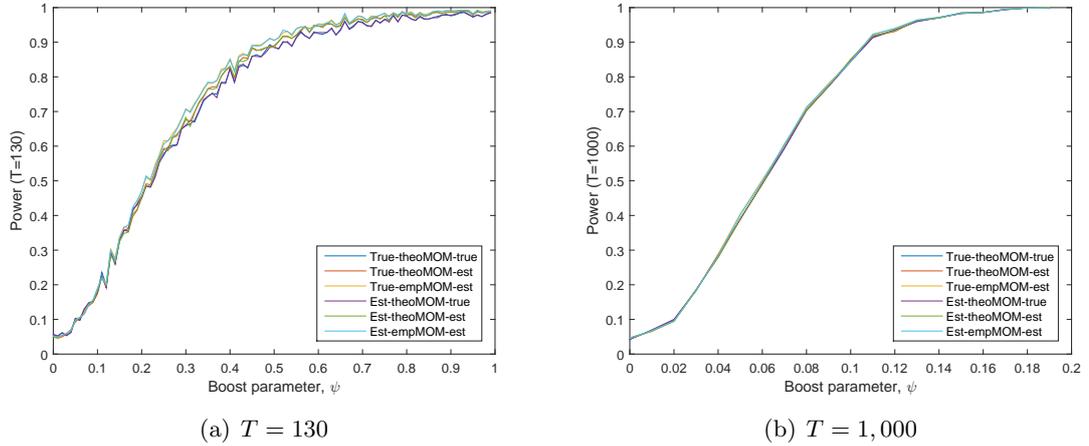


Figure 6.5: Power of the score test statistic for a linear boost and marks with an exponential distribution $X \sim \text{Exp}(\lambda = 1.00)$. This compares the use of theoretical moments with empirical moments in the score statistic, as well as true parameters and estimated parameters in the intensity function and the marks distribution.

Conclusion

The results suggest, at least in the case investigated, that the use of empirical moments compared to estimated theoretical moments does not degrade the size or power of the score statistic, and this also holds for smaller sample sizes. Knowing the true parameter values does not appear to improve size and power characteristics, compared with estimated parameters. The advantage of estimating the moments empirically allows one to avoid estimating a parametric model for marks distribution, assuming that the moments exist. This provides compelling evidence to proceed with estimating moments empirically for the evaluation of the score statistic. This is particularly useful in the process of variable or marks screening when many marks may be under consideration.

6.6.3 Robustness against moments not existing for generalized Pareto distributed marks

Objective

The purpose of this study is to investigate the extent to which the sampling distribution of the score statistic breaks down when the theoretical moments of the marks distribution, required to estimate the covariance matrix \hat{G}_m , used to calculate the score statistic, do not exist. In this context, we can consider the influence on the power of the test as a function of model misspecification, with respect to existence of theoretical moments.

DGM

We consider a linear boost (6.48), with a generalized Pareto distribution. For the generalized Pareto distribution, moments up to but not including the r th exist when $r = 1/\zeta$, where ζ is the shape parameter. Hence for $\zeta = 0.5$, the second moment does not exist, but moments of order lower than 2 do exist. In this case the score statistic for a linearly boosted generalized Pareto distributed marks is not properly defined since the variance of the score vector does not exist. Likewise, when $\zeta = 0.25$ moments up to but not including the 4th exist, in which case a the score statistic for a quadratically boosted mark is not properly defined. With that in mind, we consider the following combinations:

- Linear boost function with marks $X \sim \text{GPD}(\zeta = 0.25, \delta = 1.00)$;
- Linear boost function with marks $X \sim \text{GPD}(\zeta = 0.49, \delta = 1.00)$;
- Linear boost function with marks $X \sim \text{GPD}(\zeta = 0.50, \delta = 1.00)$;
- Quadratic boost function with marks $X \sim \text{GPD}(\zeta = 0.10, \delta = 1.00)$;
- Quadratic boost function with marks $X \sim \text{GPD}(\zeta = 0.24, \delta = 1.00)$;
- Quadratic boost function with marks $X \sim \text{GPD}(\zeta = 0.25, \delta = 1.00)$.

For some combinations that are considered, the theoretical moments are not defined. In addition, Section 6.6.2 demonstrated that empirical moments are suitable in the evaluation of the score statistic, thus we will proceed with using empirical moments. We consider the sample size of $T = 1,000$ for this study.

Results

Figure 6.6 compares the actual sampling distribution of the score statistic with the relevant $\chi_{(r)}^2$ distribution for values of ζ , which is the extremal tail index of the linearly and quadratically boosted, generalized Pareto distributed marks. Considered are various combination for which the score statistic is well defined and others for which it is not well defined. For the linear boost: when $\zeta = 0.25$, moments below the fourth exist; when $\zeta = 0.49$, the second moment just exists; and when $\zeta = 0.50$, the second moment does not exist. The case where $\zeta = 0.49$ conforms very well to the claimed asymptotic chi-squared

distribution, while for $\zeta = 0.50$ there is some deviation in the upper tails, which is not so severe as to make the upper 5% value incorrect, but may impact the upper 1% quantile. For the quadratically boosted case: the values $\zeta = 0.10$, for which moments of order below 10 exist; $\zeta = 0.24$, for which the 4th moment just exists; and $\zeta = 0.25$, for which the 4th moment does not exist. Again, the results are mixed but with no obvious break down in the approximation by the asymptotic chi-squared distribution.

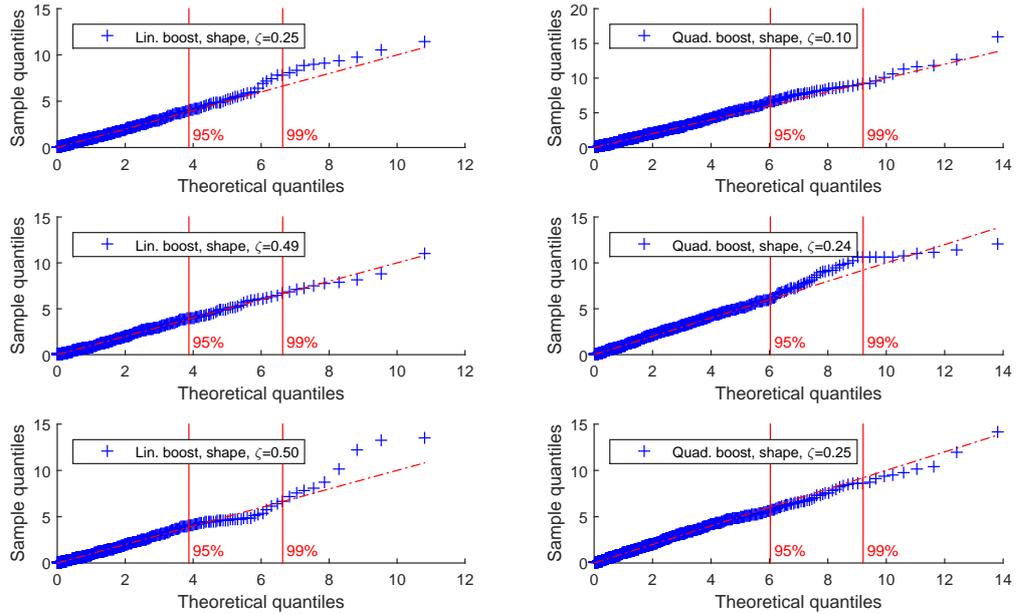


Figure 6.6: Robustness of $\chi^2_{(r)}$ distribution for the sampling distribution, comparing when moments of the GPD exist, marginally exist and do not exist. The cases considered are, linear boost with $r=1$: $X_i \sim \text{GPD}(\zeta = 0.25, \delta = 1.00)$; $X_i \sim \text{GPD}(\zeta = 0.49, \delta = 1.00)$; and $X_i \sim \text{GPD}(\zeta = 0.50, \delta = 1.00)$, and quadratic boost with $r=2$: $X_i \sim \text{GPD}(\zeta = 0.10, \delta = 1.00)$; $X_i \sim \text{GPD}(\zeta = 0.24, \delta = 1.00)$; and $X_i \sim \text{GPD}(\zeta = 0.25, \delta = 1.00)$. The sample size is $T = 1,000$.

Figure 6.7 shows the power curves for all six cases that were previously presented in Figure 6.6. First focusing on the linear boost with $\zeta = 0.49$ and $\zeta = 0.50$, there is no suggestion of breakdown in the power performance at the margin where the score statistic is theoretically not well-defined. Similarly for the quadratic boost case when $\zeta = 0.24$ and $\zeta = 0.25$, the two power curves are almost identical, suggesting a smooth transition in performance at the threshold for which moments do not exist.

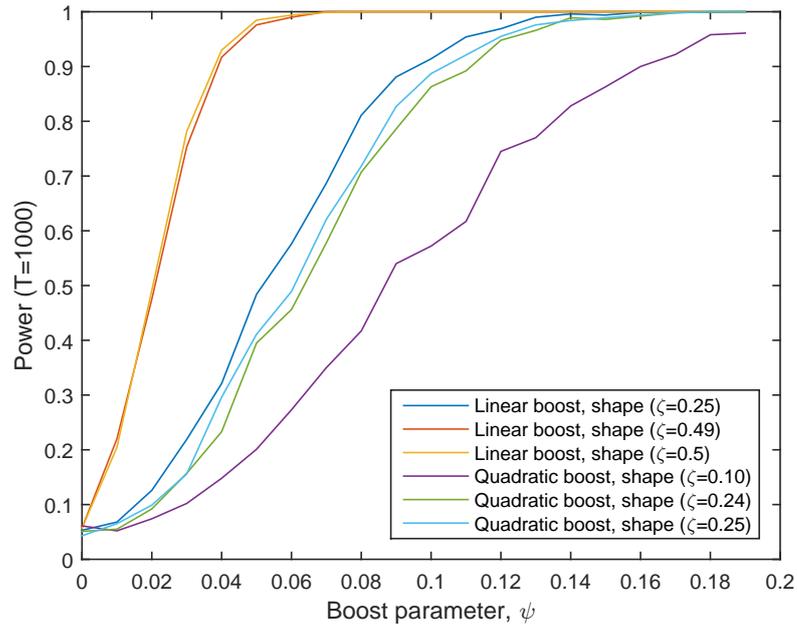


Figure 6.7: Power of the score test for various combinations of GPD distributed marks and boost functions, illustrating the impact of shape parameter varying close to values, for which the moments required to define the score statistic, don't exist or only marginally exist. The sample size is $T = 1,000$.

Conclusion

There is no obvious break down in the power performance of the score statistic when using empirical moments in place of estimated theoretical moments that are not well defined. There is no apparent degradation in the size or power of the score statistic, further motivating the use of empirical moments over theoretical moments in the evaluation of the score statistic.

6.7 Simulation Experiments with Dependence

A key feature within the LOB data sets, which was observed in Chapter 5, was the frequent violation of the i.i.d. assumption with many marks demonstrating serial dependence. In addition, select combinations of marks demonstrated joint dependence. In Section 4.2.2 we presented a simulation algorithm (Algorithm 8) to simulate a series of random events according to the specification of a given Hawkes process. Within this framework we can simulate and model joint dependence via a copula model of choice. To emulate the serial dependence features of the observed marks data, we introduced an extension to the simulation method to incorporate serial dependence structures for the marks in Algorithm 11.

For the simulation experiments that follow, we set the intensity parameters to $\theta = (0.80, 0.80, 2.10)$. We consider a linear boost in (6.48) and a sample size $T = 1,000$ ($n \approx 4000$).

6.7.1 Impact of ignored serial dependence in marks

Objective

In this case study we will apply the proposed extensions for simulating marks with serial dependence and the adjustment to the score test as proposed in Section 6.2.3. The aim is to explore the impact on the size of the score test when the serial dependence of the mark is ignored.

DGM

The marks in this simulation study are distributed with a conditional generalized Pareto distribution, $X_i \sim \text{GPD}(\zeta = 0.10, \delta_i)$. The conditional generalized Pareto distribution scale parameter is defined as $\delta_i = 0.9\delta_{i-1} + \epsilon_i$, where ϵ_i is Gaussian white noise, see Section 5.3.3. For the evaluation of $G(\mathbf{X}; \phi)$ moments are estimated empirically.

Results

Figure 6.8 shows the sampling distribution of the corrected score statistic in (6.46) with the uncorrected score statistic in (6.35). Given the serial dependence is positive and reasonably persistent, the uncorrected score statistic is inflated because the variance obtained, ignoring the extra terms due to serial dependence, is too small. This is clear from Figure 6.8(a) in which the uncorrected statistic is substantially stochastically larger than the $\chi^2_{(1)}$ distribution, while the corrected score statistic conforms to the claimed asymptotic distribution very well. This results in the uncorrected score statistic realizing an actual size which is substantially larger than that from the $\chi^2_{(1)}$ distribution, and hence the power curve is substantially inflated (Figure 6.8(b)).

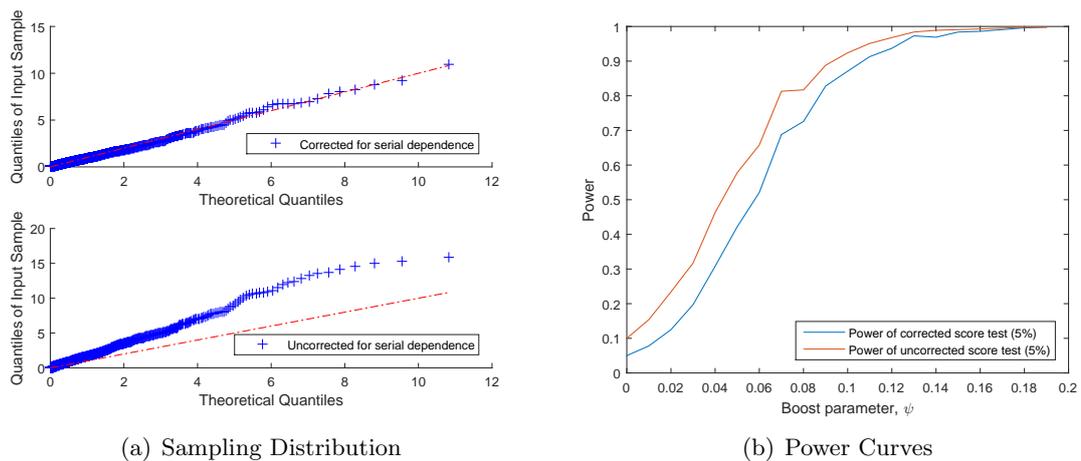


Figure 6.8: Impact on the sampling distribution and power curves of the score test statistic, by adjusting and not adjusting for serial dependence, with a sample size $T = 1,000$. The marks are conditionally GPD, $X_i \sim \text{GPD}(\zeta = 0.10, \delta_i)$, with a linear boost and empirically estimated moments.

Conclusion

Incorrectly assuming the mark random vectors are independent and identically distributed, when in fact the mark being test exhibits serial dependence, will result in an unadjusted score test statistic that is inflated. This may lead to incorrectly identifying a mark as significant. The adjustment made to the score test for serial dependence conforms to the $\chi_{(1)}^2$ distribution well.

6.7.2 Impact of ignored joint dependence in marks

Objective

We assess the impact of joint dependence between the marks on the score test, when it is both accounted for within the evaluation of the score statistic and when it is ignored.

DGM

We consider a pair of linearly boosted (6.48), jointly dependent i.i.d. marks $X_1 \sim \text{GPD}(\zeta = 0.10, \delta = 1.00)$ and $X_2 \sim \text{GPD}(\zeta = 0.10, \delta = 1.00)$. The joint dependence is modelled using a Gaussian copula, with three different models for dependence, $\rho_s \in \{0, 0.4, 0.8\}$.

For implementing the score statistic, we use both empirical moments of the marks and the estimated theoretical moments in defining $G(\mathbf{X}_m, \hat{\phi}_T)$ and in the estimated covariance matrix. We assume that the marginal marks distributions F_1 and F_2 have finite moments up to at least fourth order, such that the marginal central moment $\mu_{4j} < \infty$ for $j \in \{1, 2\}$. The resulting asymptotic variance under the null of the linearly boosted covariance Σ_G^{-1} is given in general by,

$$\Sigma_G = \begin{bmatrix} \mu_{2,1} & \mathbb{E}[(X_1 - \mathbb{E}[X_1])(X_2 - \mathbb{E}[X_2])] \\ \mathbb{E}[(X_1 - \mathbb{E}[X_1])(X_2 - \mathbb{E}[X_2])] & \mu_{2,2} \end{bmatrix}, \quad (6.50)$$

where $\mu_{2,1} = \mathbb{E}[(X_1 - \mathbb{E}[X_1])^2]$ and $\mu_{2,2} = \mathbb{E}[(X_2 - \mathbb{E}[X_2])^2]$.

In a similar manner to Section 4.2.1, the off diagonal terms are evaluated using

$$\mathbb{E}[(X_1 - \mu_{1,1})(X_2 - \mu_{1,2})] = \rho\sqrt{\mu_{2,1}\mu_{2,2}},$$

where ρ is the pairwise linear correlation between X_1 and X_2 .

For the Gaussian copula, we can obtain the pairwise linear correlations by transforming the Spearman's rank correlation, ρ_s in (4.11), obtaining

$$\mathbb{E}[(X_1 - \mu_{1,1})(X_2 - \mu_{1,2})] = \sqrt{\mu_{2,1}\mu_{2,2}} \ 2 \sin\left(\frac{\pi}{6}\rho_S(X_1, X_2)\right)$$

We imposed block diagonalization of the covariance matrix when demonstrating the case of ignored joint dependence in the score statistic.

Results

Figure 6.9 shows the power of the score test under three scenarios, whereby the strength of dependence between the marks is chosen to be Spearman's rank correlation = $(0, 0.4, 0.8)$ for each respective figure. Whilst there is no obvious breakdown in the power properties for the cases where we don't account for cross correlation terms, there is some suggestion of the power curve shifting up for the block diagonal scenarios, as the Spearman's rank correlation increases.

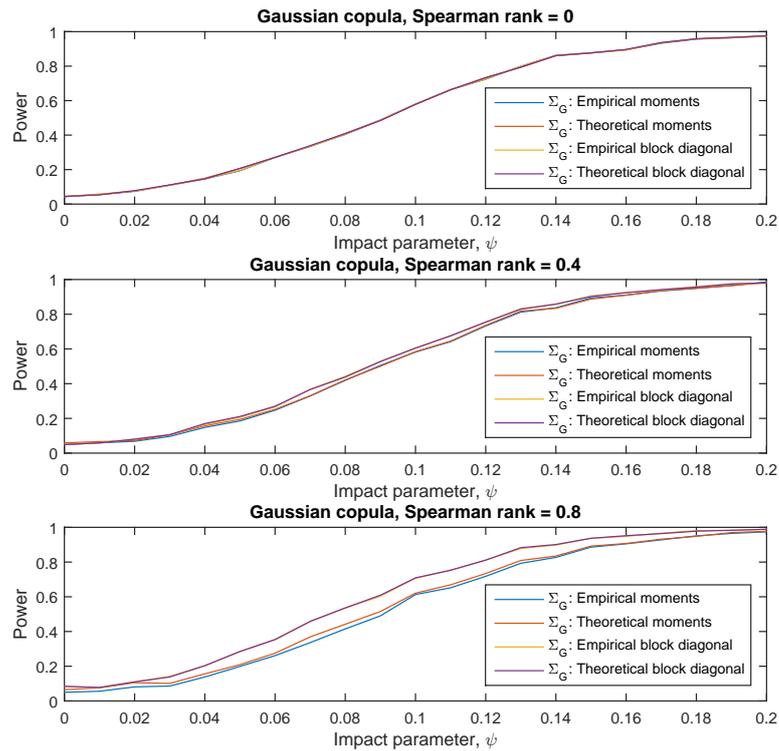


Figure 6.9: Power of the score test for a linear boost, marks with a GPD and joint dependence modelled by a Gaussian copula, with Spearman's rank correlation, $\rho_s = \{0, 0.4, 0.8\}$. Comparing the use of theoretical moments and empirical moments, with calculation of cross terms and imposed block diagonalize of the covariance matrix in the score statistic. The sample size is $T = 1,000$.

To further investigate this point, we refer to Figure 6.10 and Table 6.1, which shows the proportion of simulated scores statistics exceeding the nominal 5% upper tail probabilities from the $\chi^2_{(r)}$ distribution with $r = 2$ degrees of freedom. What is apparent is the inflation of the incorrect score statistic, that is, when the covariance Σ_G is block diagonal, becomes larger than the $\chi^2_{(2)}$ distribution as the Spearman's rank correlation increases.

Table 6.1: Upper tail probabilities for the nominal chi-squared upper 5% type I errors.

Σ_G	Spearman rank = 0	Spearman rank = 0.2	Spearman rank = 0.4	Spearman rank = 0.6	Spearman rank = 0.8
Empirical moments	0.043	0.052	0.049	0.041	0.050
Theoretical moments	0.045	0.054	0.059	0.053	0.065
Emp. block diagonal	0.044	0.051	0.050	0.067	0.084
Theo. block diagonal	0.044	0.051	0.049	0.065	0.084

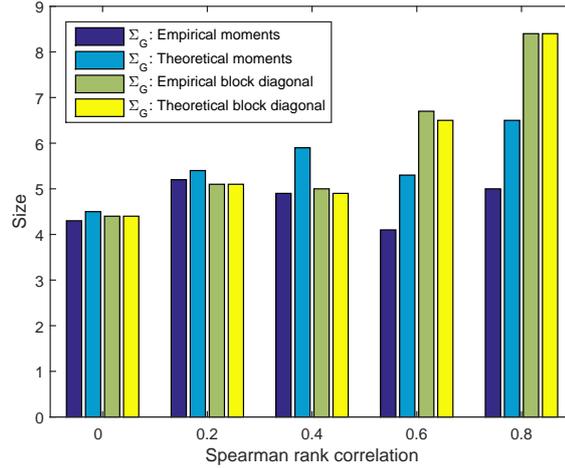


Figure 6.10: Simulated upper tail probabilities of the score test using theoretical and empirical moments, with calculation of cross terms and imposed block diagonalize of the covariance matrix. We consider the score test for a linear boost, marks with a GPD and joint dependence modelled by a Gaussian copula, with Spearman’s rank correlation $\rho_s = \{0, 0.2, 0.6, 0.8\}$. The sample size is $T = 1,000$.

Conclusion

The correct score statistic conforms to the asymptotic distribution, with largely consistent results for both the empirical and theoretical moment estimation methods. Similarly to the serial correlation case discussed previously, the results for the incorrect score statistic lead to an inflated power curve for positively correlated marks.

6.8 Simulation Experiments with Extensions

We have demonstrated that the statistic has good power properties with the use of empirical moments. This means that parametric models for the marks do not need to be specified. In addition, the score test is able capture the joint dependence between the marks and a suitable adjustment is made for the case of marks with serial dependence, all of which are features within the LOB data sets we observed in Chapter 5. To complete the analysis on the power properties of the score test, we turn our attention to assessing the size and power properties for simulation specifications that closely match the data structures observed in Chapter 5.

The key properties of the marks that were found in Chapter 5, and which we replicate for the purposes of simulations and testing the power properties of the score test with

these complex features are:

- Continuous approximations and discrete marginal distributional forms;
- Heavy tailed features;
- Serial dependence;
- Joint dependence.

It is worth noting that for cases of multiple dimensional marks, all three features, heavy tailedness, serial and joint dependence occur together.

For the simulation experiments that follow, the parameter specification is chosen to match closely what would be observe in the LOB data. Unless otherwise specified, we set the intensity parameters to, immigration rate $\eta = 0.0010$, branching ratio $\vartheta = 0.7000$ and decay rate $\alpha = 0.0100$, that is $\theta = (0.0010, 0.7000, 0.0100)$. These values are similar to those obtained from the real LOB data studied in Chapter 7, similar to what we used in Chapter 4 and the simulations presented here are explicitly relevant to the Chapter 7 applications.

For the marks parameters for simulation, we estimate the marginal distributions for the marks across 10 trading days. We use this information to specify the marginal parameter estimates and serial dependence features within the simulations. For the studies that follow, we consider a linear boost in (6.48) and moments estimated empirically. We consider simulation size of $n = 1,000$ ($T \approx 150,000$ milliseconds).

The majority of the studies in this section consider increasing serial dependence and impact on the score test. To simplify the descriptions throughout, *increasing serial dependence* (a_1) means that the parameter, a_1 in the autoregressive model in (5.24), and which is used to invoke serial dependence via a chosen parameter within the marginal distribution, is increasing.

6.8.1 Impact of increasing serial dependence with heavy tailed marks

Objective

The aim of this study is to investigate the impact on the size and power performance of the score test under the conditions of increasing serial dependence for a heavy tailed mark.

DGM

For the simulation experiments that follow, we set the intensity parameters to $\theta = (0.0010, 0.7000, 0.0100)$ and for the purposes of a comparative study, we also consider intensity parameters of $\theta = (0.0020, 0.7000, 0.0100)$. The marks have a conditional generalized Pareto distribution, $X \sim \text{GPD}(\zeta = 0.10, \delta_i)$, where the scale parameter is defined as, $\delta_i = a_1 \delta_{i-1} + \epsilon_i$, where ϵ_i is Gaussian white noise. We vary the coefficients term, $a_1 = [0.1, 0.2, \dots, 0.9]$. This study also includes the power curves for the i.i.d. case for comparison.

Results

Figure 6.11 shows the power of the score test for a generalized Pareto distributed mark with increasing serial dependence. Figure 6.11(a) demonstrates good power properties for all levels of serial dependence. Figure 6.11(b) presents the case of i.i.d. marks and two levels of serial dependence, $a_1 \in \{0.4, 0.9\}$, noting that the boost parameter ψ is shown for $\psi = [0, 0.01, \dots, 0.3]$. The power properties for serial dependence $a_1 = 0.4$, are stronger than the power properties of the score test in the case of i.i.d. marks. However, when we increase the serial dependence to levels that represent the real data, $a_1 = 0.9$, the increase in power is far greater than the proportional increase in the serial dependence parameter a_1 .

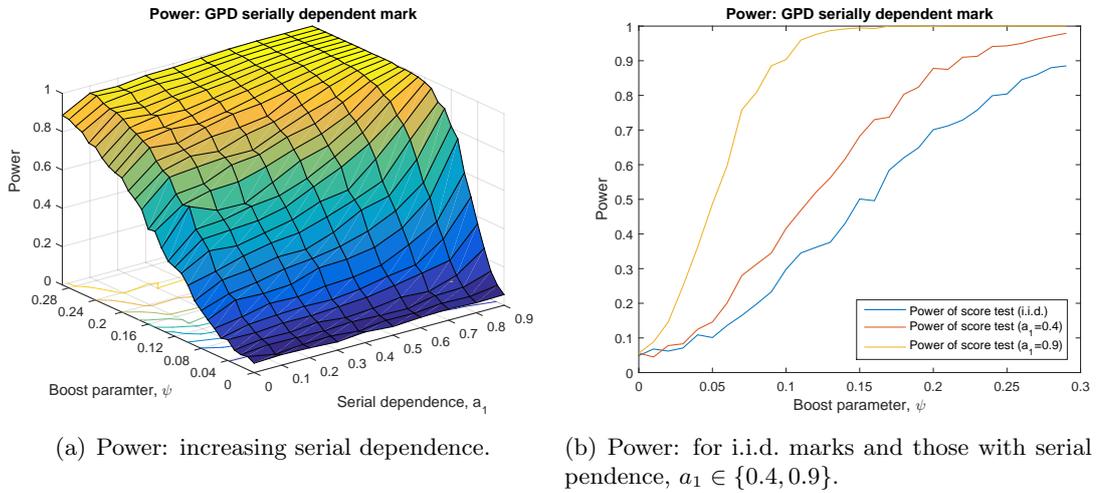


Figure 6.11: Power of the score test, with a linear boost, increasing serial dependence, for $X \sim \text{GPD}(\zeta = 0.10, \delta_i)$, with a sample size $n = 1,000$. For later comparison, note that the immigration intensity is $\eta = 0.0010$ in this study.

To explore this phenomenon further, we present the boost function for a single simulant for the i.i.d. case and for each level of the serial dependence, $a_1 \in \{0.4, 0.9\}$ (Figure 6.12). For the lower level of serial dependence, the histogram of the boost does not differ too much from the i.i.d. case, however there is a sizeable jump in the tails of the boost when the serial dependence increases to $a_1 = 0.9$. This result is expected, as the heavy tailedness of the generalized Pareto distribution, coupled with increased serial dependence leads to more sequences of extreme positive realizations of the mark random variable over time, resulting in enhanced and sustained impacts on the intensity through the boost of the marks. This increases the impact of the mark on the intensity process, thus leading to higher significance in the score test.

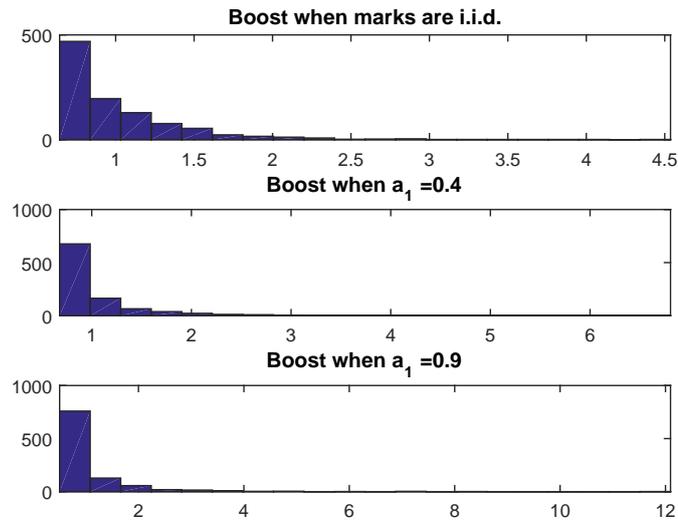
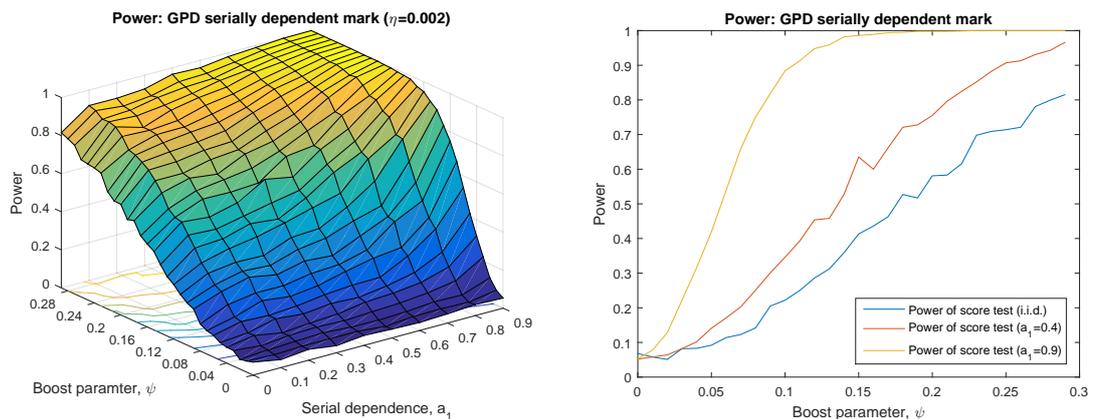


Figure 6.12: Histograms of the linear boost function for a single simulant, for i.i.d. marks and those with serial dependence $a_1 \in \{0.4, 0.9\}$, for $X \sim \text{GPD}(\zeta = 0.10, \delta_i)$, with a sample size $n = 1,000$.

We now consider the impact of doubling the immigration intensity from $\eta = 0.0010$ to $\eta = 0.0020$. Figure 6.13(a) presents good power properties for the score test when the immigration intensity is doubled. However, comparing Figure 6.13(b) to the previous Figure 6.11(b), and considering both the individual slices of the i.i.d. case and when the serial dependence is $a_1 \in \{0.4, 0.9\}$, there is a shift down in the power of the score test. The additional weight given to the immigration intensity reduces the relative influence that was driven purely by the marks on the intensity via the boost function.



(a) Power: increasing serial dependence.

(b) Power: for i.i.d. marks and those with serial dependence, $a_1 \in \{0.4, 0.9\}$.

Figure 6.13: Power of the score test, with a linear boost, increasing serial dependence, for $X \sim \text{GPD}(\zeta = 0.10, \delta_i)$, with a sample size $n = 1,000$. The immigration intensity is $\eta = 0.0020$.

Conclusion

Increasing serial dependence for a heavy tailed distributed mark results in an increasing shift up in the power curve for the score test. This is not a proportional change and is a result of increased clustering of events in the tails. When the immigration intensity is doubled, there is a noticeable shift down in the power curve of the score test. The power and size properties of the score test are reliable in the assessment of marks with heavy tails and increasing serial dependence.

6.8.2 A breakdown in power

Objective

In light of the findings of reduced power of the score test as we increase the immigration intensity, the next study inspects a single slice of boost function parameter $\psi = 0.5$ to investigate the relative proportion of the immigration intensity and branching coefficient that results in a break-down in the power of the score test. The aim is to establish some measure that will provide guidance of the break-down of power in the score test for combinations of immigration and branching coefficient parameters.

DGM

As we have observed in Figures 6.11 and 6.13, the power properties for the score test approach 1 when the boost parameter is $\psi = 0.3$, thus the parameter of the boost function of $\psi = 0.5$ ensures excellent power of the score at the previously specified levels of the immigration intensity.

For the simulation experiments that asses the power of the score test, we set the intensity parameters to, branching ratio $\vartheta = 0.7000$ and decay rate $\alpha = 0.0100$, however we varying the immigration rate such that $\eta = [0.0010, 0.0020, \dots, 0.0300]$.

When establishing a suitable measure of model stability and good power properties, we varying the immigration intensity by $\eta = [0.0020, 0.0040, \dots, 0.0300]$ and the branching coefficient by $\vartheta = [0.2000, 0.3000, \dots, 0.8000]$. For each study we consider the power of the test when the boost function is $\psi = 0.5$. The marks are conditional generalized Pareto distributed $X \sim \text{GPD}(\zeta = 0.10, \delta_i)$. For these simulation experiments we consider a sample size of $n = 1,000$.

The conditional generalized Pareto distribution scale parameter is defined as $\delta_i = a_1\delta_{i-1} + \epsilon_i$. We consider slices of the coefficients term $a_1 = [0.3, 0.6, 0.9]$ to investigate the relationship between an increasing immigration parameter, serial dependence and how this might contribute to a breakdown in power of the score test. The study also includes the power curves for the i.i.d. marks case for comparison.

Results

When the immigration intensity reaches $\eta = 0.0030$ (Figure 6.14) there is a break down in the power properties of the score test. This break down in power is consistent across all levels of serial dependence, however the stronger power properties of marks with high

serial dependence, ensures good power proprieties for higher levels of immigration intensity versus those with lower serial dependence.

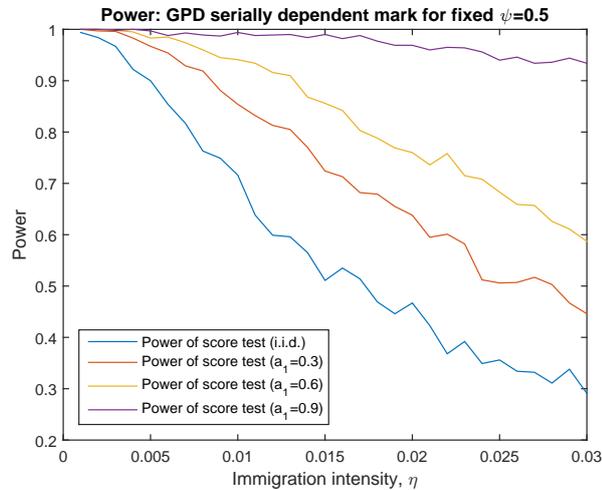


Figure 6.14: Power of the score test for a linear boost, with a boost parameter fixed at $\psi = 0.5$, increasing serial dependence, for $X \sim \text{GPD}(\zeta = 0.10, \delta_i)$, and an increasing immigration intensity. The sample size is $n = 1,000$.

The next step is to establish some measure that will provide guidance of the break-down of power in the score test for certain combinations of immigration and branching coefficient parameters. Initial guidance might suggest the use of the long run average intensity in (4.30) as a suitable measure. However, Figure 6.15 demonstrates the shape of this function for various combinations of immigration and branching coefficient, does not match the power degradation in Figure 6.16(a). In addition, the long run intensity is a measure of the long run counts of events per inter-arrival time. For a poorly calibrated marked Hawkes process, the estimated immigration and branching parameters, still result in a long run average intensity estimation that matches the counts per inter-arrival time. Thus the long run intensity does not provide useful information about overall model stability.

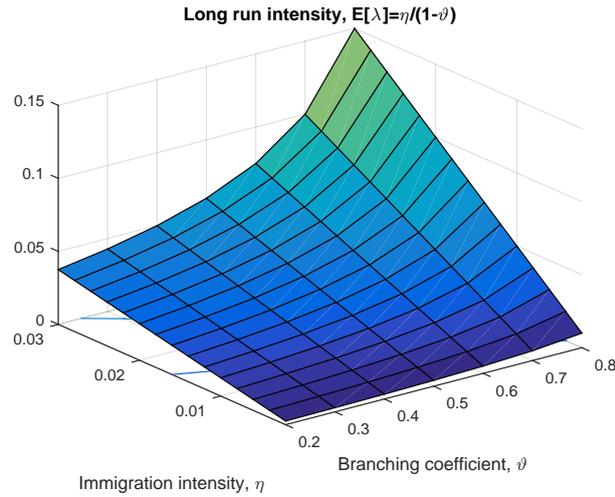


Figure 6.15: Long run intensity $\mathbb{E}[\lambda]$ for various combinations of immigration intensity η and branching coefficient ϑ .

Despite the observed degradation in the power of the score test shown in Figure 6.14 as the immigration intensity increases, it is not enough to consider this in isolation of the branching coefficient. In the case of a poorly calibrated model, the immigration intensity may not vary significantly from the well estimated models, but the branching coefficient approaches zero, implying all arrivals are attributed to the baseline intensity. And finally, using the branching coefficient in isolation of the immigration intensity as a measure of model stability, will not take into account the break-down in power that we have observed in Figure 6.14 when immigration is increased, whilst the branching coefficient remains high, $\vartheta = 0.7000$.

We propose a simple measure that considers both the interaction of the immigration intensity and branching coefficient. This measure adds further weight to the branching coefficient in the denominator to adjust for the cases where the branching coefficient is quite low, resulting in decreased power and poor calibration of the model. We will refer to this measure as a measure of model stability, IB (Immigration Branching) and defined as

$$IB = \frac{\eta}{(\eta + \vartheta)\vartheta^2}, \quad (6.51)$$

where $IB < 0.4$ implies a well calibrated model with good power properties.

The simulation study that follows, investigates the robustness of this measure across different levels of immigration intensity $\eta = [0.0020, 0.0040, \dots, 0.0300]$ and branching coefficient $\vartheta = [0.2000, 0.3000, \dots, 0.8000]$. For each study we consider the power of the test when the boost function is $\psi = 0.5$.

Figure 6.16(a) presents the power properties of the score test using marks that are conditional generalized Pareto distributed $X \sim \text{GPD}(\zeta = 0.10, \delta_i)$, with a fixed boost parameter of $\psi = 0.5$ and with combinations of immigration intensity and branching coefficient. When the branching coefficient is sufficiently high, good power properties are retained for an increasing immigration intensity. Figure 6.16(b) displays the functional

form of the measure of model stability IB . Across the 105 combinations of immigration and branching coefficient, $IB \geq 0.4$ identifies the correct combinations that result in poor power properties, with an accuracy of 96.01%.

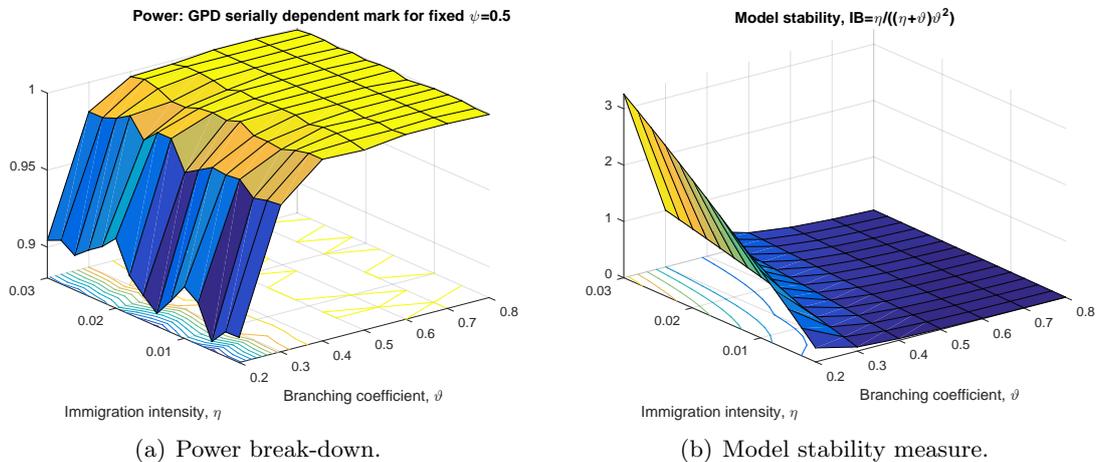


Figure 6.16: The drop in power of the score test for different combinations of immigration intensity and branching coefficient, for $\psi = 0.5$, across 1,000 simulations each. The measure of model stability IB for different combinations of immigration intensity and branching coefficient.

Conclusion

A breakdown in the power of the score test is observed when $\eta = 0.0030$, for a fixed combination of specified Hawkes process parameters and for marks with a generalized Pareto distribution. The breakdown is consistent across all levels of serial dependence, however higher serial dependence ensures better power properties as the immigration intensity increases. A measure constructed from a combination of immigration and branching ratio, IB , under the null distribution is proposed to assess the calibration and good power properties of a Hawkes process. When $IB \geq 0.4$, an accuracy of 96.01% is achieved in identifying models with poor power properties. An extension to this study will consider how this measure performs for different power properties of various marks distributions and boost functions.

6.8.3 Extended robustness tests of moments with increasing serial dependence

Objective

In Section 6.6.3 we studied the extent to which the sampling distribution of the score statistic breaks down when theoretical moments of the marks with a generalized Pareto distribution do not exist and with no obvious breakdowns observed. This is extended by investigating a combination of increased serial dependence and increased shape parameter. The shape parameter is increased well beyond second moments existing to study the effect on the power properties and any potential breakdown in the score statistic.

DGM

We consider marks with a conditional generalized Pareto distribution, with a varying shape parameter $\zeta = [0.05, 0.20, \dots, 0.95]$. Recall that when the shape parameter is $\zeta \geq 0.50$, second moments do not exist, but moments of order lower than 2 do exist. The score statistic for linearly boosted marks with a generalized Pareto distribution, is not properly defined if the variance of the score vector does not exist.

We study the power curves for i.i.d. distributed marks, $X \sim \text{GPD}(\zeta, \delta = 1.00)$ and in addition, marks that have a conditional generalized Pareto distribution scale parameter $X \sim \text{GPD}(\zeta, \delta_i)$, where $\delta_i = 0.9\delta_{i-1} + \epsilon_i$. For these simulation experiments we consider a sample size of $n = 1,000$.

Results

Figures 6.17 and 6.18 show the impact of increasing the shape parameter for marks that are both i.i.d. and serially dependent. In both cases Figure 6.17, and 6.18 show an improvement in the power of the score test as the shape parameter increases. The increased shape parameter results in marks with a distribution of longer tails, resulting in a greater impact on the intensity function. When we move from i.i.d. marks to serially dependent marks (Figure 6.18), the power is further increased for an increasing shape parameter. What is remarkable and unexpected, is that there is still no obvious break-down in the score statistic as the shape parameter increases well beyond the required levels for the second moments to exist.

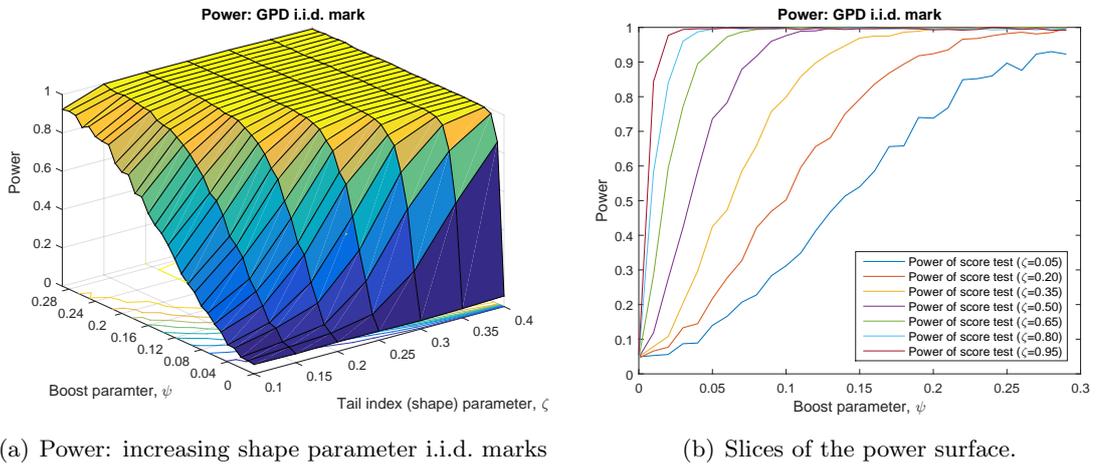
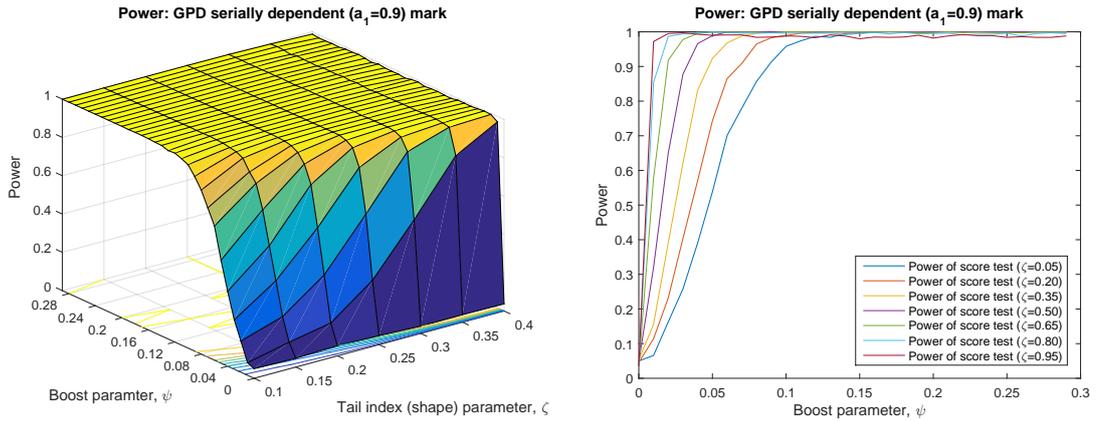


Figure 6.17: Power of the score test for a linearly boosted i.i.d. mark distributed $X_i \sim \text{GPD}(\zeta, \delta = 1.00)$ and increasing shape parameter ζ . The sample size is $n = 1,000$.

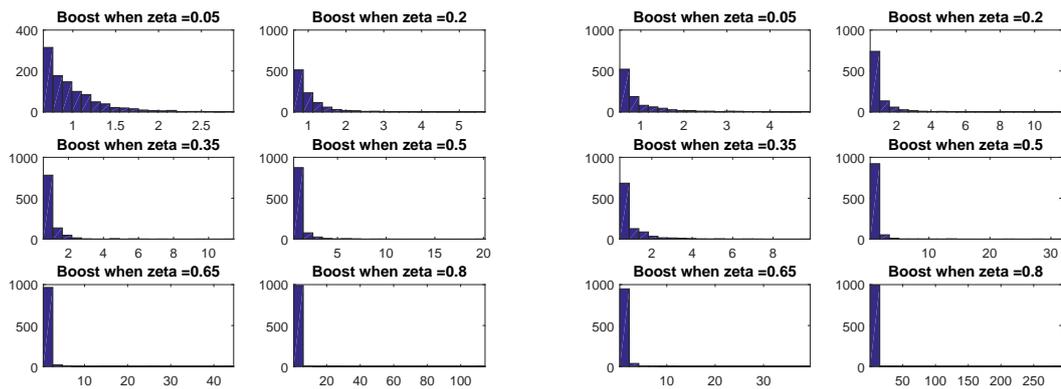


(a) Power: increasing shape parameter when marks have serial dependence.

(b) Slices of the power surface.

Figure 6.18: Power of the score test for a linearly boosted, serially dependent mark distributed $X_i \sim \text{GPD}(\zeta, \delta_i)$, where $\delta_i = 0.9\delta_{i-1} + \epsilon_i$ and an increasing shape parameter ζ . The sample size is $n = 1,000$.

As shown in the previous study of a constant shape parameter, but increasing serial dependence, we observed an increase in power of the score test due to the increased clustering of events. When this is coupled with an increasing shape parameter, this results in more events in the tail of the distribution and more cluster of those events. Figure 6.19(a) presents the increase in the boost function when increasing the shape parameter for marks that are i.i.d. Figure 6.19(b) presents the boost function under the formulation of an increase in the shape parameter, coupled with an increase in serial dependence. It is quickly apparent for all cases that there is significantly more events in the tails of the distribution. This will result in the mark having an increased impact on the intensity process.



(a) i.i.d. mark

(b) Mark with serial dependence, $a_1 = 0.9$.

Figure 6.19: Histograms of a linear boost function with a mark distributed $X_i \sim \text{GPD}(\zeta, \delta)$, for a single simulant, for each level of the shape parameter $\zeta \in \{0.05, 0.20, \dots, 0.80\}$, assuming both i.i.d. marks, where $\delta = 1.00$ and serially dependent marks, where $\delta_i = 0.9\delta_{i-1} + \epsilon_i$. The sample size is $n = 1,000$

Conclusion

For the case of a generalized Pareto distributed mark and a linear boost, the increase in the shape parameter improves the power of the score test. The increase in shape parameter, coupled with an increase in serial dependence, improves the power by increasing the clustering of events in the tails of the distribution. What is noteworthy, is that even in the face of second moments not existing and by using empirical moments, there is no break-down in the score test.

6.8.4 The discrete case with increasing serial dependence

Objective

As noted in Chapter 5, a small number of marks are best modelled by a discrete distribution and one which accounts for heavy tailedness. The negative binomial distribution was one of the appropriate distributions identified, motivating the requirement to assess the power and size characteristics of the score test when the underlying marks distribution is discrete. The aim of this study is to investigate the impact on the power and size of the score test under the conditions of increasing serial dependence, coupled with discrete heavy tailed marks.

DGM

We consider a mark with a conditional negative binomial distribution $NB(r_i, p = 0.50)$. The conditional negative binomial distribution success rate is defined as $r_i = a_1 r_{i-1} + \epsilon_i$. We vary the coefficients term $a_1 = [0.1, 0.2, \dots, 0.9]$. The study also includes the power curves for the i.i.d. case for comparison. Due to the computational time required for simulating marks that have a negative binomial distribution for size, $n = 1,000$, we reduce the size of each replicate to $n = 300$ for this study only.

Results

Figure 6.20 presents the power properties of the score test for increasing serial dependence. It is worth noting, that the serial dependence for the negative binomial distribution is invoked via the success rate r_i , hence we expect the impact of increased serial dependence to be similar to what we observed in the case of marks that have a generalized Pareto distribution. This is in fact the case, as shown in Figure 6.20. However, a distinct difference to the generalized Pareto distribution, is that we do not see the significant step up in power for the higher levels of serial dependence (Figure 6.20(b)), compared with what we observed for the GPD (Figure 6.11(b)). The negative binomial distribution tails are less pronounced than the GPD, therefore this result is expected.

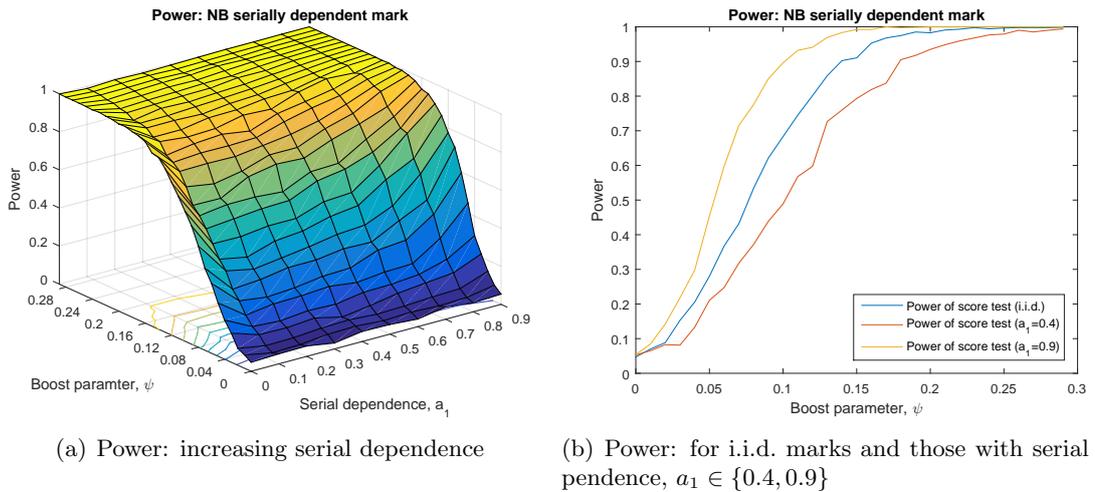


Figure 6.20: Power of the score test, with a linear boost, increasing serial dependence, for $X \sim \text{NB}(r_i, p = 0.50)$, with a sample size $n = 300$.

When we consider the histogram (Figure 6.21) of single replicate for the i.i.d. case and for each level of the serial dependence $a_1 \in \{0.4, 0.9\}$, there is consistent, but less pronounced results than what is observed with the heavier tailed generalized Pareto distribution in Figure 6.12.

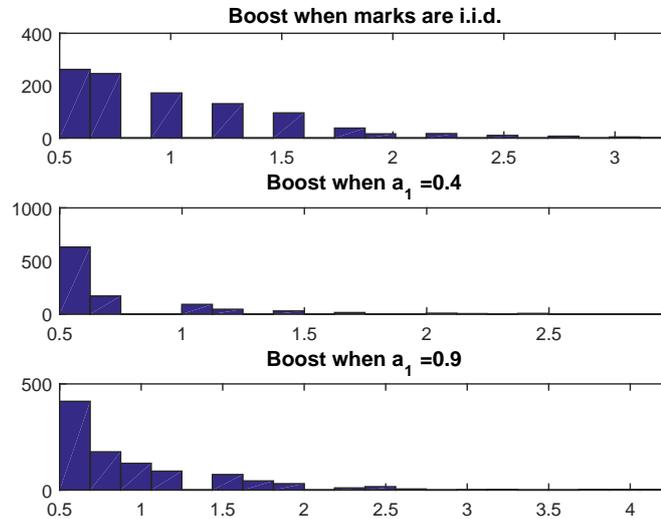


Figure 6.21: Histograms of the linear boost function, for a single simulant, i.i.d. marks and those with serial dependence $a_1 \in \{0.4, 0.9\}$, for $X \sim \text{NB}(r_i, p = 0.5)$, with a sample size $n = 300$.

Conclusion

Similarly to the generalized Pareto distribution, the negative binomial distribution exhibits stronger power properties when there is an increase in serial dependence. However, the power curves do not step up significantly with the introduction of serial dependence,

rather they exhibit a proportional increase in power, due to the fact that the tails are less pronounced than the generalized Pareto distribution. Even with the small sample size used for this study, the score test exhibits excellent size and power properties.

6.8.5 Power properties of the score test for high dimensional marks

Objective

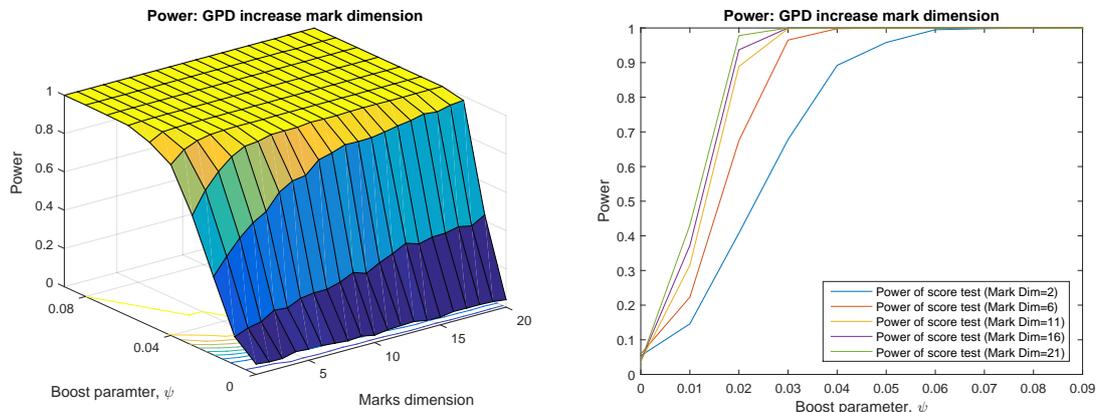
In light of the large number of potential marks that can be derived from the LOB, we now present a case study that extends on Section 6.6.1 and explores the size and power properties of the score statistic as we increase the marks dimension for the case of i.i.d marks. We previously observed a noticeable increase in the power of the score test for an increase in marks dimensions for both light and heavy tailed marks. The aim of this section is to explore this phenomenon further and to test the persistent with a greater dimension of marks.

DGM

We consider increasing dimensions of marks $d \in \{1, 2, \dots, 21\}$. The marks are distributed with a generalized Pareto distribution $X_i \sim \text{GPD}(\zeta = 0.10, \delta = 1.00)$. All marks have the same parameter specification and recall from Section 6.5, they are combined multiplicative within the boost function.

Results

Figure 6.22 shows the impact of an increase in i.i.d. generalized Pareto distributed marks dimension from 2 to 21 on the power properties of the score test. The power of the score is significantly improved as we increase the dimension. The effect is most noticeable when the dimension of the marks is low. Moving from $d = 16$ to $d = 21$, for example, has only a small incremental shift up in the power curves. For very high dimension of marks, the size and power properties of the score test are strong.



(a) Power: increasing marks dimension

(b) Power: for marks dimension $d \in \{2, 6, 11, 16, 21\}$

Figure 6.22: Power of the score test, with a linear boost, increasing marks dimension, and assuming i.i.d. GPD marks $X_i \sim \text{GPD}(\zeta = 0.10, \delta = 1.00)$. The sample size is $n = 1,000$.

The marks are combined in the boost function multiplicatively. The increase in significant marks into the Hawkes process process will increase the clustering of events. We can see how pronounced this impact is by considering the single replicate examples of the boost function for a range of different dimensions of marks in Figure 6.23. There is an evident increase in the boost function as the dimensions of the marks increase.

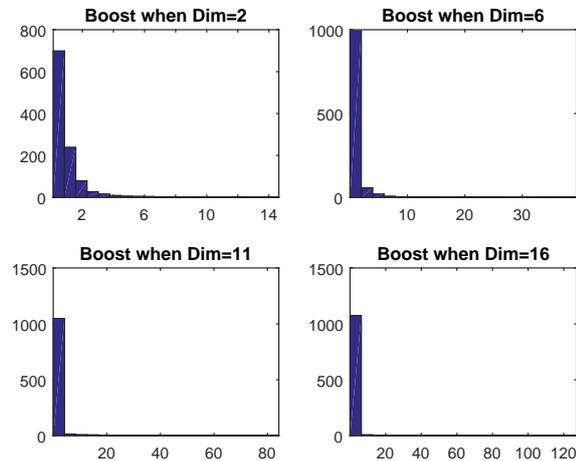


Figure 6.23: Histograms of the linear boost function for a single simulant, i.i.d. marks with dimension $d \in \{2, 6, 11, 16, 21\}$, for $X_i \sim GPD(\zeta = 0.10, \delta = 1.00)$. The sample size is $n = 1,000$.

Conclusion

Consistent with earlier simulation experiments in Section 6.6.1 for a heavy tailed distribution, the increase in marks dimension increases the power of the score test. However, the increase is most significant for incremental increases of low dimension, reducing as this moves to higher dimension. A high dimension mark vector does not degrade the size and power characteristics of the score test.

6.8.6 Impact of increasing the joint dependency between bivariate i.i.d. marks

Objective

The aim of this section is to study the effect on the power of the score test, of increasing the joint dependence between bivariate i.i.d. marks that are distributed with a heavy tailed distribution. We apply different copula models for joint dependence to assess whether the copula models degrade the performance of the score test.

DGM

We consider two dimensional i.i.d. marks with joint dependence. The marks are distributed with a generalized Pareto distribution $X_i \sim GPD(\zeta = 0.10, \delta = 1.00)$. We consider two cases where the joint dependence between the marks is modelled by a Gaussian copula

and a Gumbel copula, with a Spearman's rank correlation of $\rho_s \in \{0, 0.1, \dots, 1\}$. The choice of a Gumbel is to capture the dependence in the tails. Moments are estimated empirically and thus the correlation between the marks is also evaluated empirically. For these simulation experiments we consider a sample size of $n = 1,000$.

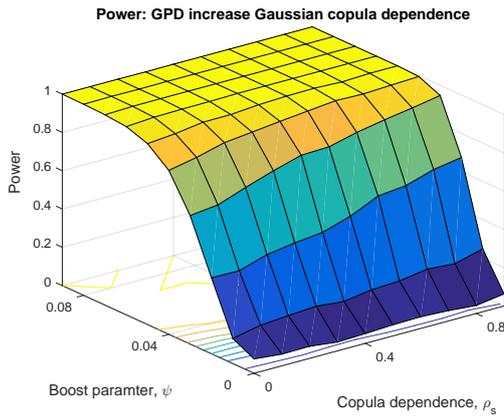
Results

For correlated marks data, both copula models (Figures 6.24 and 6.25) show an increase in the power of the score test when the correlation is higher. This is consistent with the results we observed in Section 6.6.1, when we studied the case of light tailed, jointly dependent marks with a Gaussian copula. When marks are correlated and combined multiplicatively, events will occur more frequently together, which will in turn have a bigger impact on the intensity function, than if independent.

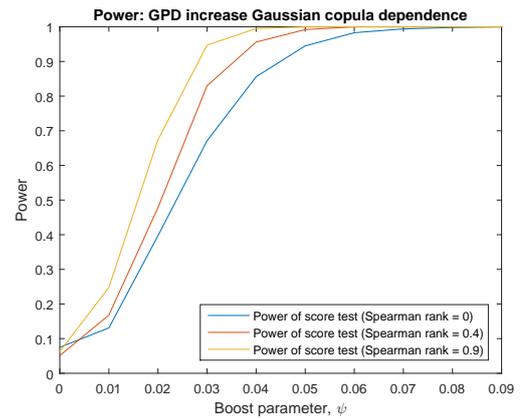
The Gumbel copula captures the dependence in the upper tail. The upper tail dependence of the Gumbel copula is $\lambda_U = 2 - 2^{1/\rho^\theta}$. In the bivariate case, the explicit expression for the Gumbel copula density is given by

$$\begin{aligned} c(u_1, u_2) &= \frac{\partial^2}{\partial u_1 \partial u_2} C(u_1, u_2) \\ &= C(u_1, u_2) u_1^{-1} u_2^{-1} \left[\sum_{k=1}^2 (-\ln u_k)^\rho \right]^{2\left(\frac{1}{\rho}-1\right)} (\ln u_1 \ln u_2)^{\rho-1} \\ &\quad \times \left[1 + (\rho - 1) \left[\sum_{k=1}^2 (-\ln u_k)^\rho \right]^{-\frac{1}{\rho}} \right]. \end{aligned}$$

Following on from the argument above, and given that the underlying distribution is heavy tailed, the impact of higher correlation in the tails will have a greater impact on boosting the intensity function compared to a Gaussian copula model, as shown in Figures 6.24(b) and 6.25(b). In addition, the impact of an increase in correlation between the marks is proportional to the increase in power for the Gaussian copula model, but this is not the case for the Gumbel copula model (Figure 6.25(b)). The increase in power for a Spearman's rank of $\rho = 0.4$, compared with uncorrelated marks is significant, however the effect diminishes as the correlation increases further.

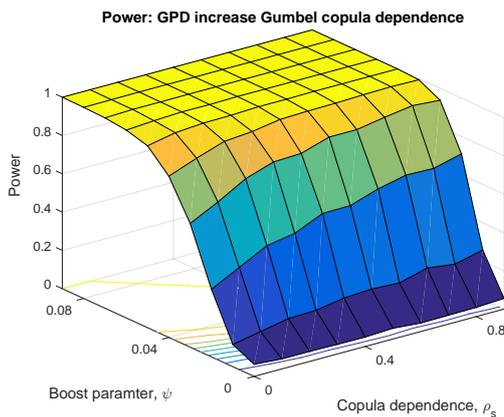


(a) Power: increasing copula dependence

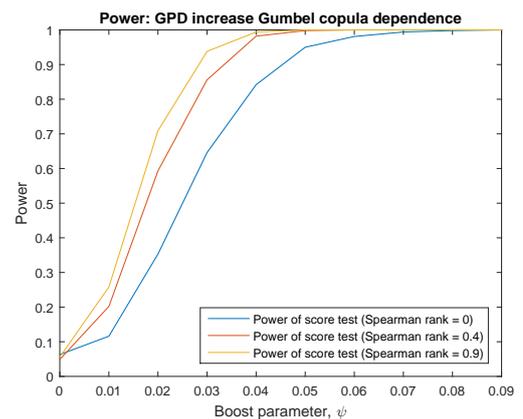


(b) Power: for Spearman's rank, $\rho_s \in \{0, 0.4, 0.9\}$

Figure 6.24: Power of the score test, with a linear boost, bivariate marks with Gaussian copula dependence, assuming i.i.d. GPD marks, $X_i \sim GPD(\zeta = 0.10, \delta = 1.00)$. The sample size is $n = 1,000$.



(a) Power: increasing copula dependence



(b) Power: for Spearman's rank, $\rho_s \in \{0, 0.4, 0.9\}$

Figure 6.25: Power of the score test, with a linear boost, bivariate marks with Gumbel copula dependence, assuming i.i.d. GPD marks $X_i \sim GPD(\zeta = 0.10, \delta = 1.00)$. The sample size is $n = 1,000$.

The histograms in Figure 6.26 of the boost function for a single simulation for each level of Spearman's rank correlation, shows an increase in tail behaviour for both the Gaussian and Gumbel copula as the correlation increases. In the case of the Gumbel copula, and in support of the discussions above, the increase in tail behaviour is further amplified, compared with the Gaussian copula when there is more correlation in the tails.

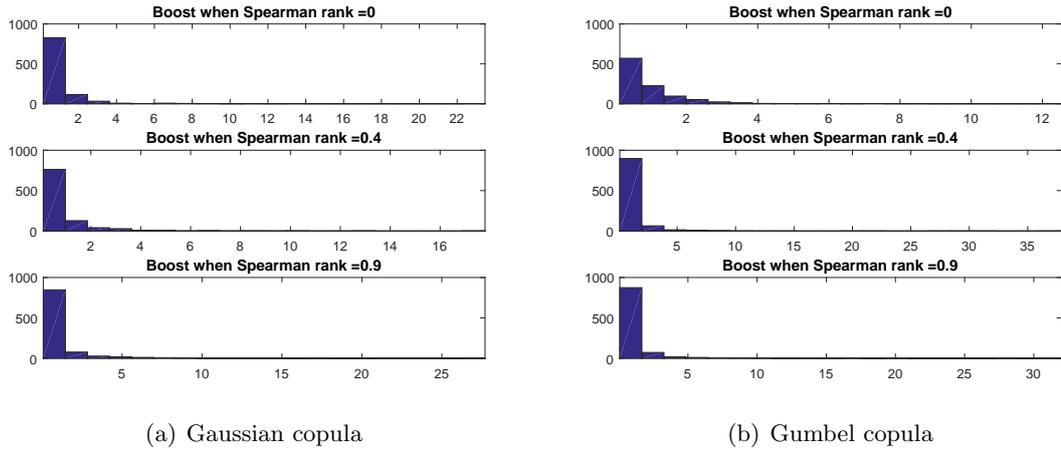


Figure 6.26: Histograms of the linear boost function for a single simulant, jointly dependent i.i.d. marks $X_i \sim GPD(\zeta = 0.10, \delta = 1.00)$, with Spearman's rank, $\rho_s \in \{0, 0.4, 0.9\}$. The joint dependence is modelled by a Gaussian and a Gumbel copula, respectively. The sample size is $n = 1,000$.

Conclusion

For marks that are jointly dependent and modelled by a Gaussian or Gumbel copula model, exhibit an increase in power of the score test when the Spearman's rank correlation increases. The effect for the Gaussian copula is proportional with the increase in correlation, however for the Gumbel copula the increase in power is both stronger than the Gaussian copula and shows diminishing effect of increased power as the correlation increases. This is in part due to the combined effect of a heavy tailed marks distribution and correlation in the tails. We observed good size and power properties of the score test in the presence of joint dependence and when evaluating the correlation between the marks empirically.

Only the bivariate case was investigated when studying the power properties of jointly dependent marks. The adjustment required to normalize the boost function (Section 4.2.1), has been developed for a two dimensional mark vector only. It is expected that future work will extend upon this. The combined effect of joint dependence and increasing mark dimensions will likely further increase the power of the score test when marks have a heavy tailed distribution, but to what extent remains a question for future research.

6.8.7 Impact of increasing the joint dependency between bivariate serially dependent marks

Objective

The aim of this study is to explore the effect of serially dependent, bivariate, heavy tailed marks with joint dependence on the power of the score test. This final study is vital for the application to real data that follows, as it replicates the features of the marks that we observed in Chapter 5 and provides a robust study of the power properties of score test under realistic conditions.

DGM

We consider bivariate, serially dependent marks with joint dependence. The marks are combined multiplicative, each with a linear boost in (6.48). They are distributed with a generalized Pareto distribution $X_i \sim \text{GPD}(\zeta = 0.10, \delta_i)$. The conditional generalized Pareto distribution scale parameter is defined as $\delta_i = 0.9\delta_{i-1} + \epsilon_i$. The joint dependence is modelled via a Gaussian copula model with varying Spearman's rank correlation $\rho_s \in \{0, 0.1, \dots, 1\}$. Again we utilize empirical moments and the simulation experiments are of sample size $n = 1,000$.

Results

From the extensive studies of the power properties, we expect that serially dependent marks, with a generalized Pareto distribution and joint dependence will exhibit strong size and power characteristics for the score test. As we can see in Figure 6.27, this is in fact the case. As the correlation of the copula increases, so does the power of the score test.

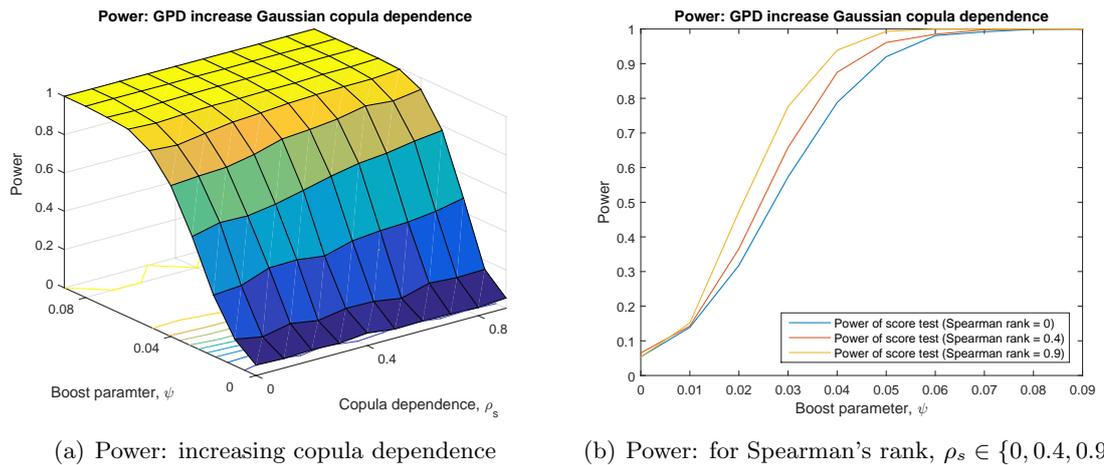


Figure 6.27: Power of the score test, with a linear boost, bivariate marks with Gaussian copula dependence, assuming serially dependent marks with a conditional GPD $X \sim \text{GPD}(\zeta = 0.10, \delta_i)$, where $\delta_i = 0.9\delta_{i-1} + \epsilon_i$. The sample size is $n = 1,000$.

Further supporting the findings above, the impact to the boost function of combining serial and joint dependent marks in a multiplicative form, leads to increased tails in the histogram of the boost function as the joint correlation increases (Figure 6.28).

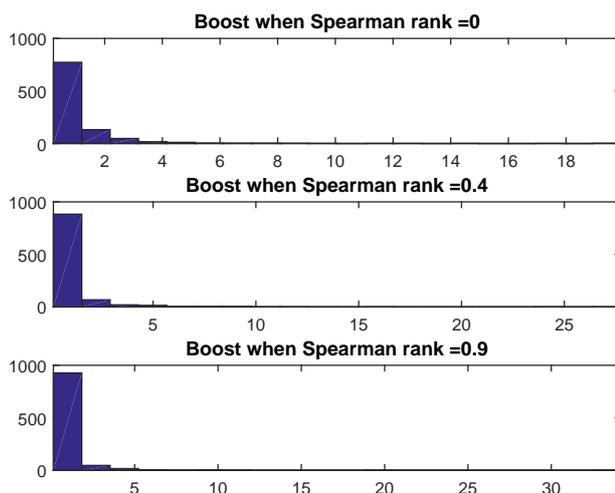


Figure 6.28: Histograms of the linear boost function for a single simulant, jointly dependent marks with Spearman’s rank $\rho_s \in \{0, 0.4, 0.9\}$. The joint dependence is modelled by a Gaussian copula. The marks are serially dependent, with a conditional GPD $X \sim \text{GPD}(\zeta = 0.10, \delta_i)$, where $\delta_i = 0.9\delta_{i-1} + \epsilon_i$. The sample size is $n = 1,000$.

Conclusion

Under realistic assumptions of heavy tailed marks with serial and joint dependence features, the score test has strong size and power properties and power increases as the joint dependence increases. This provides assurance that the score test is reliable for the use of marks constructed from LOB data.

6.9 Conclusion

The limit order book displays self-excitation behaviour within a high frequency setting, making it appropriate for the modelling of Hawkes process. Upon initial inspection, the inclusions of marks into the Hawkes process improves the fit of the models and makes the models more amenable to application. There is a large catalogue of potential marks that could be created, however there is limited literature guiding the selection of appropriate marks. To our knowledge, there does not exist a detection method for marks in Hawkes process.

The score test provides a method to detect the existence of marks without parameter estimation via the joint likelihood. The score statistic is asymptotic chi-squared with r degrees of freedom. Under various combinations of sample sizes, boost function and marks density, we show that the chi-squared distribution can be used for obtaining sufficiently accurate quantiles for application.

The score test has good power characteristics under various marks distributions. The use of empirical moments does not degrade the size or power of the score statistic and due to the greater flexibility and ease of implementation, empirical moments are advised. Volume based marks are often heavy tailed and have serial dependence. A procedure is

proposed for adjusting the covariance matrix, used to normalise the score vector, for serial dependence.

From these studies we can conclude that the score test used for marks with joint, serial dependence and heavy tailed marginals, will have excellent size and power properties. This will be further amplified in a multivariate setting beyond the two dimensional studies and again further increased with the use of copulas that capture the joint tail dependence. Coupled with the ease of implementation, and the consideration of the complex data sets of the LOB, this makes the score test a powerful and flexible tool to use when identify appropriate marks for the Hawkes process.

Chapter 7

Score test application and the decoupled approximate likelihood method

This research started with the exploration of heavy tailed features of volumes on the LOB and the appropriate methods of modelling these volumes in Chapter 2. The research was further extended to investigate models for the LOB, which could incorporate these volume features and other important aspects of the LOB. The Hawkes process was identified as a potentially useful model for the dynamics of the intensity function, because it allows for irregularly spaced time sequences, a multivariate framework, multiple dependent marks and the ability to capture the impact of marks on the intensity. As discussed in substantial detail in Chapter 3, a critical first step to successfully apply these models to the LOB is to identify the appropriate events in terms of order book levels and event types of orders, that constitute the event times of the Hawkes process.

Prior to the application of the Hawkes process to the LOB, the identification of marks was studied in Chapter 5. What became apparent when reviewing the literature on marked Hawkes process, was the lack of guidance that past literature provides on the possible marks that can be incorporated. This motivated the cataloguing of an extensive (but not exhaustive) list of potential marks, with some based on financial theory, that could be considered and the detailing of their properties and dependence features (Chapter 5).

The LOB data sets are complex and large in size. This poses many challenges when fitting a Hawkes process with multivariate marks (presented in Chapter 4). These issues are further exacerbated by the large list of potential marks. This motivated the development of a novel detection method based on the likelihood score statistic which was proposed in Chapter 6. The score test can be used when a general decay function in the intensity process is replaced by various boost functions with multiple marks being assessed. Although not developed in this thesis, the score test can be extended to assess marks for a multivariate Hawkes process. The advantage of this statistic is that the Hawkes process can be fit under the hypothesis that the intensity is independent of the marks. The statistic was shown to have good size and power properties with the use of empirical

moments. This means that parametric models for the marks do not need to be specified. In addition, the score test is able capture the joint dependence between the marks and a suitable adjustment is made for the case of marks with serial dependence, all of which are features within the data sets we observed in Chapter 5.

This chapter brings together the various strands of research within this thesis to apply the score test to the marks identified in Chapter 5. We will begin by introducing the key data sets that we will consider in this study and specify the event process that will underpin the marks creation and intensity function of the Hawkes process. Modelling an entire day or multiple days of events is not appropriate with a single stationary model. This is due to the non-stationarity of the intensity process and computational challenges the vast data sets present, which is discussed in detail in Section 4.2.4). We will study the appropriate intra-day time segment, to assess the marks impact via the score test and to fit the Hawkes process. The assessment of the marks via the score test will be conducted with the identification of appropriate marks for modelling in Section 7.2. The second section of this chapter will explore the challenges of parameter estimation via the log-likelihood with increasing model complexity and marks dimension, joint and serially dependent marks and the presence of heavy tailed marks. The final section will propose a decoupled approximate likelihood method using empirical moments for the estimation of the Hawkes process. This will be compared with results from the log-likelihood and provide a more robust framework for modelling the Hawkes process, as the distribution for the marks does not need to specified.

7.1 Score test application to real data

We illustrate the application of the score test to the LOB using the final 10 trading days in July 2015 for SILVER and for comparative purposes, NIKKEI. These two assets have been studied extensively throughout this research in Chapters 2, 3 and 5. Recall, SILVER is a commodities futures with the underlying asset being silver. It is listed on COMEX, which is a primary futures and options market for trading metals such as gold, silver, copper and aluminium. Whereas NIKKEI is a futures contract, with the underlying asset being the Nikkei 225 Equity Index. NIKKEI is listed on the Singapore Exchange (SGX), and the CME, but we do not consider LOB data from the CME for this asset. The stock market index comprises of stocks listed on the Tokyo Stock Exchange (TSE). These two assets present vastly different features, are highly liquid and provide a good example of the application of the score test to real data. Futures contracts have expiration dates as opposed to stocks that trade in perpetuity. In this research we study the *front* futures contract. The historical data used within this research contains a continuous futures history or the *on-the-run* front contract, so joining contract data is not necessary. Refer to Section 2.1 for further details on roll dates.

Within this chapter we consider an event process, which is defined using the matched LOB data, with event types $e \in \{LO, MO, C\}$, levels $l \in \{1, \dots, 5\}$ and for the bid side only. This choice is informed by the conclusions reached in Section 3.2.1.

To demonstrate the robustness and versatility of the score test, it will be applied to

assess the suitability of all the marks that we defined in Chapter 5 and listed below in Table 7.1. Recall, from Chapter 6, the score test can be defined using empirical moments for the function $G(\mathbf{X})$ and $\text{Cov}(G(\mathbf{X}))$, rather than estimates based on theoretical moments evaluated at sample estimates of the marks distribution parameter ϕ . Use of empirical estimates does not degrade the size and power performance of the score statistic. This is an important point, given findings from Chapter 5 which demonstrated that finding the best distribution for the marks can be difficult for many marks. In addition, we have marks that may be best represented by either continuous or discrete distributions. In view of this, our application of the score test will exclusively use empirical moments. There are several advantages of using empirical moments: there is no need to find the best marginal or joint distribution for the marks; whether the mark is discrete or continuous can be ignored; and a model for serial dependence may not be needed.

The transformations of the marks, which were proposed in Table 5.3, are applied to the marks in this section. Table 7.1 presents all marks as they were defined in Section 5.2 and this table is the naming convention used throughout this chapter.

Table 7.1: Shortened naming convention with equation reference for the endogenous marks constructed from matched LOB data. The numbers following the name prior to the equation reference are used in the charts that follow.

Depth based:	Centred depth/price based:	Price based:
Bid/Ask depth-1 (5.1)	Imbalance-9(5.9)	Rel price LO-12 (5.17)
Bid/Ask opp. side depth-2 (5.2)	Rel imbalance-10 (5.10)	Rel price C-13 (5.18)
Volume based:	Mid-price-11 (5.11)	Count based:
Bid/Ask vol MOLOC-3 (5.3)	Mid-price returns-14 (5.12)	Bid/Ask count MOLOC-19 (5.19)
Bid/Ask vol MOLO-4 (5.4)	Spread-15 (5.13)	Bid/Ask count LO-20 (5.20)
Bid/Ask vol MO-5 (5.6)	Traded MO price-16 (5.14)	Bid/Ask count C-21 (5.21)
Bid/Ask vol LO-6 (5.5)	Volatility mid-price-17 (5.15)	
Bid/Ask vol C-7 (5.7)	Volatility mid-price ret.-18 (5.16)	
Inside vol MO-8 (5.8)		

7.1.1 Selection of intra-day time segments

A constant immigration intensity is a requirement of the stationary Hawkes process. It is possible to incorporate a functional form of the immigration intensity to account for time-varying nature of the data, for example Bowsher (2007), Toke (2011), Gao et al. (2017), Omi et al. (2017) and Chen and Hall (2013) construct a Hawkes process with an exogenous immigration intensity and Stindl and Chen (2018) incorporate a renewal process. However, as noted in Filimonov and Sornette (2015), Omi et al. (2017) if the dynamics of $\eta(t)$ are convoluted, many parameters may be needed to capture its complexity. Even in the case of a constant boost (no marks) this makes the estimation less robust and often degenerate (Filimonov and Sornette, 2015), with the same level of the likelihood function resulting in several solutions with different parameter values. To date, we are not aware of any research or methods that simultaneously address the challenges of estimating a marked Hawkes process with time varying immigration. The research in this thesis focuses on detecting and modelling marks and because of this, intra-day variability in the immigration component will be adjusted for using a somewhat ad-hoc approach.

For each of the assets we consider, the total number of events on a single side of the

LOB and for levels 1:5 is massive. For a single trading date 31-July-2015, SILVER has 60,691 event times recorded within the constraints of *liquid market hours* (Section 2.1.1). For NIKKEI, the total number of event times is 41,107. With such vast amounts of data, we proceed by slicing the day into time segments, rather than using the entire day of data to apply the score test and fit the marked Hawkes process models. This is primarily motivated by the non-stationarity of the intensity process observed throughout the day and the computational challenges of fitting the model to such large datasets. The first aim of this section is to determine the appropriate intra-day time segment size, which balances the following criteria:

Criterion 1. *A sufficiently large sample size within the time segments to ensure good power properties of the score test and reasonable log-likelihood fits;*

Criterion 2. *To minimize the time segment size, to ensure approximate stationarity of the intensity function;*

Criterion 3. *To ensure there is a sufficient proportion of time segments over which a selected mark is significant.*

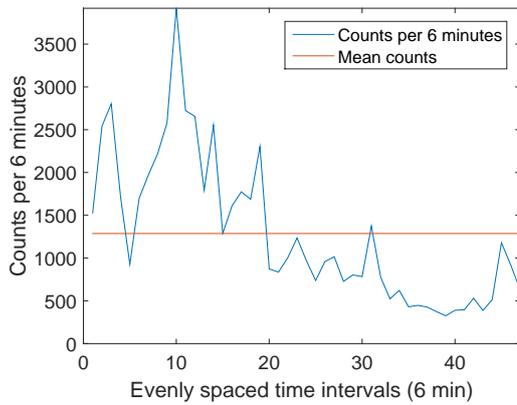
As a part of this assessment, we will address the second aim of this section, which is to select two suitable marks for the application study in Section 7.2.

To address Criterion 1, we refer back to the simulation studies in Section 4.5.2 which studied the effect of increasing sample size on the robustness of the parameter estimates for a Hawkes process with a generalized Pareto distributed mark. The bias in the parameter estimates versus the true values was low for sample sizes of, $n = 600$ and greater. In Section 6.6.1 we studied the score test statistic over a range of sample sizes, and whilst we observed improvement in the power properties of the score test as the sample size increased, the score test performed well for the smallest sample size studied, which was approximately $n = 500$.

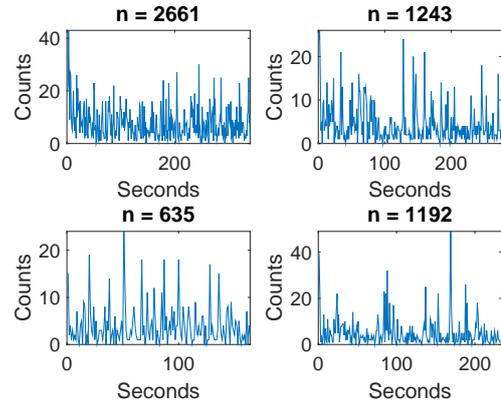
When modelling the intra-day LOB activity of a financial instrument at the minimum time granularity of one millisecond, there will be presence of non-stationary within the time series as a result of regime switching rather than local clustering. If the regime switching is not accounted for, the Hawkes process, which assumes stationarity, will interpret this as an increase or decrease in clustering. Formal testing for stationarity of the intensity process is not yet developed. The assessment is made by inspecting the data visually.

Figures 7.1(a) and 7.2(a) display the count of events across evenly spaced time intervals of 6 minutes each for SILVER and NIKKEI. What is immediately apparent is the time-varying nature of the counting process throughout the day. We note that the underlying stock market index for NIKKEI, the Nikkei 225 is an equity index, of which the constituent instruments are listed on the Tokyo Stock Exchange (TSE). The TSE closes between 11:30pm and 12:30pm local time, corresponding to the dip in activity for NIKKEI. Either side of the ‘lunchtime dip’ in activity, NIKKEI displays a series that is closer to stationarity than SILVER. Therefore, we are able to consider much larger time segments for this study. Future work will be required to incorporate an immigration rate as a deterministic function of time to account for the lunchtime dip, and which is beyond the scope of this thesis.

Figures 7.1(b) and 7.2(b) present four plots of 6 minute slices for SILVER and 40 minute slices for NIKKEI, with counts on one second evenly spaced time intervals. The slices are chosen at the first, second, third and fourth quarter intervals of the day. We can see that the non-stationarity diminishes with smaller time segments, making modelling intra-day time segments, rather than the entire day or multiple days appropriate. The time segments of 6 minutes for SILVER and 40 minutes for NIKKEI, provide sufficient counts of events within each segment (Criterion 1), whilst ensuring approximate stationarity (Criterion 2). The mean count of events within each time segment across the entire day are, $\bar{n} = 1287$ for SILVER and $\bar{n} = 4714$ for NIKKEI.

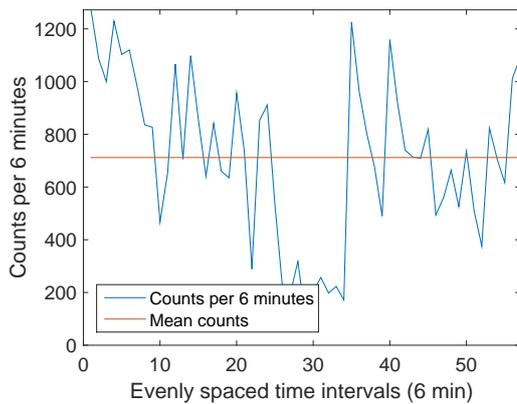


(a) Counts per 6 minutes throughout the day

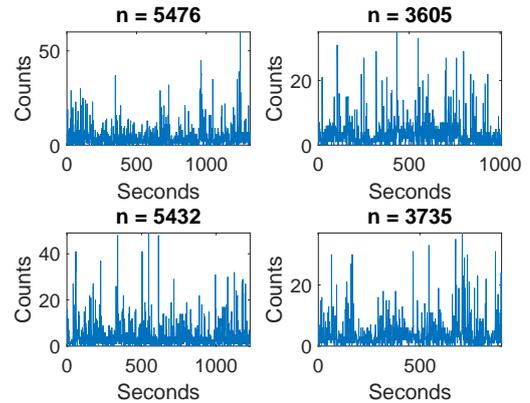


(b) Counts per 1 second for four selected 6 minute time segments throughout the day

Figure 7.1: Counts of events across even time intervals of 6 minutes for the entire day. The four plots show the counts of events across even time intervals of 1 second, for a duration of 6 minutes. The asset is SILVER, bid side and trading date 31-Jul-2015. Event types correspond to events $e \in \{LO, MO, C\}$ and on levels $l \in \{1, \dots, 5\}$.



(a) Counts per 6 minutes throughout the day



(b) Counts per 1 second for four selected 20 minute time segments throughout the day

Figure 7.2: Counts of events across even time intervals of 6 minutes for the entire day. The four plots show the counts of events across even time intervals of 1 second, for a duration of 40 minutes. The asset is NIKKEI, bid side and trading date 31-Jul-2015. Event types correspond to events $e \in \{LO, MO, C\}$ and on levels $l \in \{1, \dots, 5\}$.

Criterion 3 is addressed by evaluating the proportion of time segments for which a select mark is significant and to ensure this proportion is sufficiently large. For each mark and for each time segment considered, we use the Ljung-Box Q-test with a default lag of three, to assess for serial dependence within the marks. If the test rejects the null hypothesis, we implement the serially dependent adjusted score test in (6.46), otherwise we use the standard method in (6.35). For the calculation of $G(\mathbf{X}; \phi)$, we use the empirical estimate for moments as discussed above.

The assessment of the marks significance via the score test is conducted across time segments for each of the 10 trading days. We consider multiple time segments to test the robustness of significance of the mark for varying time segments from $n \in \{1,000, 2,000, \dots, 10,000\}$ and which will inform Criterion 3, selecting an appropriate intra-day time segment. The colour maps presented in Figures 7.3 display the proportion of significant marks across all time segments. We use the score test to evaluate each mark for each time segment and then calculate the proportion of time segments for which the mark is significant out of the total number of time segments. The lighter the colour in the map, the more times the mark boost parameter is significantly different from zero.

For both SILVER (Figure 7.3(a)) and NIKKEI (Figure 7.3(b)) the sample sizes of 3000 for SILVER and 8000 for NIKKEI present a large increase in significance of each mark. This is likely to be a by-product of data aggregation. SILVER shows higher robustness of significance of marks across all time segments considered, compared with NIKKEI. If a particular mark is not significant with the smaller time segments, this tends to persist across the larger time segments. NIKKEI on the other hand, presents a greater number of marks significantly impacting the intensity process.

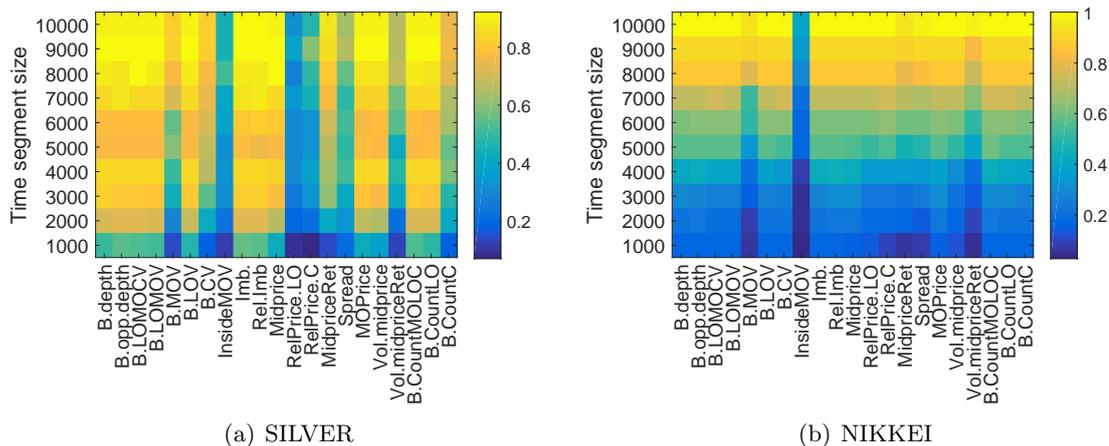


Figure 7.3: *Individual marks.* The proportion (*right hand column scale*) of segments a mark is significant by the score test, across all time segments size $n \in \{1,000, 2,000, \dots, 10,000\}$, for 10 trading days 20-Jul-2015 to 31-Jul-2015, bid side, for SILVER and NIKKEI. Marks are evaluated using matched LOB data, with event types $e \in \{LO, MO, C\}$ and levels $l \in \{1, \dots, 5\}$.

Table 7.2 presents the proportions in the heat map above, as a percentage of significant marks, as defined by the score test, across select time segments of which don't

appear to have excessively high proportions of significant marks due to data aggregation. For SILVER we consider time segment size $n \in \{1,000, 2,000\}$ and for NIKKEI $n \in \{5,000, 6,000, 7,000\}$. The proportion of marks that significantly impact the intensity process for SILVER, varies much more than what is observed for NIKKEI. Whilst there are marks that are significant for both SILVER and NIKKEI, there are some which are asset specific. For example, Rel price LO and Rel price C are frequently significant for NIKKEI, whereas for SILVER it is not a key mark to consider in the Hawkes process. These findings highlight the importance of individual asset assessment of a marks impact on the intensity process. The choice of time segments of 6 minutes for SILVER ($\bar{n} = 1,287$) and 40 minutes for NIKKEI ($\bar{n} = 4,714$) ensure there is sufficient proportion of time segments over which a selected mark is significant, meeting Criterion 3.

Table 7.2: The proportion of significant marks, as defined by the score test, across time segment size, $n \in \{1,000, 2,000\}$, for SILVER and $n \in \{5,000, 6,000, 7,000\}$, for NIKKEI, across 10 trading days from 20-Jul-2015 to 31-Jul-2015. Marks are evaluated using matched LOB data, with event types $e \in \{LO, MO, C\}$, levels $l \in \{1, \dots, 5\}$ and for the bid side only.

Mark	SILVER			NIKKEI	
	1,000 events	2,000 events	5,000 events	6,000 events	7,000 events
Bid depth	50.6%	71.6%	56.8%	61.7%	72.5%
Bid opp. side depth	54.6%	73.1%	56.8%	63.3%	72.5%
Bid vol MOLOC	53.4%	72.6%	56.8%	63.3%	74.5%
Bid vol MOLO	51.8%	71.6%	56.8%	63.3%	72.5%
Bid vol MO	14.2%	28.8%	33.8%	50.0%	52.9%
Bid vol LO	48.0%	69.2%	56.8%	63.3%	72.5%
Bid vol C	18.4%	42.7%	54.1%	63.3%	74.5%
Inside vol MO	12.1%	23.6%	17.6%	16.7%	19.6%
Imbalance	55.1%	72.6%	56.8%	61.7%	72.5%
Rel imbalance	53.9%	73.1%	56.8%	61.7%	72.5%
Mid-price	42.6%	65.9%	55.4%	61.7%	72.5%
Rel price LO	8.7%	19.2%	52.7%	61.7%	72.5%
Rel price C	7.1%	22.1%	55.4%	65.0%	74.5%
Mid-price returns	14.2%	40.8%	45.9%	60.0%	70.6%
Spread	20.8%	33.7%	48.6%	60.0%	68.6%
Traded MO price	42.3%	67.8%	56.8%	63.3%	72.5%
Volatility mid-price	37.4%	64.4%	50.0%	61.7%	72.5%
Volatility mid-price ret.	12.5%	22.1%	40.5%	50.0%	62.7%
Bid count MOLOC	53.4%	71.2%	59.5%	63.3%	74.5%
Bid count LO	48.5%	71.2%	55.4%	63.3%	74.5%
Bid count C	18.9%	41.3%	55.4%	61.7%	72.5%

Below is a list of marks with the highest proportion of significance time segments for 1,000 events (Table 7.2) for SILVER only, which closely matches the number of events in the chosen time segment of 6 minutes. The marks that have the highest proportion of significant time segments are all volume based. The marks that are selected for the application of modelling SILVER LOB data with a Hawkes process in Section 7.2, are Bid vol MOLOC and Bid depth. This choice is also motivated by Rambaldi et al. (2017) and Kirchner (2017b) who have incorporated event volumes into a Hawkes process. The Bid depth has not been studied as a mark in a Hawkes process for LOB data, however it has been identified as an important stylized feature of the LOB (Section 5.1).

- Imbalance in (5.9) (55.1%);

- Bid opp. side depth in (5.2) (54.6%);
- Rel imbalance in (5.10) (53.9%);
- Bid vol MOLOC in (5.3) (53.4%) (**selected for application**);
- Bid count MOLOC in (5.19) (53.4%);
- Bid vol MOLO in (5.4) (51.8%);
- Bid depth in (5.1) (50.6%) (**selected for application**).

7.1.2 Score test of pairwise marks

We begin our application of the score test to multivariate marks by considering all pairwise combinations of the marks selected above. Recall from Section 5.3.5, the marks are typically strongly pairwise correlated. This correlation could impact the effectiveness of the score test for detecting marks and needs to be accounted for in the score test. Likewise, for likelihood fitting of the marked process, normalization of the boost requires modelling of bivariate dependence, typically using a copula as discussed in Section 4.2.1.

Figure 7.4 presents the proportion of significant bivariate marks across the time segments of 6 minutes for SILVER and 40 minutes for NIKKEI. The pairs of marks are indexed by i and j , where $i \in \{1, \dots, 20\}$ and $j \in \{2, \dots, 21\}$. Table 7.1 presents the names corresponding to the numerical values i and j can take. The combinations of pairs of marks considered are $\{(1, 2), (1, 3), \dots, (2, 3), (2, 4), \dots, (20, 21)\}$. As with the heat charts in the case of single marks, a lighter colour represents a higher proportion of times the bivariate marks are significantly different from the null hypothesis. Comparing SILVER with NIKKEI, a higher number of bivariate marks are significant for NIKKEI. This is consistent with the evaluation of the univariate marks in Figure 7.3.

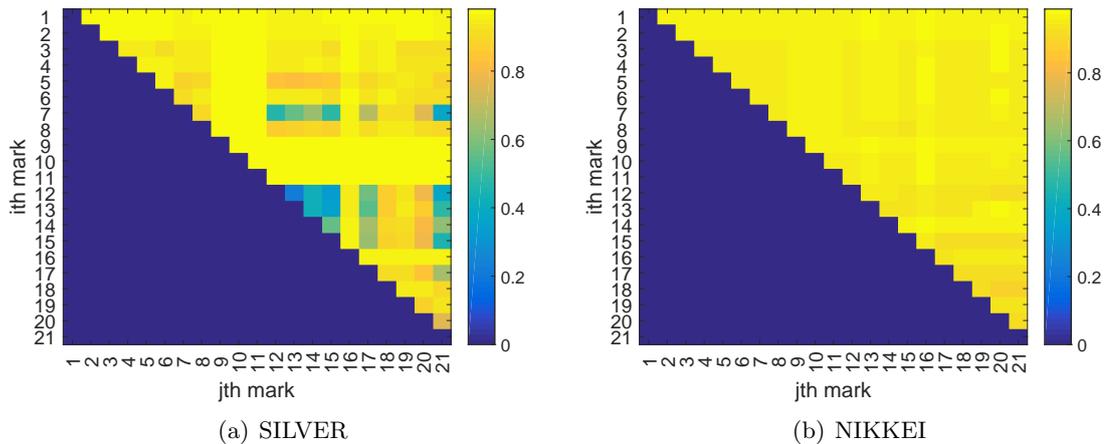


Figure 7.4: *Pairwise marks*. The proportion of time segments, which pairs of marks are significant via the score test, for time segments of 6 minutes, for SILVER and 40 minutes, for NIKKEI, across 10 trading days from 20-July-2015 to 31-July-2015. Combinations of $\{i, j\}$ marks are $i \in \{1, \dots, 20\}$ and $j \in \{2, \dots, 21\}$. Table 7.1 shows the names corresponding to the numerical values i and j . Marks are constructed using matched bid side LOB data, with event types $e \in \{LO, MO, C\}$ and levels $l \in \{1, \dots, 5\}$.

In Section 7.2.2 we present the fitting of the Hawkes process with bivariate marks selected above, Bid vol MOLOC and Bid depth. The combination of these marks ($i = 1, j = 3$), are depicted in Figure 7.4 for column $i = 1$, row $j = 3$. The assessment of this pair by the score test, results in significance of 96.7% of the time segments considered across 10 trading days.

7.1.3 Score test of higher dimensional marks

In future application, it is possible that many marks that exhibit joint dependence will be incorporated into the Hawkes process. It is not possible to plot all combinations of marks, however to demonstrate the flexibility of the score test, we present the proportion of significant marks, which are assessed jointly, and are sequentially introduce in an increasing number. A test of this kind would not be computational feasible if assessment was conducted with a model of the log-likelihood, again highlighting the significant practical application of the score test.

Figure 7.5 displays the proportion of significant marks of dimension $d \in \{2, \dots, 21\}$, across the defined time segments of 6 minutes for SILVER and 40 minutes for NIKKEI. The marks are indexed by i and j with Table 7.1 presenting the names corresponding to the numerical values of i and j . The combinations of marks considered are $\{(1, 2), (1, 2, 3), \dots, (1, \dots, 21), \dots, (2, 3), (2, 3, 4), \dots, (2, \dots, 21), \dots, (20, 21)\}$. When the mark, inside MO Volume, is introduced into the sequence of marks for both SILVER and NIKKEI, the sequence reduces in significance from approximately 94.8% of time segments, to 47.5% of time segments (Figure 7.5). Combinations of relative price of LO and C and mid-price returns have contributed in a reduction of the number of time segments that are significant for SILVER (Figure 7.5(a)).

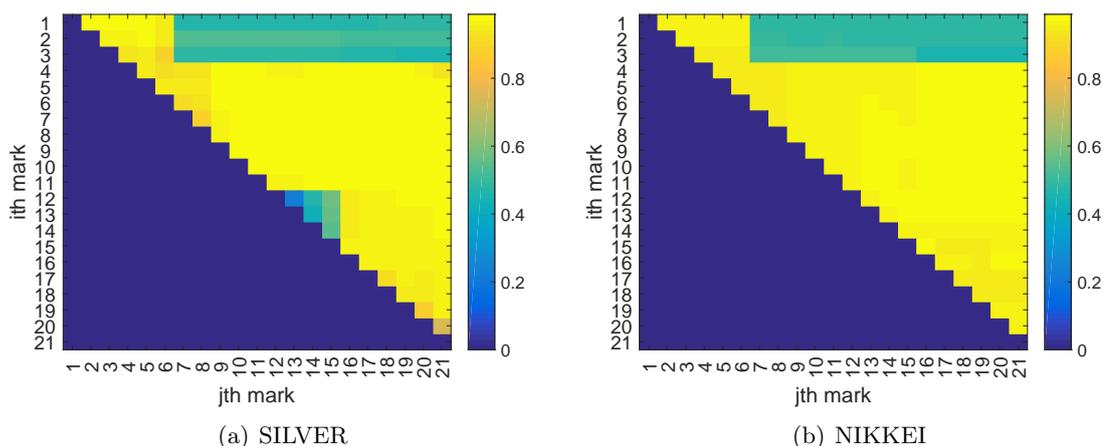


Figure 7.5: *Multiple marks.* The proportion of significant marks via the score test, for time segment $n = 1,000$, for SILVER and $n = 5,000$, for NIKKEI, across 10 trading days 20-July-2015 to 31-July-2015, for combinations $j:i$ of marks. Marks are constructed using matched bid side LOB data, with event types $e \in \{LO, MO, C\}$ and levels $l \in \{1, \dots, 5\}$. Table 7.1 shows the names corresponding to the numbers in the chart.

In conclusion, we select a time segment of 6 minutes for SILVER and 40 minutes for

NIKKEI, which ensures that the sample size of events is large enough to give reliable score test results and reasonable likelihood fits. We observe that the intensity of events is clearly not stationary throughout the day for both assets. We expect this to be the same for other trading days and assets. The time segment choice produces approximately stationary counts. Finally, this conservative choice of time segments are sufficiently large enough to ensure an adequate proportion of segments for selected marks, are significant and not an artefact of aggregation.

From assessment of the marks for SILVER using the chosen time segment and across 10 trading days, Bid vol MOLOC and Bid depth have been selected as marks for the application of the Hawkes process in Section 7.2. When the score test was applied to bivariate marks, the assessment of the combination of Bid vol MOLOC and Bid depth resulted in a significance of 96.7% of the time segments considered. A demonstration of the score test for higher dimensional marks, where $d \in \{2, \dots, 21\}$ presented the flexibility of the test and speed at which this could be performed. As noted, a test of this kind would not be computationally feasible if conducted with a model for the log-likelihood.

7.2 Fitting marked Hawkes process models

The section that follows will explore the challenges of the log-likelihood parameter estimation with increasing model complexity and marks dimension, joint and serially dependent marks and in the presence of heavy tailed marks. The final section will propose a decoupled approximate likelihood method using empirical moments for the estimation of the Hawkes process parameters. This will be compared with results from the log-likelihood method, providing a more robust framework for modelling the Hawkes process, as the distribution for the marks does not need to be specified. The aim of this section is to identify the non-trivial challenges of fitting a marked Hawkes process when there is serial dependence and to present alternative methods for parameter estimation.

For the application of fitting the Hawkes process, we consider SILVER for a single trading date 31-July-2015. Future studies will consider a much broader asset selection and time frames. However, it is anticipated that the challenges and findings we present here are not a function of the asset or day selected, but apply to most liquid LOBs of different assets. This will form an element of future research.

In Section 7.1.1 we identified non-stationarity in the intensity process and proposed an appropriate intra-day time segment of 6 minutes for SILVER to ensure a sufficiently large sample size for assessment via the score test and model fitting, whilst minimizing the time window to reduce non-stationarity in the intensity process. As we noted, the recommended time segment presented approximate stationarity and for the trading day 31-July-2015, represents 47 time segments which will be modelled by a Hawkes process.

Based on the findings in Sections 7.1.1 and 7.1.2, we selected two marks to demonstrate the fitting of the Hawkes process. As noted in the Section 7.1, the event process is defined as event types $e \in \{LO, MO, C\}$, levels $l \in \{1, \dots, 5\}$ and for the bid side only. The marks selected are:

- Bid vol MOLOC;
- Bid depth.

The model stability measure $IB = \eta/[(\eta + \vartheta)\vartheta^2]$ defined in (6.51) will play a key role in identifying intra-day time segments that are suitable for the application of the score test and ensure a well calibrated models. Throughout this section, Hawkes process parameter estimates will be presented for the time segments that meet the requirement of $IB < 0.4$.

7.2.1 Modelling a Hawkes process with a one dimensional mark

We extend the Hawkes process by incorporating a mark with a linear boost function. This highlights a number of issues when estimating parameters via the log-likelihood in (4.15). Recall, that both marks exhibit serial dependence, as demonstrated in Section 5.3.5. Whilst we are able to account for this serial dependence by adjusting the score test, the current development of the log-likelihood estimation assumes the marks are i.i.d. We will explore the impact of assuming serial independence in the section that follows.

Simulation results in the presence of a serially dependent mark

The simulation experiment to follow will uncover the challenges that are present when fitting a Hawkes process with a heavy tailed and serially dependent mark. We fit the intensity process in which the marks are linearly boosted

$$h(X; \psi) = 1 + \psi X. \quad (7.1)$$

Theoretically derived moments from a specified parametric distribution in the construction of the boost function will be used. The intensity parameter estimates specified in the simulation are, $\eta = 0.0020$, $\vartheta = 0.7000$, $\alpha = 0.0100$ and the boost parameter is $\psi = 0.5$. The marks in this simulation study $X \in \mathbb{R}$ are serially dependent and distributed with a conditional generalized Pareto distribution $X \sim \text{GPD}(\zeta = 0.05, \delta_i)$. The conditional generalized Pareto distribution scale parameter is defined as $\delta_i = 0.9\delta_{i-1} + \epsilon_i$, where ϵ_i is Gaussian white noise. Refer to Section 5.3.3 for details of the simulation method for a Hawkes process in the case of serially dependent marks. For this study we simulate 1,000 replicates of sample size $n = 1,000$ each. Similarly to Section 4.5, we will assess the quality of the fit visually through boxplots and by estimating the *bias* of the parameter estimates, which is evaluate as a percentage difference of the mean of the parameter estimates across all replicates and their true value.

Table 7.3 presents the true parameters specified in the 1,000 simulations and the mean and parameter estimates from maximizing the log-likelihood. We note that the scale parameter of the generalized Pareto distribution δ_i varies for each simulant, due to the incorporation of serial dependence. We take the mean of the scale parameter used in the simulations.

Table 7.3: Impact of a mark with serial dependence on estimating the parameters for the Hawkes process, whilst incorrectly assuming i.i.d. marks. The sample size is $n = 1,000$, for 1,000 replicates. The mark is conditionally GPD, $X \sim \text{GPD}(\zeta = 0.05, \delta_i)$, with a linear boost and with estimated theoretical moments.

Parameter	True value	Mean Estimate	Max Estimate
Immigration, η	0.0020	0.0020	0.0026
Branching, ϑ	0.7000	0.7034	0.8401
Decay, α	0.0100	0.0101	0.0132
Shape, ζ	0.0500	0.5019	0.6759
Scale, δ	1.8372	1.0699	1.4558
Boost, ψ	0.5000	0.5267	1.1889

Figure 7.6 and Table 7.3 show that the immigration intensity, branching coefficient, and decay parameters are well estimated. The parameter estimate bias from this simulation compares closely with the bias from the simulation study in Section 4.5.2, which is presented in brackets, immigration intensity 2.15% (2.28%), branching coefficient -0.48% (-0.53%) and decay function parameter 1.27% (1.60%). The boost function parameter has a higher bias of 5.34% compared with the intensity parameter estimates, but is consistent with what we observed in Section 4.5.2, (6.62%). We can see that for a univariate model, the presence of serial dependence in the marks does not degrade the reliability of the intensity and boost parameter estimates using the likelihood in (4.15), which falsely, in this case, assumes the marks are i.i.d.

However, the generalized Pareto marginal parameter estimates present significant bias under the incorrect assumption of i.i.d., with the shape parameter estimates being over-estimated by a factor of 9.04, and the scale parameter estimates being under-estimated by -39.08% , which is significantly higher than the (-0.07%) observed in Section 4.5.2. The significant upward bias of the shape parameter is driven by the increase in tail events due to the autocorrelation in the mark time series.

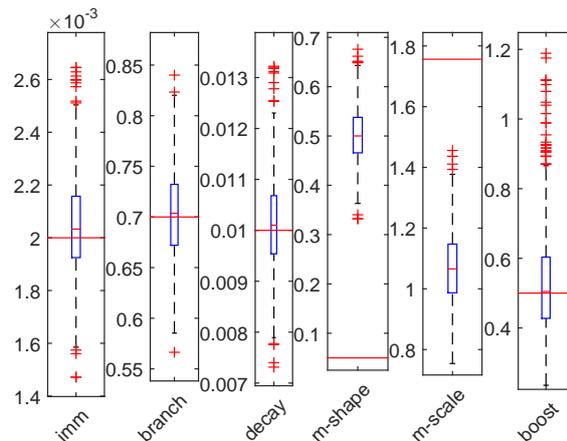


Figure 7.6: Boxplot of parameter estimates for a Hawkes process, with a one dimensional, serially dependent mark, under the incorrect assumption of i.i.d., and with a sample size $n = 1,000$, for 1,000 replicates. The mark is conditionally GPD $X \sim \text{GPD}(\zeta = 0.05, \delta_i)$, with a linear boost and with estimated theoretical moments.

Challenges with bootstrapping methods in the presence of serial dependence

In Section 4.4, we proposed a bootstrapping method for the estimation of the standard errors, which are used to calculate the p-values for the Hawkes process. This method is suitable for a range of Hawkes process with: a constant boost; a boost function with i.i.d. marks; and a boost function with i.i.d. marks and joint dependence. However, it becomes problematic for a Hawkes process with serially dependent marks. This is further confounded in higher dimension when multivariate marks are both serially and jointly dependent.

Whilst it is possible to simulate a mark with serial dependence (Section 5.3.3), the estimation of the model parameters in the log-likelihood (4.15) assumes the marks are i.i.d. The method of the bootstrapping involves using the MLE parameters to specify the simulation parameters and then re-estimating the parameters of the replicates. As noted in Section 7.2.1, the impact of simulating a serially dependent process using estimates of the parameters, and then re-estimating the parameters again with the i.i.d. assumption, results in an inflated shape parameter and a decreased scale parameter estimate.

In Figure 7.7 we present a histogram of a mark which has a shape parameter of $\zeta = 0.5529$. We simulate a Hawkes process with a linear boost and a mark that is conditionally generalized Pareto distributed $X \sim \text{GPD}(\zeta = 0.5529, \delta_i)$. The estimation of the marks distribution under the assumption of i.i.d. results in an inflated shape parameter of $\zeta = 0.8582$.

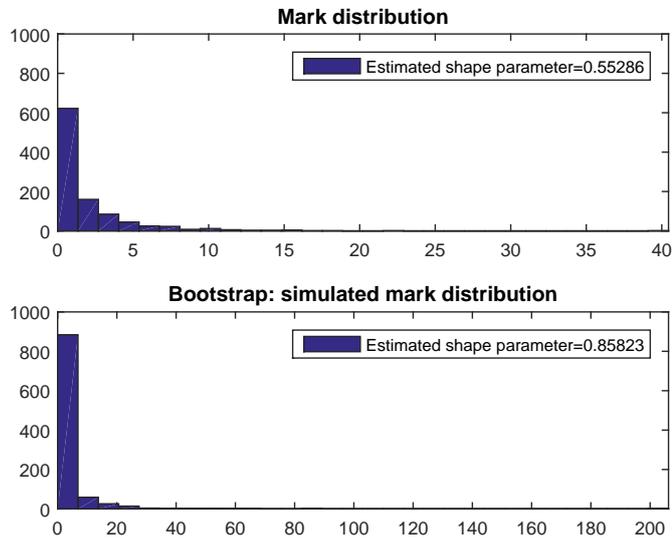


Figure 7.7: Histogram of the mark distribution, with a shape parameter $\zeta = 0.5529$. Histogram of a bootstrapped simulant of a serially dependent mark, with an MLE estimated shape parameter $\zeta = 0.8583$, under the incorrect assumption of i.i.d..

For the case of a Hawkes process with multivariate marks, the standard errors obtained from bootstrapping will not be correct due to the misspecification of the Hawkes process, and therefore cannot be used. One solution is to extend the MLE method for the estimation of a Hawkes process parameters to include the estimation of serially dependent marks.

However to our knowledge the theory required for a Hawkes process with serially dependent marks is not developed and is non-trivial.

Hawkes process with a univariate mark

In this section we consider two models, with the first (*Model 1*) incorporating the mark: Bid depth. This mark is transformed by taking the absolute value of the first level difference. The second model (*Model 2*) incorporates the mark: Bid vol MOLOC. Both marks enter the Hawkes process via a linear boost function and are modelled by a generalized Pareto distribution. Both marks exhibit serial dependence, which cannot be accounted for within the log-likelihood of the Hawkes process, therefore standard errors will not be calculated.

Figure 7.8 presents two barcode plots for each model considered. The top panels shows the time segments where the score test results confirms the mark is significant. We can see that for 46 of the 47 time segments within the single trading date 31-July-2015, the score test evaluated both marks as significant. Ideally we would compare the significance of the mark evaluated by the score test, with the p-value of the boost function via the MLE of the model. However, that is not possible for reasons outlined above.

The bottom panels present the model stability estimate $IB < 0.4$, to identify time segments that result in a well calibrated model, which is the case for all time segments. It is worth noting, that in the study of the Hawkes process with a constant boost, 46 of the 47 time segments resulted in a well calibrated model. The inclusion of a mark has improved the calibration of the model for this particular trading day.

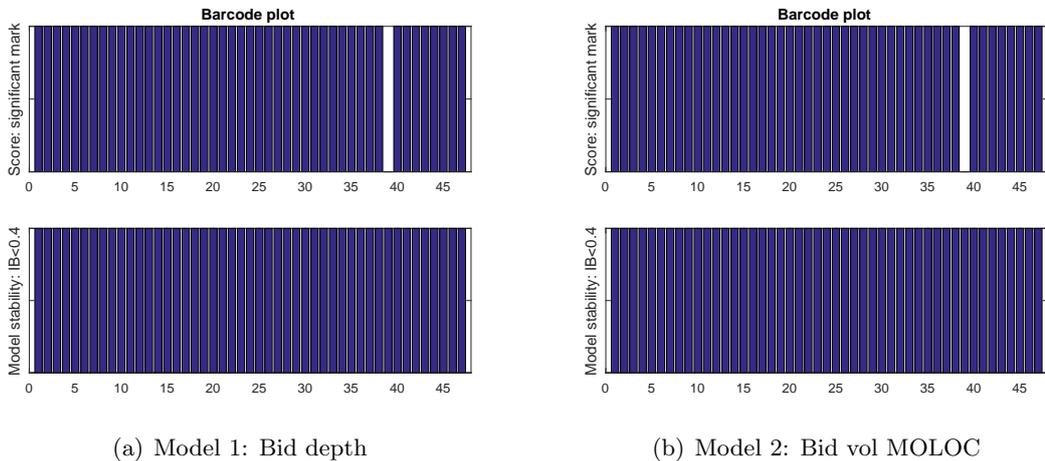


Figure 7.8: Barcode plot for each time segment, with an indicator $I = 1$ when the score test is significant, and an indicator $I = 1$ when the Hawkes process, with a linear boost and a univariate mark, is well calibrated ($IB < 0.4$). The asset is SILVER, on 31-July-2015, with event types $e \in \{LO, MO, C\}$, levels $l \in \{1, \dots, 5\}$ and bid side only.

Figures 7.9 and 7.10 present boxplots of the parameter estimates for *Model 1* and *Model 2*. The models are estimated across all time segments for a single trading day. The intensity parameters, $\theta = (\eta, \vartheta, \alpha)$ are consistent for both models as expected. The mean of the *Model 1* boost parameter is $\bar{\psi} = 0.4195$, however for later time segments there

is significant fluctuation in the parameter estimates, with one time segment resulting in an estimate of $\hat{\psi} = 3.5916$ (Figure 7.9). Comparing this with *Model 2*, the mean boost parameter estimate is $\bar{\psi} = 0.0654$, with a maximum of $\hat{\psi} = 0.1870$ across all time segments (Figure 7.10). The second mark Bid vol MOLOC has far less impact on boosting the intensity process than the first mark Bid depth.

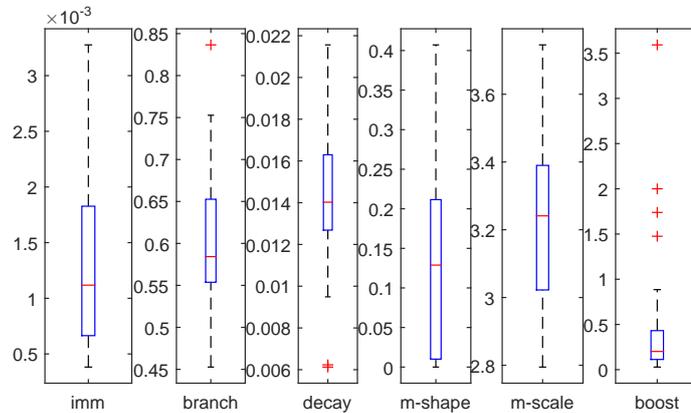


Figure 7.9: *Model 1*: Bid depth. Boxplot of the parameter estimates for a Hawkes process, with a linear boost and a single mark. The models are estimated across 47 time segments, for SILVER bid side, on 31-July-2015, with event types $e \in \{LO, MO, C\}$ and levels $l \in \{1, \dots, 5\}$.

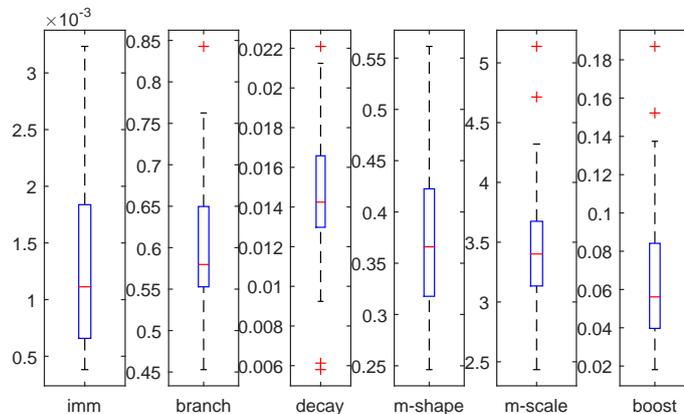


Figure 7.10: *Model 2*: Bid vol MOLOC. Boxplot of the parameter estimates for a Hawkes process, with a linear boost and a single mark. The models are estimated across 47 time segments, for SILVER bid side, on 31-July-2015, with event types $e \in \{LO, MO, C\}$ and levels $l \in \{1, \dots, 5\}$.

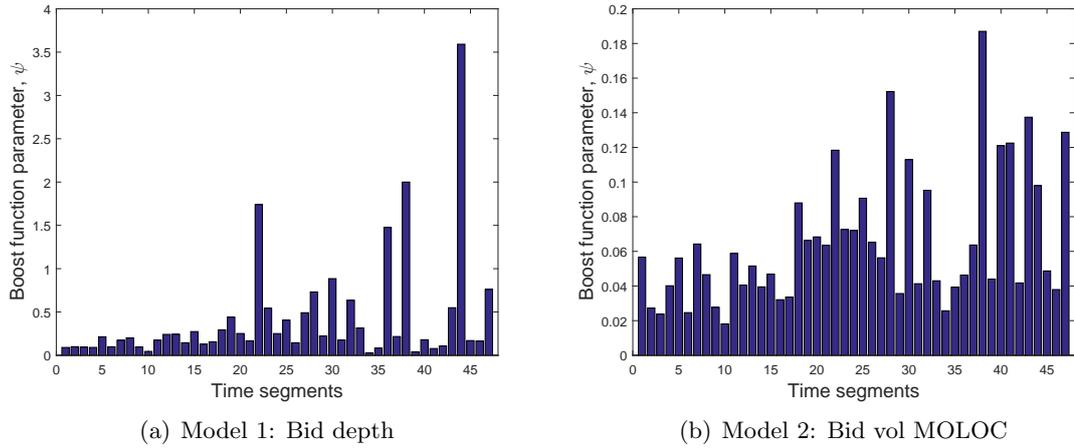


Figure 7.11: Boost function parameter estimates across all 47 time intervals for a Hawkes process, with a linear boost, a single mark, for SILVER bid side, on 31-July-2015, with event types $e \in \{LO, MO, C\}$ and levels $l \in \{1, \dots, 5\}$.

For the purpose of illustration of the Hawkes process and comparing *Model 1* and *Model 2* in greater detail, we select the time segment which is presented in Figure 7.12 as the blue vertical bar. The inserted chart, which shows the counts of events across evenly spaced seconds, are stationary event times.

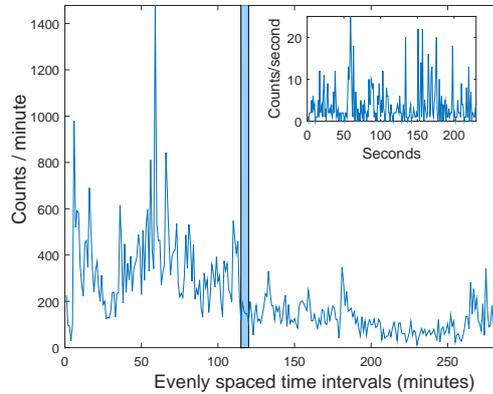


Figure 7.12: Counts of events across even 1 minute time intervals. Insert plots are counts across even 1 second time intervals. Counts are for SILVER bid side, on trading day 31-July-2015, event types $e \in \{LO, MO, C\}$ and levels $l \in \{1, \dots, 5\}$. The blue bar represents the time segment to be modelled by a Hawkes process, with linear boost and a single mark.

The score test statistic for the mark in *Model 1*, Bid depth is $S = 32.8170$, with a p-value of 0. The score statistic for the second mark in *Model 2*, Bid vol MOLOC is $S = 42.0690$, with a p-value of 0. For both marks we strongly reject the null hypothesis that the mark has no impact on the intensity process.

Table 7.4 presents the parameter estimates for *Model 1* and *Model 2* for the same time segment for SILVER on the trading date 31-July-2015. Both models exhibit similar intensity parameters $\theta = (\eta, \vartheta, \alpha)$. For a Hawkes process with a linear boost function and a single mark, we require only the first moments to exist $\zeta < 1$, which is the case

for both marks. However, it is noted that the shape parameter estimates are likely to be over-estimated due to the serial dependence within both marks.

Table 7.4: *Model 1* and *Model 2* parameter estimates for a Hawkes process with a linear boost function, estimated for a single time segment (6 minutes), for SILVER on 31-July-2015, with event types $e \in \{LO, MO, C\}$, levels $l \in \{1, \dots, 5\}$ and for the bid side only.

Parameter	Model 1: Bid depth	Model 2: Bid vol MOLOC
Immigration, η	0.0010	0.0010
Branching, ϑ	0.5872	0.5792
Decay, α	0.0147	0.0158
Mark-shape, ζ	0.1838	0.4958
Mark-scale, δ	3.1196	3.2173
Boost, ψ	0.2517	0.0683

Figure 7.13 shows the intensity function as previously defined in (4.14) and using the parameter estimates in Table 7.4. The intensity plots for both models show an estimated Hawkes process with a constant boost function (red line) and the Hawkes process with a boosted intensity (blue line). In both cases we can see that when a mark occurs, the impact on the intensity increases at that event time. The impact on the intensity process appears to be more pronounced for *Model 1* due to the higher estimated boost function parameter (Figure 7.13(a)).

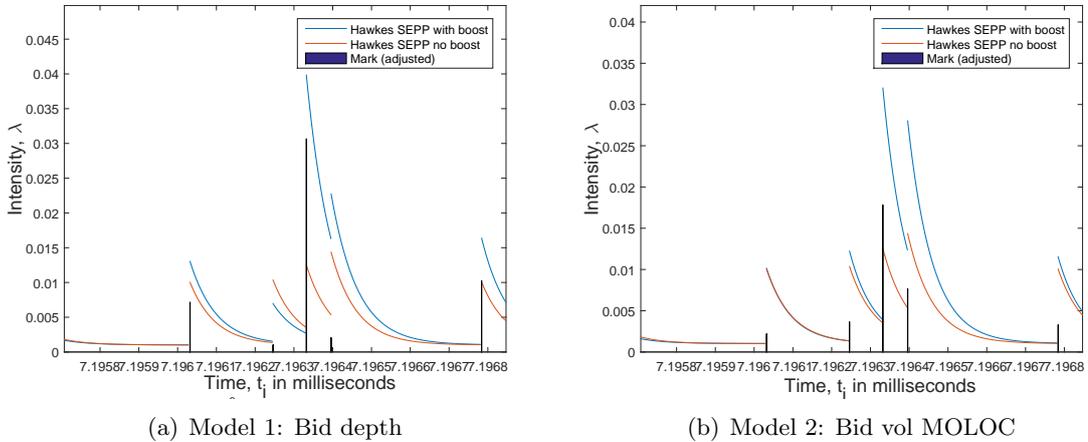


Figure 7.13: A subset of the intensity function, with decay versus time, for SILVER. The subset represents 9 seconds of trading during 31-July-2015 and is modelled by a Hawkes process, with a linear boost function (blue) and a Hawkes process, with a constant boost function (red). Model 1 has a mark: Bid depth and *Model 2* has a mark: Bid vol MOLOC.

A graphical representation of both models goodness of fit is a plot of the residual process, as defined in Section, 4.2.3. Figures 7.14(a) and 7.15(a) show the inter-arrival times, with the empirical quantiles plotted on the vertical axis. The theoretical quantiles are plotted on the horizontal axis. Since the residual process should be Poisson with unit intensity, we expect the durations to be exponentially distributed with parameter one. We can see that both models provide a reasonable fit, but with some deviation in the tails.

Figures 7.14(b) and 7.15(b) show the plot of points of the counting function versus the normalized transformed times. Based on the properties of the Poisson distribution, we

expect that the location of events, conditioned on the number of events to be uniformly distributed in the observation interval (Liniger, 2009). There are two confidence bands of 95% and 99% level for the Poisson null hypothesis and the counting process does not exceed these confidence bands.

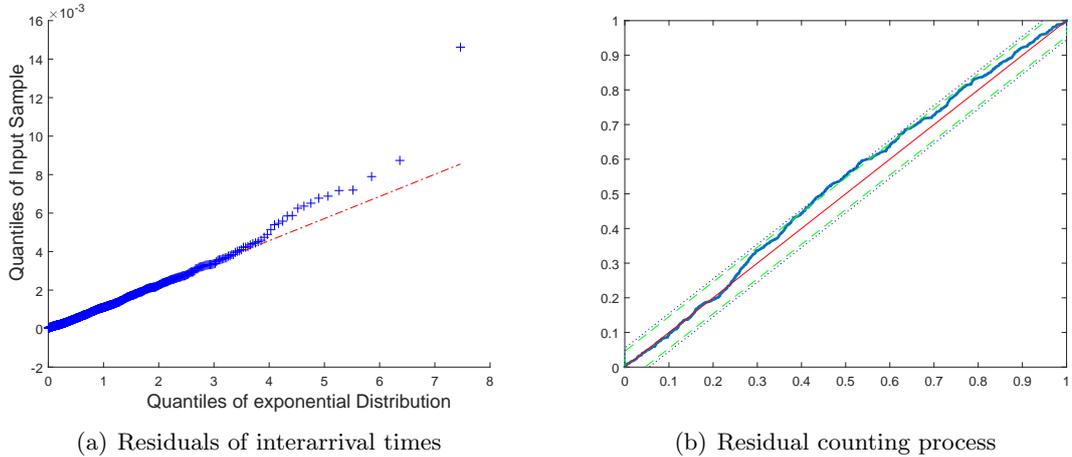


Figure 7.14: QQ-plots of the residual inter-arrival times and the residual counting process, for the Hawkes process, with a linear boost and univariate mark, Bid depth. The asset is SILVER on the trading date 31-July-2015, with event types $e \in \{LO, MO, C\}$, levels $l \in \{1, \dots, 5\}$ and bid side only. The two confidence bands for the Poisson null hypothesis are, 95% (black-dashed) and 99% (green-dashed).

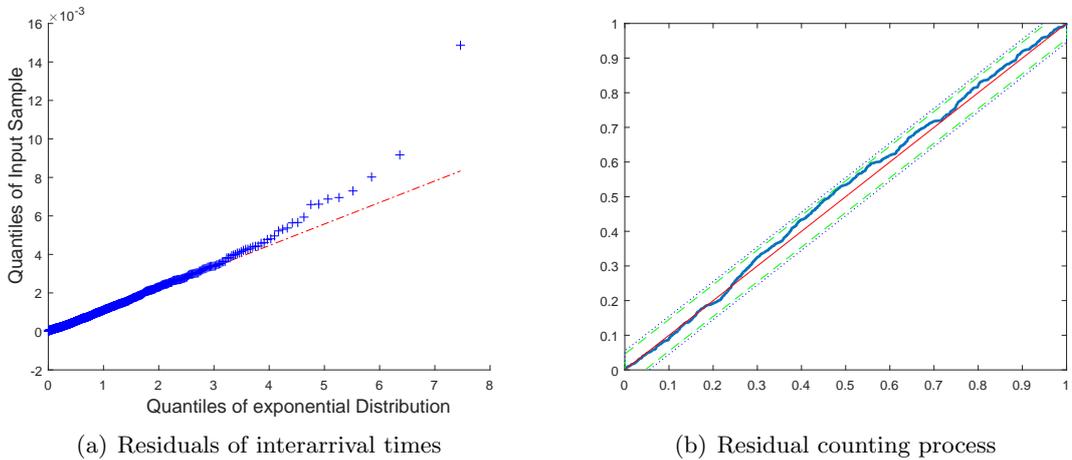


Figure 7.15: QQ-plots of the residual inter-arrival times and the residual counting process, for the Hawkes process, with a linear boost and univariate mark, Bid vol MOLOC. The asset is SILVER on the trading date 31-July-2015, with event types $e \in \{LO, MO, C\}$, levels $l \in \{1, \dots, 5\}$ and bid side only. The two confidence bands for the Poisson null hypothesis are, 95% (black-dashed) and 99% (green-dashed).

7.2.2 Modelling a Hawkes process with two dimensional marks

In higher dimensions, consideration of both serial dependence and joint dependence when fitting the Hawkes process is critical. Recall for the log-likelihood method, where the marks are assumed to be i.i.d. but jointly dependent, we can adjust the boost function normalization to account for the joint dependence (Section 4.2.1). Recall that to normalise the boost with joint dependence requires the second moment to exist. The marks exhibit heavy tails and are best modelled by a generalized Pareto distribution. For second moments to exist we require that the shape parameter of the generalized Pareto distribution to be $\zeta < 0.5$.

Simulation results in the presence of serial dependence and joint dependence

The simulation studies that follow, extend those in Section 7.2.1, by highlighting the additional challenges of parameter estimation via the log-likelihood function when the marks are serially and jointly dependent and exhibit heavy tailed features. The intensity parameter estimates specified in the simulation are, $\eta = 0.0020$, $\vartheta = 0.7000$, $\alpha = 0.0100$. We fit the intensity process linear boost function in (7.1), with boost function parameters $\psi_1 = 0.5$ and $\psi_2 = 0.5$. The marks ($X_i \in \mathbb{R}^2$) in this simulation study are serially dependent and distributed with a conditional generalized Pareto distribution $X_i \sim \text{GPD}(\zeta = 0.05, \delta_i)$. The conditional generalized Pareto distribution scale parameter is defined as $\delta_i = 0.9\delta_{i-1} + \epsilon_i$. The joint dependence is modelled via a Gaussian copula model with a Spearman's rank correlation of $\rho_s = 0.5$. For this study we simulate 1,000 replicates of sample size $n = 1,000$ each.

In Section 4.5.5 we discovered that the *approximate likelihood method by decoupling the marks parameters*, which uses empirical moments for the normalization of the boost function, rather than theoretical moments, led to a lower bias in parameter estimates for the Hawkes process with bivariate and jointly dependent generalized Pareto distributed i.i.d. marks. A summary of the key results from that section follows.

- The estimation of the model parameters in the log-likelihood (4.15), when the true shape parameter is close to 0.5 (close to breaching the assumption of second moments existing) resulted in unreliable parameter estimates (Table 4.6). We note that 50% of the replicates in the simulation study that follows, have a shape parameter estimate that is higher than the required second moments (Table 7.5).
- Imposing an upper bound of 0.5 on the optimization procedure improved the parameter estimates, however the boost parameter estimates were greatly inflated and there was a significant downward bias in the branching coefficient (Table 4.6).
- We finally considered replacing theoretical moments with empirical moments in the calculation of the normalization of the boost function and this produced reliable parameter estimates (Table 4.6).
- Following on from that, and considering different copula models with a low shape parameter $\zeta = 0.05$, we discovered that using empirical moments for the normalization

of the boost function gave a significant reduction in the bias for the boost parameter estimates and the immigration intensity parameters across all models (comparing Table 4.7 with Table 4.8).

Based on our findings in Section 4.5.5, we proceed by replacing theoretical moments with empirical moments for the simulations and model fitting that follows.

Table 7.5: Impact of bivariate marks with serial and joint dependence on estimating the parameters for the Hawkes process, whilst incorrectly assuming i.i.d. marks. The sample size is $n = 1,000$, for 1,000 replicates. The marks are conditionally GPD, $X_i \sim \text{GPD}(\zeta = 0.05, \delta_i)$ and with a linear boost function. The joint dependence is modelled by a Gaussian copula with $\rho_s = 0.5$.

Parameter	True value	Mean Estimate	Max Estimate	Percentage $\zeta < 0.5$
Immigration, η	0.0020	0.0020	0.0025	
Branching, ϑ	0.7000	0.6991	0.8977	
Decay, α	0.0100	0.0101	0.0139	
Mark 1-shape, ζ	0.0500	0.5014	0.6807	48.60%
Mark 1-scale, δ	1.8254	1.0856	1.6553	
Mark 2-shape, ζ	0.0500	0.4996	0.6904	51.00%
Mark 1-scale, δ	1.8306	1.0791	1.5222	
Boost 1, ψ	0.5000	0.5244	2.4116	
Boost 2, ψ	0.5000	0.5300	1.2869	

Table 7.5 and Figure 7.16 show that the incorporation of an additional mark and joint dependency does not appear to impact the estimation of the intensity and boost parameters. The bias in the intensity and boost parameter estimates is, immigration intensity 0.91%, branching coefficient -0.12% , decay function parameter 0.65% (all less than 1%), boost 1 parameter 4.88% and boost 2 parameter 6.01%. These results are in-line with the bias observed in Section 4.5.5 for the Hawkes process with bivariate jointly coupled i.i.d. marks with a generalized Pareto distribution.

The shape parameters again exhibit an upward bias, increasing the shape by a factor of 9.02 and 8.99 and a downward bias in the scale parameters of -40.13% and -39.84% , respectively. Increasing the mark dimension to $d = 2$ and including copula dependence, does not degrade the reliability of the intensity and boost function parameter estimates. The bias in the marginal parameter estimates remains high and consistent with what we observed in the one dimensional marked Hawkes process.

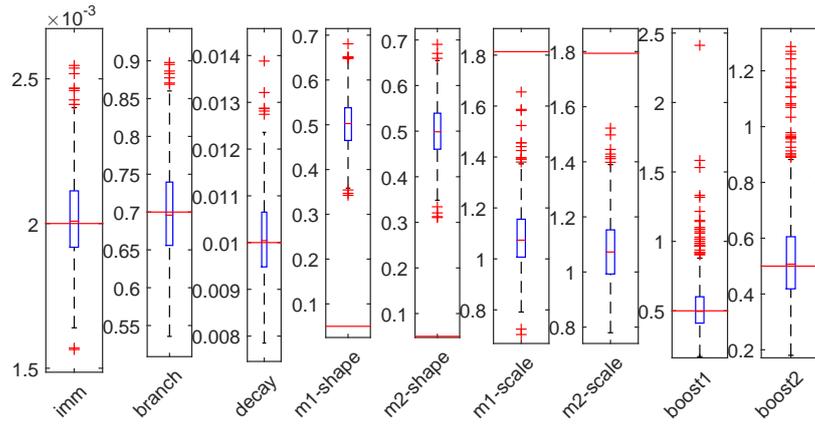


Figure 7.16: Boxplot of parameter estimates for a Hawkes process, with bivariate serially and jointly dependent marks $X_i \in \mathbb{R}^2$, under the incorrect assumption of i.i.d., and with a sample size $n = 1,000$, for 1,000 replicates. The marks are conditionally GPD $X_i \sim \text{GPD}(\zeta = 0.05, \delta_i)$ and with a linear boost function. The joint dependence is modelled by a Gaussian copula with $\rho_s = 0.5$.

As we highlighted in Section 7.2.1 in the one dimensional mark case, it is not possible to use a bootstrapping technique to estimate standard errors for the evaluation of p-values for the Hawkes process when the mark is serially dependent. The incorporation of an additional mark and joint dependence between the marks, further amplifies these challenges due to the requirement of second moments.

Hawkes process with bivariate and jointly dependent marks

We extend the study in Section 7.2.1 by modelling a Hawkes process with bivariate marks, Bid depth and Bid vol MOLOC, and jointly coupled via a Gaussian copula. As we've highlight, both marks are serially dependent.

In Figure 7.17 (top panel) we present score test results for the bivariate marks and see that for all time segments, the score test confirms that the bivariate marks are significant. The bottom panel in Figure 7.17 shows that all time segments result in a well calibrated model, as defined by the model stability measure $IB < 0.4$.

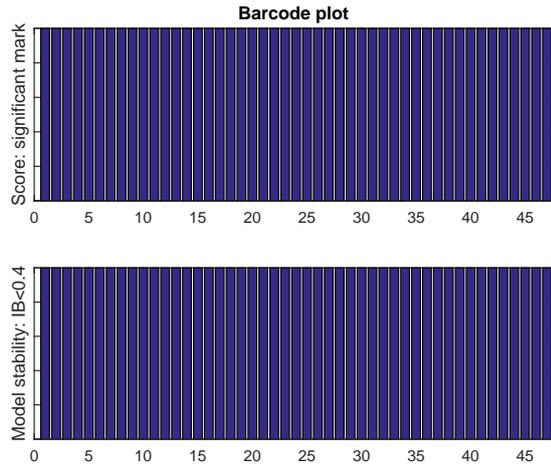


Figure 7.17: Barcode plot for each time segment, with an indicator $I = 1$ when the score test is significant, and an indicator $I = 1$ when the Hawkes process, with a linear boost and bivariate marks is well calibrated ($IB < 0.4$). The asset is SILVER, on 31-July-2015, with event types $e \in \{LO, MO, C\}$, levels $l \in \{1, \dots, 5\}$ and bid side only.

We present the parameter estimates for all estimated models across the time segments in Figure 7.18. The parameters are well estimated across all time segments, with estimated intensity parameters being similar to the estimated intensity parameters of the univariate models considered in Section 7.2.1.

The mean boost parameter estimates for the first mark, Bid depth is $\bar{\psi}_1 = 0.3627$, which is slightly lower than *Model 1*, $\bar{\psi} = 0.4195$. However, there is a significant down-weight in the second mark, Bid vol MOLOC boost parameter $\bar{\psi}_2 = 0.0087$, compared with *Model 2*, $\bar{\psi} = 0.0654$.

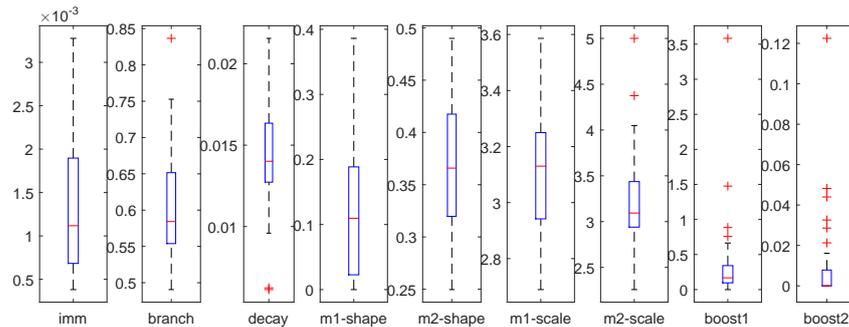


Figure 7.18: Boxplot of the parameter estimates for a Hawkes process, with a linear boost and bivariate marks. The models are estimated across 47 time segments, for SILVER on 31-July-2015, with event types $e \in \{LO, MO, C\}$, levels $l \in \{1, \dots, 5\}$ and bid side only.

When Bid depth is combined multiplicatively in a boost function with the mark, Bid vol MOLOC the second mark has very little impact on boosting the intensity function for many of the time segments (Figure 7.19(a)), which is consistent with the mean of the estimated boost parameter down-weights we observed in Figure 7.18.

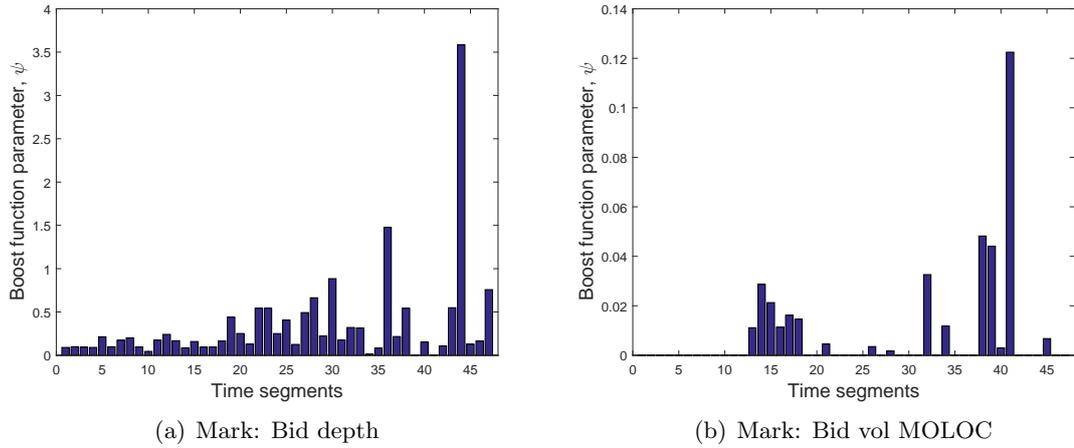


Figure 7.19: Boost function parameter estimates across all 47 time intervals for a Hawkes process, with a linear boost and bivariate marks, for SILVER bid side, on 31-July-2015, with event types $e \in \{LO, MO, C\}$ and levels $l \in \{1, \dots, 5\}$.

We model two time segments, *Time segment 1* and *Time segment 2* presented in Figure 7.20 as the blue vertical bars to study the change in parameter estimates and to evaluate the goodness-of-fit for the Hawkes process with jointly dependent bivariate marks. These particular time segments are chosen, because they exhibit different levels of clustering behaviour, with *Time segment 1* (Figure 7.20(a) insert chart) displaying far greater clustering than *Time segment 2* (Figure 7.20(a) insert chart).

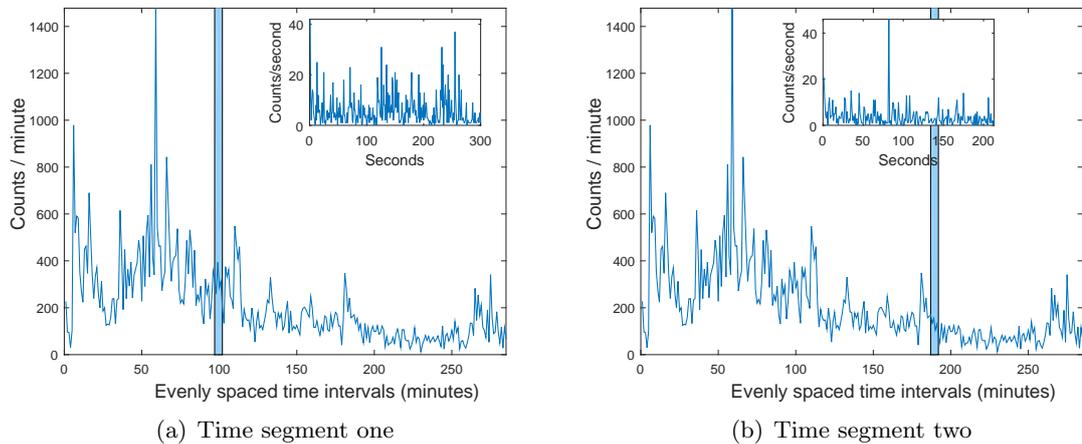


Figure 7.20: Counts of events across even time intervals of one minute. Insert plots are counts of events across even time intervals of one second. The counts of events are for SILVER on a single trading date 31-July-2015. The event process is defined as event types $e \in \{LO, MO, C\}$, levels $l \in \{1, \dots, 5\}$ and bid side only. The blue bars represents the two time segment that will be modelled by a Hawkes process, with linear boost and bivariate marks.

The score test statistic for the bivariate marks is $S = 36.63$, with a p-value of $1.11E-08$ for *Time segment 1*, and $S = 45.04$, with a p-value of $1.65E-10$ for *Time segment 2*. We strongly reject the null hypothesis of the bivariate marks having no impact on the intensity function for both time segments.

The increased clustering behaviour of *Time segment 1* (Figure 7.20(a)) compared to *Time segment 2*, is captured well by the Hawkes process and reflected in the parameter estimates for both time segments in Table 7.6. We see a drop in both immigration intensity and branching coefficient for *Time segment 2*, in addition the boost parameter estimates for both marks have decreased. Consistent with what we observed across all time segments, the parameter estimate for the boost function in *Time segment 1* for the first mark, Bid depth has a greater contribution in boosting the intensity than the second mark, Bid vol MOLOC. The first mark appears to be capturing the majority of the increased clustering behaviour, with the boost parameter of the second mark not changing substantially from *Time segment 1* to *Time segment 2*.

Table 7.6: *Time segment 1* and *Time segment 2* parameter estimates for a Hawkes process with a linear boost function and bivariate marks that are jointly dependent, for SILVER on 31-July-2015, with event types $e \in \{LO, MO, C\}$, levels $l \in \{1, \dots, 5\}$ and for the bid side only.

Parameter	Time segment 1	Time segment 2
Immigration, η	0.0017	0.0006
Branching, ϑ	0.6514	0.6324
Decay, α	0.0119	0.0170
Mark 1-shape, ζ	6.4088E-08	0.2639
Mark 1-scale, δ	3.0559	2.9030
Mark 2-shape, ζ	0.3729	0.4253
Mark 1-scale, δ	3.3586	2.6533
Boost 1, ψ	0.0960	0.0120
Boost 2, ψ	0.0162	0.0118

The parameters specified in Table 7.6 are used to plot the intensity functions for *Time segment 1* (Figure 7.21(a)) and *Time segment 2* (Figure 7.21(b)) using (4.14). The first time segment represents 100 milliseconds and the second time segments represents one second. The intensity plots for each time segments shows the estimated Hawkes process with a constant boost (red line) and the Hawkes process with a boosted intensity from bivariate multiplicative marks (blue line). Whilst the charts reflect very small time frames within the time segment, it is apparent that the impact of the boost on the intensity function, compared with a Hawkes process with a constant boost, is significant for both time segments, with an even greater shift up of the intensity for *Time segment 1*. This highlights the importance of including significant marks into the Hawkes process to accurately capture the intensity of event arrivals into the LOB.

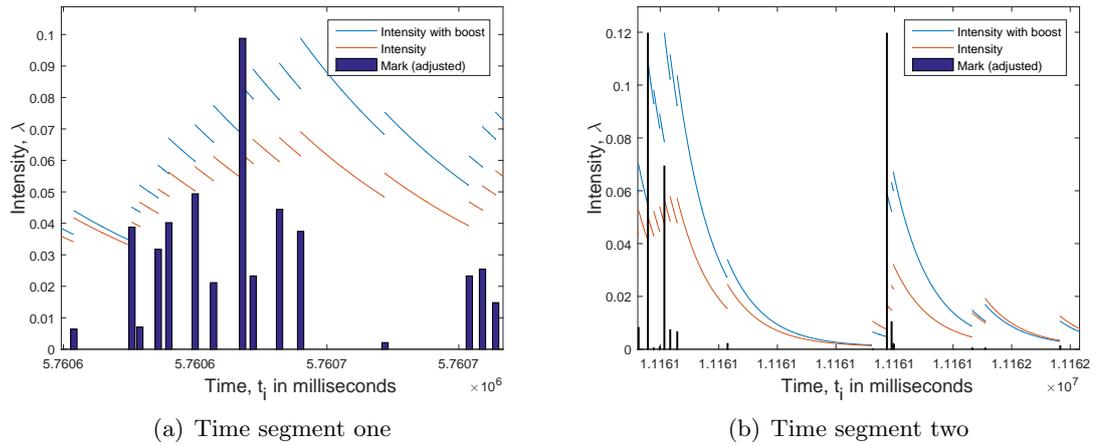


Figure 7.21: A subset of the intensity function, with decay versus time, for SILVER. The two time segments represent 100 milliseconds and one second on 31-July-2015. The two models are: a Hawkes process with linearly boost bivariate marks, Bid depth and Bid vol MOLOC (blue); and a Hawkes process with a constant boost (red).

Figures 7.22 and 7.23 present the goodness of fit plots of the model for both time segments. From the residual plots we can see that the model provides a good fit in both cases.

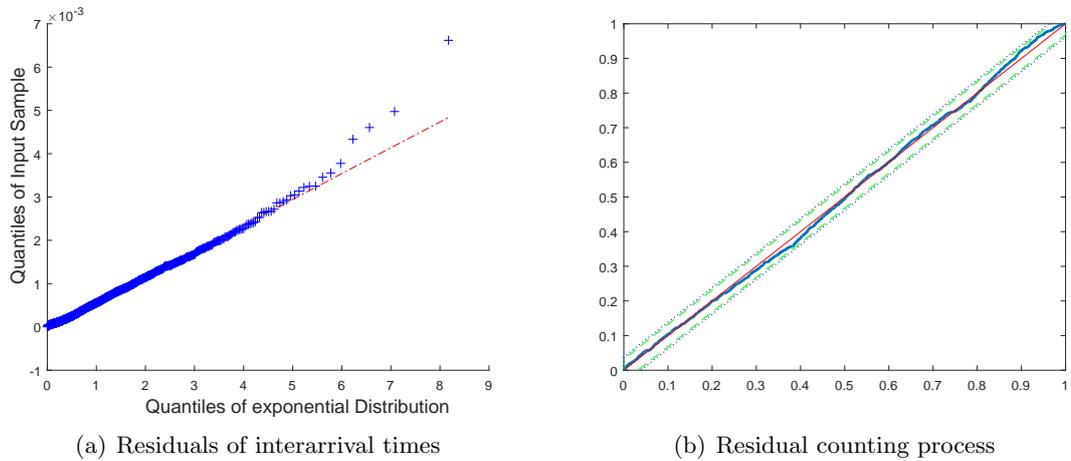


Figure 7.22: QQ-plots of the residual inter-arrival times and the residual counting process, for *Time segment one*. The Hawkes process has a linear boost and bivariate marks, Bid depth and Bid vol MOLOC. The asset is SILVER on the trading day 31-July-2015, with event types $e \in \{LO, MO, C\}$ and levels $l \in \{1, \dots, 5\}$. The two confidence bands for the Poisson null hypothesis are: 95% (black-dashed) and 99% (green-dashed).

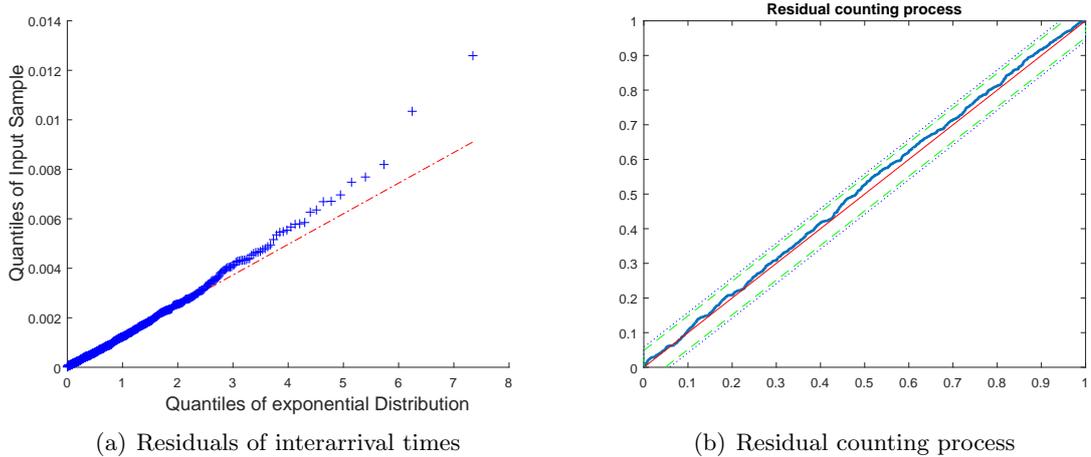


Figure 7.23: QQ-plots of the residual inter-arrival times and the residual counting process, for *Time segment two*. The Hawkes process has a linear boost and bivariate marks, Bid depth and Bid vol MOLOC. The asset is SILVER on the trading day 31-July-2015, with event types $e \in \{LO, MO, C\}$ and levels $l \in \{1, \dots, 5\}$. The two confidence bands for the Poisson null hypothesis are: 95% (black-dashed) and 99% (green-dashed).

7.2.3 A decoupled approximate likelihood method

In light of the difficulty of correctly identifying the appropriate parametric distributions for a high dimensional mark vector and the added complexity of serial dependence and joint dependence, we propose a **decoupled approximate likelihood method** to address these challenges. This method follows on from the development of the theory and findings of the score test. The method will use empirical moments, rather than theoretically derived moments from a specified parametric distribution in the construction of the boost function (as per Section 7.2.2). This method extends upon the *approximate likelihood method by decoupling the marks parameters*, by not requiring the parametric specification of the marks. The decoupled approximate likelihood method will include the intensity component of the log-likelihood, providing estimates for the intensity parameters $\theta = (\eta, \vartheta, \alpha)$ and the parameters for the scalar boost function ψ .

Defining the decoupled approximate likelihood method

Recall from Chapter 4, informally we define the log-likelihood as (4.15), but with the third term replaced by the conditional density $f(\mathbf{x}_1, \dots, \mathbf{x}_n | t_1, \dots, t_n; \phi)$ of the marks given the event times, such that

$$\begin{aligned}
 l_g(\nu) = & \int_{[0, T] \times \mathbb{X}} \ln \lambda_g(t; \nu) N_g(dt \times d\mathbf{x}) - \Lambda_g(T; \nu) \\
 & + \ln f(\mathbf{x}_1, \dots, \mathbf{x}_n | t_1, \dots, t_n; \phi)
 \end{aligned} \tag{7.2}$$

where ϕ represents all the parameters of the joint conditional distribution including any parameters needed to model serial dependence.

Regarding joint estimation of the intensity parameters and the marks process param-

eters, the likelihood informally proposed in (7.2) could be maximised over all parameters. When the auto-covariance structure of the marks is assumed to be that of a discrete time process, conditioning on the event times to obtain $f(\mathbf{x}_1, \dots, \mathbf{x}_n | t_1, \dots, t_n; \phi)$, is not required. This is due to the actual time differences $t_j - t_i$ between all pairs of event times not being needed, and the index of time order is all that is required.

For a number of examples of potential marks we have encountered in application to the limit order book, the marks have serial dependence, heavy tails and copula dependence between elements in the mark vector at each time. These are complex time series models, and as such there is very little methodology for obtaining maximum likelihood estimates for the parameters ϕ by maximising $\ln f(\mathbf{x}_1, \dots, \mathbf{x}_n | t_1, \dots, t_n; \phi)$. Computational difficulties compound this further when the joint maximization of $l_g(\nu)$ in (7.2) is attempted. As per the pragmatic approach employed by the *approximate likelihood method by decoupling the marks parameters* method, we use empirical moments to normalize the boost function required to define the process intensity $\lambda_g(t)$ appearing in the first two components of the log-likelihood. That is, we use $\tilde{g}(\mathbf{x}; \psi) = \frac{h(\mathbf{x}; \psi)}{\bar{h}(\mathbf{x}; \psi)}$, where $\bar{h}(\mathbf{x}; \psi)$ is the function h evaluated using empirical moments in place of theoretical moments. This has the effect of ‘decoupling’ the parameters ϕ from the parameters, θ and ψ leading to a decoupled approximate likelihood

$$l_g(\theta, \psi) = \int_{[0, T] \times \mathbb{X}} \ln \lambda_{\tilde{g}}(t; \nu) N_g(dt \times d\mathbf{x}) - \Lambda_{\tilde{g}}(T; \nu) \quad (7.3)$$

for (θ, ψ) .

Simulation study

The aim of the simulation study is to assess whether we can model a Hawkes process with heavy tailed, serially and joint dependent marks and still achieve reliable parameter estimates, without having to specify the parametric distribution of the marks.

The case study that follows, will proceed in a similar fashion to Section 7.2.2, simulating a Hawkes process with bivariate, serially and joint dependent marks. The intensity parameter estimates specified in the simulation are, $\eta = 0.0020$, $\vartheta = 0.7000$, $\alpha = 0.0100$. We fit the intensity process with a linear boost function in (7.1), with boost function parameters $\psi_1 = 0.5$ and $\psi_2 = 0.5$. The marks $X_i \in \mathbb{R}^2$ in this simulation study are serially dependent with a conditional generalized Pareto distribution $X_i \sim \text{GPD}(\zeta = 0.05, \delta_i)$. The conditional generalized Pareto distribution scale parameter is defined as $\delta_i = 0.9\delta_{i-1} + \epsilon_i$. The joint dependence is specified via a Gaussian copula with $\rho_s = 0.5$. For this study we simulate 1,000 replicates of sample size $n = 1,000$ each.

We will only present the intensity parameters and the boost parameters that we obtain from the *approximate likelihood method by decoupling the marks parameters* in (4.15) $(\hat{\theta}^{(1)}, \hat{\psi}_1^{(1)}, \hat{\psi}_2^{(1)})$, as a comparison to those estimated via the proposed decoupled approximate likelihood method in (7.3) $(\hat{\theta}^{(2)}, \hat{\psi}_1^{(2)}, \hat{\psi}_2^{(2)})$. Within each chart, the red horizontal line corresponds to the true value used in the simulation.

Figure 7.24 shows the parameter estimates for the two likelihood methods. The percentage error between the parameter estimates for the two methods is $(\hat{\theta}^{(1)} - \hat{\theta}^{(2)})/\theta$, $(\hat{\psi}_1^{(1)} - \hat{\psi}_1^{(2)})/\psi$ and $(\hat{\psi}_2^{(1)} - \hat{\psi}_2^{(2)})/\psi$. Both methods produce almost identical results across the 1,000 replicates, with the mean percentage error of: immigration intensity, $-9.7826E - 05\%$; branching coefficient, $2.2658E - 05\%$; decay, $-1.3675E - 05\%$; boost 1, $-2.1824E - 04\%$; and boost 2, $-1.5972E - 04\%$.

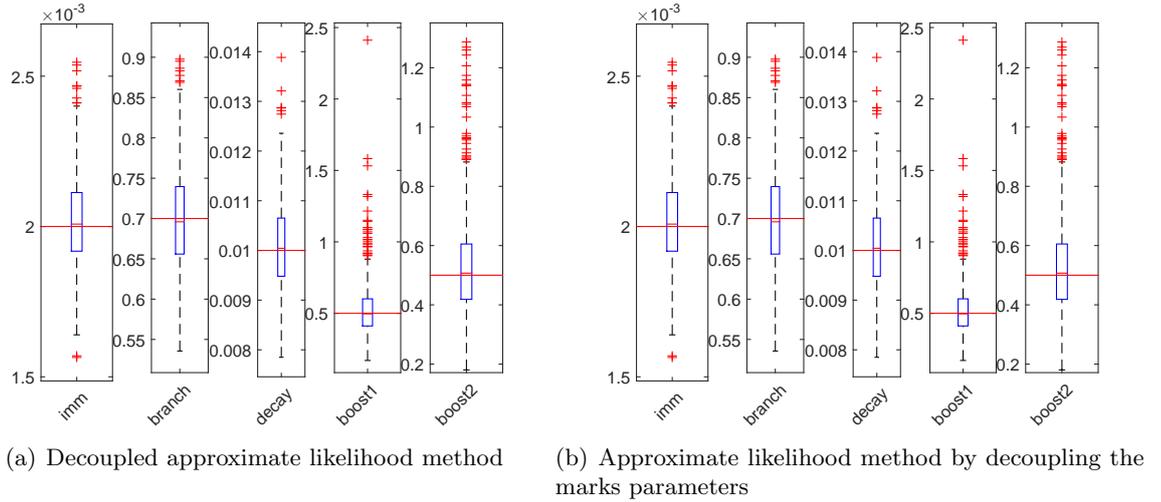


Figure 7.24: Boxplot of parameter estimates from 1,000 simulations, with bivariate serially dependent GPD marks that are jointly dependent. The parameters are estimated via: the decoupled approximate likelihood method; and the *approximate likelihood method by decoupling the marks parameters*, with specifications of a jointly dependent GPD marks, evaluated with empirical moments.

In practice one would assess the distributional features of the data before applying a parametric distribution to the marks. However, it is not always clear what the correct distributional form is due to the complexity of the data, as discovered in Chapter 5. In addition, when assessing hundreds of potential marks, automated methods that would be used in practice may make an incorrect assumption of the distributional properties of some marks.

Three different methods have been used to estimate the parameters of the Hawkes process. The first was the log likelihood method using estimated theoretical moments. In Chapter 5 we discovered that when the marks are jointly dependent, this method results in a higher bias in the parameter estimates. The second method considered was the *approximate likelihood method by decoupling the marks parameters*, which replaced the estimated theoretical moments with empirical moments. This reduced the bias of the boost parameter estimates and the immigration intensity. When using this method to estimate the parameters of the Hawkes process with marks that exhibit serial dependence, thus incorrectly assuming i.i.d., this method still resulted in reliable intensity and boost parameter estimates, despite the marginal parameter estimates exhibiting high bias (Section 7.2.2). The third method, the decoupled approximate likelihood method, also uses empirical moments, but does not require the specification of the parametric distribution

of the marks.

In Figure 7.25 we present the parameter estimates for the simulated data used in the previous example, but modelling the replicates with a Hawkes process with bivariate marks coupled via a Gaussian copula and marginally exponentially distributed, when in fact the marks are heavy tailed. It is quickly apparent that the intensity and boost parameter estimates are unreliable when the distribution specified for the marks is a very poor approximation. However, the decoupled approximate likelihood method does not require the distribution of the marks to be specified.

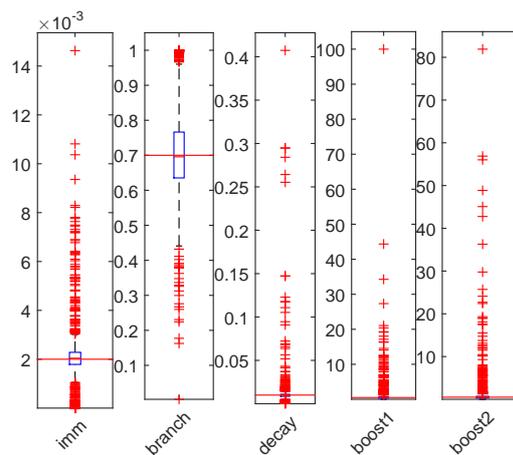


Figure 7.25: Boxplot of parameter estimates (misspecified) from 1,000 simulations, with bivariate serially dependent GPD marks that are jointly dependent. The replicates are modelled using a Hawkes process, with jointly dependent exponentially distributed marks, evaluated with empirical moments.

The second method will be insufficient in practice, despite the decoupling of the marks parameters from the intensity and boost parameters. The specification of a marks distribution, when far from approximating the distributional form of the mark (specifying a light tailed distribution for the marks, when in fact they exhibit heavy tails) will create biases in the intensity and boost parameter estimates. Therefore, the decoupled approximate likelihood method is the most appropriate for estimating the parameters of a Hawkes process with multivariate marks, for the modelling LOB data.

Application of the decoupled approximate likelihood method to real data

We apply the two likelihood methods discussed in this section to estimate the parameters of a Hawkes process with bivariate marks, Bid depth and Bid vol MOLOC for SILVER, across 47 intra-day time segments, on the trading date 31-July-2015. The models are equivalent to what was studied in Section 7.2.2. The three cases we consider are:

1. Parameters estimated by the decoupled approximate likelihood method;
2. Parameters estimated by the *approximate likelihood method by decoupling the marks parameters*, with the marks specification being a generalized Pareto distribution and joint dependence modelled by a Gaussian copula;

- Parameters estimated by the *approximate likelihood method by decoupling the marks parameters*. The bivariate marks have an exponential distribution specification, when in fact the marks are heavy tailed. The coupling of the marks is via a Gaussian copula.

Figure 7.26 presents boxplots of the parameter estimates across all time segments for the decoupled approximate likelihood method and the *approximate likelihood method by decoupling the marks parameters*, with a marginal generalized Pareto distribution for the marks. We see that the decoupled approximate likelihood method produces reliable intensity and boost parameter estimates, and which are in agreement with the *approximate likelihood method by decoupling the marks parameters*.

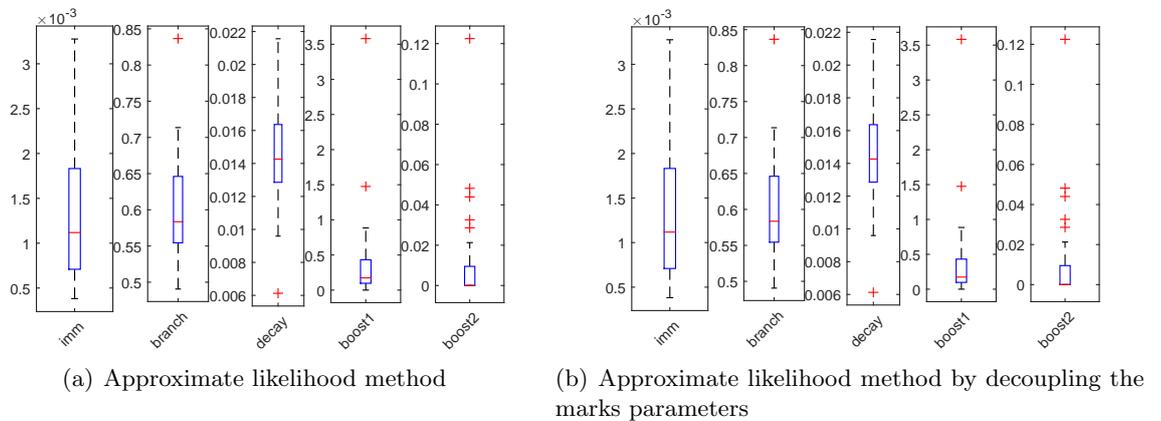


Figure 7.26: Boxplot of the parameter estimates evaluated by the decoupled approximate likelihood method, and the *approximate likelihood method by decoupling the marks parameters*, that specified the bivariate marks as GPD, with joint dependency modelled via a Gaussian copula. The models are estimated across 47 time segments, for SILVER on 31-July-2015, with event types $e \in \{LO, MO, C\}$, levels $l \in \{1, \dots, 5\}$ and bid side only.

In Figure 7.2.3 we have applied the *approximate likelihood method by decoupling the marks parameters*, but in this case with an exponential distribution specification for the marks in the Hawkes process. As we have studied in Section 5.3.5, this is not an appropriate approximate distribution for these two marks, with the marks being heavy tailed. The parameter estimates present significant bias, with a number of time segments producing outliers for both the intensity parameter estimates and boost parameter estimates. In one time segment we can see that this model incorrectly identifies a time segment at criticality ($\hat{\vartheta} \approx 1$).

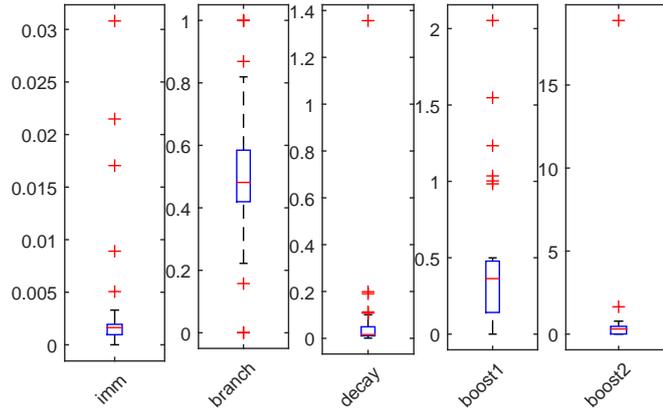


Figure 7.27: Boxplot of the parameter estimates evaluated by the *approximate likelihood method by decoupling the marks parameters*, for a Hawkes process with bivariate marks specified as exponentially distributed, with joint dependency modelled via a Gaussian copula. The models are estimated across 47 time segments, for SILVER on 31-July-2015, with event types $e \in \{LO, MO, C\}$, levels $l \in \{1, \dots, 5\}$ and bid side only.

Table 7.7 shows the mean parameter estimates across the two different likelihood methods and the two parametric specifications for the marks for the Hawkes process, evaluated by the *approximate likelihood method by decoupling the marks parameters*. As noted above, the mean parameter estimates for the decoupled approximate likelihood method and the *approximate likelihood method by decoupling the marks parameters* with generalized Pareto distributed marks are in agreement. However, the mean parameter estimates for the *approximate likelihood method by decoupling the marks parameters* with exponentially distributed marks are vastly different, with the immigration intensity and decay parameters inflated, and the branching coefficient with a downward bias. The boost parameters are unreliable, being vastly different from the analysis conducted on these particular marks throughout this section.

Table 7.7: Mean parameter estimates for a Hawkes process with a linear boost function, estimated across all time segments, for SILVER on 31-July-2015, with event types $e \in \{LO, MO, C\}$, levels $l \in \{1, \dots, 5\}$ and for the bid side only. We consider two methods of estimation and two parametric specifications for the Hawkes process, GPD and exponential for the *approximate likelihood method by decoupling the marks parameters*

Parameter	Approximate likelihood method	Approximate likelihood method by decoupling the marks parameters	
		GPD	Exp
Immigration, η	0.0013	0.0013	0.0032
Branching, ϑ	0.5982	0.5982	0.5122
Decay, α	0.0143	0.0143	0.0694
Boost 1, ψ_1	0.3626	0.3627	0.4265
Boost 2, ψ_2	0.0087	0.0087	0.7104

The decoupled approximate likelihood method is computationally more efficient than the log-likelihood, with a run time of approximately 14 times faster than the *approximate likelihood method by decoupling the marks parameters*. The flexibility it introduces, allows the ability to model the complex data sets of the LOB, without making assumptions

about the parametric distribution of the marks. The parameter estimates obtained from the decoupled approximate likelihood method matched those from the log-likelihood (with the appropriate marginal distribution for the marks) in the application of real limit order book data under all model specifications we have considered in this study. The decoupled approximate likelihood method provides a powerful alternative to the log-likelihood.

7.3 Conclusion

This chapter has brought together all of the strands of research from within this thesis. From the comprehensive description of the vast data sets and the construction of the event arrival process described in detail in Chapter 3, we were able to identify key LOB data sets to model with a Hawkes process. The score test developed in Chapter 6 was applied to the marks identified in Chapter 5. A total of 21 marks were assessed for illustrative purposes, however this is only a small fraction of the marks that could potentially be constructed from the LOB data. Conclusions have been made about the validity of the marks in the Hawkes process via: the choice of empirical moments, which doesn't require the specification of a parametric distribution; adjustments made to the score test in the event of serial dependence, which is observed in many marks; and dependence features being captured without having to specify a copula.

We observed persistence across time segments when using the score test to assess the significance of each mark across various time segment sizes. There are some marks that were significant across assets SILVER and NIKKEI, however there is a degree of variability which would support the individual assessment for each asset in practice. The score test identified highly significant marks across the 10 trading days considered and the results advised the selection of Bid depth and Bid vol MOLOC to be incorporated as marks into a bivariate Hawkes process.

The intensity of events are not stationary, however over small time windows the intensity was approximately stationary. An intra-day time segment of 6 minutes for SILVER and 40 minutes for NIKKEI was selected, ensuring that the sample size of events is large enough to give reliable score test results and reasonable likelihood fits. This conservative choice of time segments was sufficiently large enough to ensure an adequate proportion of segments for selected marks are significant and not an artefact of aggregation.

All pairwise combinations of the 21 marks were assessed by the score test. An illustration of sequentially increasing dimensions of marks demonstrated the power of the score test when testing many marks and the speed at which the assessment under the null hypothesis can be conducted. A test of this kind is not computationally feasible if the assessment was conducted with a model of the log-likelihood.

The Hawkes process was applied to real data. Increasingly complex models were introduced and brought to light the challenges of modelling serially dependent marks, which were further explored in simulation. Simulating a Hawkes process with a linear boost, serially dependent conditional generalized Pareto distributed mark and estimating the parameters by the log-likelihood under the false assumption of i.i.d. marks, resulted in well estimated intensity and boost parameter estimates. The marks marginal distribution

parameter estimates were unreliable, with a significant upward bias of a factor of 9 for the shape parameter, attributed to the increase in tail events, driven by the autocorrelation. This was further confounded in the Hawkes process with serially dependent bivariate marks that are coupled via a Gaussian copula model and conditionally generalized Pareto distributed. Due to the shape parameter inflation and downward bias of the scale parameter in the marginal distribution, it is not possible to establish standard errors via bootstrapping when marks exhibit serial dependence.

The decoupled approximate likelihood method addressed the challenges of correctly identifying the appropriate parametric distributions for a mark vector in the Hawkes process. The parameter estimation is robust in the presence of serial and joint dependent marks and matches those from the log-likelihood method in the application of real limit order book data, under all model specifications we have considered in this study. It is both computationally more efficient than estimating the parameters by the log-likelihood, with a speed of approximately 14 times faster than the log-likelihood method. It provides a powerful alternative to the log-likelihood in the estimation of parameters of the Hawkes process.

Chapter 8

Conclusion

8.1 Summary and contributions

The research work presented in this thesis forms important contributions towards modelling features of the limit order book, a complex dynamical system of orders and cancellations in continuous time at multiple price levels. Through the development of a comprehensive description of the limit order book data, and methods of data aggregation, an appropriate model being the Hawkes self-exciting point process was established. The properties of this model were studied via extensive simulation experiments, challenges were identified and recommendations made. Appropriate marks were established and their statistical properties studied. This led to the development of a score test as a powerful and effective tool for assessment of the marks impact on the intensity function. Application of these methods were applied to limit order book data for futures, with a decoupled approximate likelihood method proposed, providing promising results. This research has produced a comprehensive set of sophisticated tools to model the limit order book data. The development of this research will be of great importance to both investors and market regulators, leading to a deeper understanding of how the limit order book evolves, thus creating a greater understanding of the price discovery mechanism.

8.1.1 The limit order book volume process

The first contribution of this thesis was in response to an incomplete understanding of the features and distributional properties of the volume profiles of the limit order book. Initially, the heavy tailed features of limit order book volumes on level 1 to 5, aggregated to short, evenly spaced time intervals were investigated and found to require a variety of heavy tailed distributional models to adequately capture their statistical features. The methods developed, established the necessary foundations to model other heavy tailed features which occur in many marks that were incorporated in the Hawkes process.

8.1.2 Accurately describing the limit order book and identifying marks

There is an enormous amount of literature on statistical properties of the limit order book, however there is very little literature that gives a detailed description of the limit order book and the limitations and considerations necessary in the decision of the appropriate model for the limit order book. This thesis presented the complicated structure of limit order book data and trade data, and the challenges associated with matching and modelling this data. A detailed analysis of the process steps, by which the physically operating order book is transformed into data suitable for analysis was presented. The description of this process is novel and important because it establishes clearly the limitations of available limit order book data for point process modelling. The limit order book was found to consist of events that frequently occur at the same time, breaching a core assumption of any regular point process, particularly the multivariate Hawkes self-exciting point process often used in this field. To overcome the challenges arising from simultaneity of events in the observed limit order book, events that occur at the same time were treated as a single event and the additional information of the level and type of orders occurring was incorporated in the marks attached to these events. The univariate Hawkes self-exciting point process, which can be enhanced via the inclusion of marks, was proposed as a way to model event clustering on the limit order book.

Substantial empirical research has been conducted on the features of the limit order book, but it has not been quantified in a meaningful way for incorporation into a model for the limit order book. In addition, the applications of Hawkes processes to financial data have almost exclusively ignored marks, and of the limited studies that include marks, only low marks dimension is considered, with no guidance on the construction, properties and challenges associated with marked Hawkes processes. Throughout this research it became apparent that these questions are non-trivial, giving cause for the lack of literature available.

Whilst guided by available literature, an empirical contribution of this thesis was the identification and description of a large number of potential marks which could potentially impact the intensity of the point process. The inclusion of marks, which provide a summary of the nature, level and associated information of the amalgamated events to be modelled, can enrich the Hawkes process. However, the number and complexity of potential marks highlighted the necessity for new methods for selecting which marks have significant impacts on the intensity process. Furthermore, it was concluded that the method for detecting marks should be robust to assumptions such as distributional shape of marks and serial dependence.

8.1.3 A score test for the detection of marks

Joint estimation via maximum likelihood of the Hawkes process parameters and those needed to describe the marks distribution is challenging and in view of the number and complexity of the marks identified as being relevant for modelling the limit order book, a method for screening marks that is computationally straightforward to implement was developed. This new approach was based on the score test, which only requires the single

fitting of the unboosted Hawkes process to the sequence of observed event times, together with the estimates of the moments of the functions of marks under assessment. The moments can be obtained parametrically or non-parametrically. The test has an asymptotic chi-squared distribution under the null hypothesis that the marks do not impact intensity and extensive simulations were presented to confirm the utility of this for realistic models and sample sizes. Additionally, extensive simulation studies of the power of the new test statistic were presented. Extensions of this method to serially dependent marks were developed, something which proves to be essential in detecting appropriate marks in the limit order book. Coupled with the ease of implementation, this makes the score test a powerful and flexible tool to use when identifying appropriate marks for the Hawkes process and forms a core contribution of this thesis.

8.1.4 Hawkes process with multivariate marks for the limit order book

Many of the simulation and estimation challenges of the Hawkes process that arise due to the ultra high frequency data of limit order books and the statistical properties of the marks, have not been studied and have prevented the application of marked Hawkes process in most literature. Through simulation experiments, across a broad range of mark dimensions, distributional forms, dependence models and boost functions, we identified the challenges of fitting a complex model structure and proposed methods to overcome these challenges. A guideline was provided for the methods of simulation and estimating for these processes regarding the software design in MATLAB, that resulted in an efficient and robust implementation.

Through the application of the score test to futures data, marks that have significant impact on the intensity process were identified. Fitting the Hawkes process with multivariate marks to limit order book data via a likelihood based method presented substantial practical challenges, which were investigated via simulation for a variety of model formulations. For the estimation of the model in the presence of serial dependence, a new method was proposed. Through simulation and application to real data, the decoupled approximate method of likelihood estimation was proposed. This method reduces the modelling assumptions on the statistical properties of the marks, and leads to estimation of the Hawkes process parameters and the boost function, which has good performance.

8.2 Future work

The Hawkes process with a power-law decay function presents significant computational challenges. The extensive running time of the calculations required in likelihood optimization, due to the non-recursive process required for evaluation, is further amplified with the ultra high frequency data sets considered for LOB modelling. The exploration of the performance of the sum of exponentials decay that is often utilized to represent a power-law form, should be considered for the marked Hawkes process. In addition, there are substantial challenges with edge effects, due to the long memory properties of the power-law decay function. A deeper understanding of the implications of the edge effects for this process

are required before further application to model ultra high-frequency financial data.

Extensions of the univariate Hawkes process with multivariate marks to a multivariate Hawkes process with multivariate marks should be considered. There are certainly applications whereby the modelling of some subset of the LOB would not result in simultaneity of event times, which is the requirement of any simple point process. In light of the modelling challenges presented in the fitting of the univariate Hawkes process, careful analysis of the multivariate case will need to be conducted.

Given the non-stationarity in the event times for longer time-scales, extensions to a time varying immigration intensity should be considered. In addition, the Hawkes process with a renewal process presented by Stindl and Chen (2018), extended to include multivariate marks will be an excellent candidate study for future research.

Further study on endogenous marks that can be constructed from the LOB and driven by a specific research focus should be considered. Identification and construction of exogenous marks will be highly interesting, especially for the study of futures and changes in the intensity based on the underlying asset features that are not constructed from the futures LOB. In addition, future research should extend the studies on different boost function formulations.

Current theoretical developments for the score test are under-way. Extensions to marked multivariate processes for the score test will extend in a straightforward manner, however, details of implementation and computational challenges with application to real data need to be developed.

Extensions to formulations of the boost function normalization in the presence of jointly dependent marks to higher dimensions, are required to facilitate the modelling of many marks. In addition, the utilization of the components of the information matrix described in Section 6.1 to calculate the standard errors numerically for the decoupled approximate likelihood method are required.

Finally, extensive research will ensue, with the application of the marked Hawkes process to LOB data, to facilitate a deeper study of the many research questions regarding price formation. For example, investigation of the trade off between immigration and the branching ratio, and the extent to which including or excluding marks impacts the estimates. That is, if relevant marks are not included in the intensity, whether this can lead to false conclusions of the exogeneity and endogeneity questions which were discussed in the literature review.

Bibliography

- Abergel, F., Anane, M., Chakraborti, A., Jedidi, A., Muni Toke, I., 2016. *Limit Order Books*. Cambridge University Press.
- Abergel, F., Jedidi, A., 2013. A mathematical approach to order book modelling. *International Journal of Theoretical and Applied Finance* 16.
- Achab, M., Bacry, E., Gaiffas, S., Mastromatteo, I., Muzy, J.F., 2017. Uncovering causality from multivariate Hawkes integrated cumulants. *The Journal of Machine Learning Research* 18, 6998–7025.
- Achab, M., Bacry, E., Muzy, J.F., Rambaldi, M., 2018. Analysis of order book flows using a non-parametric estimation of the branching ratio matrix. *Quantitative Finance* 18, 199–212.
- Ahn, H.J., Bae, K.H., Chan, K., 2001. Limit orders, depth and volatility: Evidence from the Stock Exchange of Hong Kong. *Journal of Finance* 56, 767–788.
- Aït-Sahalia, Y., Cacho-Diaz, J., Laeven, R.J., 2015. Modelling financial contagion using mutually exciting jump processes. *Journal of Financial Economics* 117, 585–606.
- Aït-Sahalia, Y., Hurd, T.R., 2016. Portfolio choice in markets with contagion. *Journal of Financial Econometrics* 14, 1–28.
- Aït-Sahalia, Y., Jacod, J., 2009. Testing for jumps in a discretely observed process. *The Annals of Statistics* 37, 184–222.
- Alder, R., Feldman, R., Taqqu, M.S., 1998. *A practical guide to heavy-tails: Statistical techniques for analysing heavy-tailed distributions*. Birkhäuser.
- Alfonsi, A., Blanc, P., 2016. Dynamic optimal execution in a mixed-market-impact Hawkes price model. *Finance and Stochastics* 20, 183–218.
- Andersen, T., Bollerslev, T., Diebold, F., Labys, P., 2000. Great realizations. *Risk* 13, 105–108.
- Anderson, P., Borgan, O., Gill, R., Keiding, N., 1996. *Statistical Models Based on Counting Processes*. Springer Series in Statistics.
- Anderson, T.W., 1962. On the distribution of the two-sample Cramer-von Mises criterion. *The Annals of Mathematical Statistics* 33, 1148–1159.

- Angus, J.E., 1994. The probability integral transform and related results. *SIAM Review* 36, 652–654.
- Bacry, E., Dayri, K., Muzy, J.F., 2012. Non-parametric kernel estimation for symmetric Hawkes processes. Application to high frequency financial data. *The European Physical Journal B* , 85:157.
- Bacry, E., Delattre, S., Hoffmann, M., Muzy, J.F., 2013a. Modelling microstructure noise with mutually exciting point processes. *Quantitative Finance* 13, 65–77.
- Bacry, E., Delattre, S., Hoffmann, M., Muzy, J.F., 2013b. Some limit theorems for Hawkes processes and application to financial statistics. *Stochastic Processes and their Application* 123, 2475–2499.
- Bacry, E., Jaisson, T., Muzy, J.F., 2016. Estimation of slowly decreasing Hawkes kernel: application to high frequency order book dynamics. *Quantitative Finance* 16, 1179–1201.
- Bacry, E., Mastromatteo, I., Muzy, J.F., 2015. Hawkes processes in finance. *Market Microstructure and Liquidity* 1.
- Bacry, E., Muzy, J.F., 2014. Hawkes model for price and trades high-frequency dynamics. *Quantitative Finance* 14, 1147–1166.
- Bacry, E., Muzy, J.F., 2016. First- and second-order statistics characterization of Hawkes processes and non-parametric estimation. *IEEE Transactions On Information Theory* 62, 2184–2202.
- Balkema, A., Embrechts, P., Nolde, N., 2010. Meta densities and the shape of their sample clouds. *Journal of Multivariate Analysis* 101, 1738–1754.
- Bartolozzi, M., Mellen, C., Matteo, T.D., Aste, T., 2007. Multi-scale correlations in different futures markets. *The European Physical Journal B* 58, 207–220.
- Bauwens, L., Hautsch, N., 2009. Modelling financial high frequency data using point processes, in: Mikosch, T., Kreiß, J.P., Davis, R.A., Andersen, T.G. (Eds.), *Handbook of Financial Time Series*. Springer Series, pp. 953–979.
- Beirlant, J., Goegebeur, Y., Segers, J., Teugels, J., Waal, D., Ferro, C., 2004. *Statistics of Extremes: Theory and Applications*. Wiley, New York.
- Bensalah, Y., 2000. Steps in applying extreme value theory to finance: A review. Working Paper 2000-20 , 151–175.
- Berkowitz, J., Kilian, L., 2000. Recent development in bootstrapping time series. *Econometric Reviews* 19, 1–48.
- Biais, B., Hillion, P., Spatt, C., 1995. An empirical analysis of the limit order book and the order flow in the paris bourse. *The Journal of Finance* 50, 1655 – 1689.

- Black, J., 2010. Financial markets, in: Cane, P., Kritzer, H.M. (Eds.), *The Oxford Handbook of Empirical Legal Research*. Oxford University Press, pp. 151–175.
- Bouchaud, J.P., Gefen, Y., Potters, M., Wyart, M., 2004. Fluctuations and response in financial markets: the subtle nature of ‘random’ price changes. *Quantitative Finance* 4, 176–190.
- Bouchaud, J.P., Mézard, M., Potters, M., 2002. Statistical properties of stock order books: Empirical results and models. *Quantitative Finance* 2, 251–256.
- Bowsher, C.G., 2007. Modelling security market events in continuous time: Intensity based, multivariate point process models. *Journal of Econometrics* 141, 876–912.
- Breitenlechner, M., Gächter, M., Sindermann, F., 2015. The finance-growth nexus in crisis. *Economic Letters* 132, 31–33.
- Bremaud, P., Massoulié, L., 1996. Stability of nonlinear Hawkes processes. *The Annals of Probability* 24, 1563–1588.
- Breusch, T.S., Pagan, A.R., 1980. The Lagrange multiplier test and its applications to model specification in econometrics. *The Review of Economic Studies* 47, 239–253.
- Broyden, G.C., 1970. The convergence of a class of double-rank minimization algorithms. *Journal of the Institute of Mathematics and Its Applications* 6, 76–90.
- Brunnermeier, M.K., Pedersen, L.H., 2009. Market liquidity and funding liquidity. *The Review of Financial Studies* 22, 2201–2238.
- Busch, D., 2017. MiFID II and MiFIR: stricter rules for the EU financial markets. *Law and Financial Markets Review* 11, 126–142.
- Calcagnile, L.M., Bormetti, G., Treccani, M., Marmi, S., Lillo, F., 2018. Collective synchronization and high frequency systematic instabilities in financial markets. *Quantitative Finance* 18, 237–247.
- Cao, C., Hansch, O., Wang, X., 2008. Order placement strategies in a pure limit order book market. *The Journal of Financial Research* 31, 113–140.
- Cao, C., Hansch, O., Wang, X., 2009. The information content of an open limit-order book. *Journal of Futures Markets* 29, 16–41.
- Caprio, G., Beck, T., Claessens, S., 2013. *The Evidence and Impact of Financial Globalization*. Boston: Academic Press.
- Cartea, À., Jaimungal, S., Penalva, J., 2015. *Algorithmic and High-Frequency Trading (Mathematics, Finance and Risk)*. Cambridge University Press.
- Castillo, E., Hadi, A.S., 1997. Fitting the generalized Pareto distribution to data. *Journal of the American Statistical Association* 92, 1609–1620.

- Chakraborti, A., Toke, I.M., Patriarca, M., Abergel, F., 2011. Econophysics review: I. Empirical facts. *Quantitative Finance* 11, 991–1012.
- Challet, D., Stinchcombe, R., 2001. Analyzing and modelling 1+1d markets. *Physica A: Statistical Mechanics and its Applications* 300, 285–299.
- Chavez-Demoulin, V., Davison, A., McNeil, A., 2005. Estimating value-at-risk: a point process approach. *Quantitative Finance* 5, 227–234.
- Chavez-Demoulin, V., McGill, J.A., 2012. High-frequency financial data modeling using Hawkes processes. *Journal of Banking & Finance* 36, 3415–3426.
- Chen, F., Hall, P., 2013. Inference for a nonstationary self-exciting point process with an application in ultra-high frequency financial data modeling. *Journal of Applied Probability* 50, 1006–1024.
- Cheng, I.H., Xiong, W., 2014. Financialization of commodity markets. *Annual Review of Financial Economics* 6, 419–441.
- Chicheportiche, R., Bouchaud, J.P., 2012. Weighted Kolmogorov-Smirnov test: accounting for the tails. *Physical Review E* 86.
- Chordia, T., Roll, R., Subrahmanyam, A., 2001. Market liquidity and trading activity. *The Journal of Finance* 56, 501–530.
- Chordia, T., Roll, R., Subrahmanyam, A., 2002. Order imbalance, liquidity and market returns. *Journal of Financial Economics* 65, 111–130.
- Chordia, T., Roll, R., Subrahmanyam, A., 2008. Liquidity and market efficiency. *Journal of Financial Economics* 87, 249–268.
- Clinet, S., Yoshida, N., 2017. Statistical inference for ergodic point processes and application to limit order book. *Stochastic Processes and their Applications* 127, 1800–1839.
- Comerton-Forde, C., Putnins, T., 2015. Dark trading and price discovery. *Journal of Financial Economics* 118, 70–92.
- Cont, R., 2001. Empirical properties of asset returns: Stylized facts and statistical issues. *Quantitative Finance* 1, 223–236.
- Cont, R., 2007. Volatility clustering in financial markets: Empirical facts and agent-based models, in: Teyssière, G., Kirman, A. (Eds.), *Long Memory in Economics*. Springer, Berlin, Heidelberg, pp. 289–309.
- Cont, R., 2011. Statistical modeling of high-frequency financial data. *IEEE Signal Processing Magazines* 28, 16–25.
- Cont, R., Stoikov, S., Talreja, R., 2010. A stochastic model for order book dynamics. *Operations Research* 58, 549–563.

- Cont, R., Tankov, P., 2004. *Financial Modelling with Jump Processes*. Chapman and Hall/CRC.
- Cruz, M.G., Peters, G.W., Shevchenko, P.V., 2015. *Fundamental Aspects of Operational Risk and Insurance Analytics: A Handbook of Operational Risk*. John Wiley & Sons, Inc.
- Daley, D., Vere-Jones, D., 2007. *An introduction to the theory of point processes*. Springer-Verlag New York. 2 edition.
- Dassios, A., Zhao, H., 2011. A dynamic contagion process. *Advances in Applied Probability* 43, 814–846.
- Davis, R.A., Mikosch, T., 2009. The extremogram: A correlogram for extreme events. *Bernoulli* 15, 977–1009.
- Davis, R.A., Mikosch, T., Cribben, I., 2012. Towards estimating extremal serial dependence via the bootstrapped extremogram. *Journal of Econometrics* 170, 142–152.
- Dayri, K.A., 2012. *Market Microstructure and Modeling of the Trading Flow*. Ph.D. thesis. Ecole Polytechnique.
- Denuit, M., Dhaene, J., Goovaerts, M., Kaas, R., 2005. *Actuarial Theory for Dependent Risks. Measures, Order and Models*. John Wiley & Sons, Ltd.
- Diggle, P., Heagerty, P., Liang, K.Y., Zeger, S., 2002. *Analysis of Longitudinal Data*. Oxford University Press. 2 edition.
- Dungey, M., Martin, V.L., 2007. Unravelling financial market linkages during crises. *Journal of Applied Econometrics* 22, 89–119.
- Dunsmuir, W.T.M., Clinet, S., Peters, G.W., Richards, K.L., 2018. Asymptotic distributions of the score test for detecting marks in Hawkes processes. In Preparation .
- Ellul, A., Holden, C.W., Jain, P., Jennings, R., 2003. Determinants of order choice on the New York Stock Exchange. Working Paper, Indiana University .
- Embrechts, P., Kirchner, M., 2018. Hawkes graphs. *Theory of Probability and its Application* 62, 132–156.
- Embrechts, P., Klüppelberg, C., Mikosch, T., 1997. *Modelling Extremal Events for Insurance and Finance*. volume 33. Springer-Verlag.
- Embrechts, P., Liniger, T., Lin, L., 2011. Multivariate Hawkes processes: An application to financial data. *Journal of Applied Probability* 48A, 367–378.
- Embrechts, P., Resnick, S., Samorodnitsky, G., 1999. Extreme value theory as a risk management tool. *North American Actuarial Journal* 3, 30–41.
- Engle, R.F., Lunde, A., 2003. Trades and quotes: A bivariate point process. *Journal of Financial Econometrics* 1, 159–188.

- Engle, R.F., Russell, J.R., 1998. Autoregressive conditional duration: A new model for irregularly spaced transaction rates. *Econometrica* 66, 1127–1162.
- Epps, T.W., 1979. Comovements in stock prices in the very short run. *Journal of the American Statistical Association* 74, 291–298.
- Epstein, G.A., 2005. *Financialization of the world economy*. Edward Elgar Publishing Limited.
- Errais, E., Giesecke, K., Goldberg, L., 2010. Affline point processes and portfolio credit risk. *SIAM Journal on Financial Mathematics* 1, 642–665.
- Fama, E., 1965. The behaviour of stock market prices. *The Journal of Business* 38, 34–105.
- Fama, E., Roll, R., 1968. Some properties of symmetric stable distributions. *Journal of the American Statistical Association* 63, 817–836.
- Fauth, A., Tudor, C., 2012. Modeling first line of an order book with multivariate marked point processes. Working Paper arXiv:1211.4157 .
- Filimonov, V., Sornette, D., 2012. Quantifying reflexivity in financial markets: Toward a prediction of flash crashes. *Physical Review E* 85, 056108–1 056108–9.
- Filimonov, V., Sornette, D., 2015. Apparent criticality and calibration issues in the Hawkes self-excited point process model: application to high-frequency financial data. *Quantitative Finance* 15, 1293–1314.
- Filimonov, V., Wheatley, S., Sornette, D., 2015. Effective measure of endogeneity for the autoregressive conditional duration point processes via mapping to the self-excited hawkes process. *Communications in Nonlinear Science and Numerical Simulation* 22, 23–37.
- Fletcher, R., 1970. A new approach to variable metric algorithms. *Computer Journal* 13, 317–322.
- Francq, C., Zakoïan, J., 2013. Estimating the marginal law of a time series with applications to heavy tailed distributions. *Journal of Business & Economic Statistics* 31, 412–425.
- Gao, X., Zhou, X., Zhu, L., 2017. Transform analysis for hawkes processes with applications in dark pool trading. *Quantitative Finance* 18, 265–282.
- Geman, H., 2005. *Commodities and Commodity Derivatives: Pricing and Modeling Agricultural, Metals and Energy*. Wiley Finance.
- Geman, H. (Ed.), 2008. *Risk Management in Commodity Markets: from Shipping to Agriculturals and Energy*. Wiley Finance.
- Godsill, S., 2000. Inference in symmetric alpha-stable noise using MCMC and the slice sampler. *Acoustics, Speech, and Signal Processing VI*, 3806–3809.

- Goldfarb, D., 1970. A family of variable metric updates derived by variational means. *Mathematics of Computation* 24, 23–26.
- Goldsmith, R.W., 1959. *The Comparative Study of Economic Growth and Structure*. NBER.
- Goldsmith, R.W., 1969. *Financial structure and development*. Yale University Press, New Haven, CT.
- Goncalves, S., Kilian, L., 2004. Bootstrap autogressions with conditional heteroskedasticity of unknown form. *Journal of Econometrics* 123, 89–120.
- Gopikrishnan, P., Plerou, V., Gabaix, X., Stanley, H.E., 2000. Statistical properties of share volume traded in financial markets. *Physical Review E* 62.
- Gottesman, A., 2016. *Derivatives Essentials: An introduction to forwards, futures, options and swaps*. John Wiley & Sons, Inc.
- Gould, M., Bonart, J., 2016. Queue imbalance as a one-tick-ahead price predictor in a limit order book. *Market Microstructure and Liquidity* 2.
- Gould, M.D., Porter, M.A., Williams, S., McDonald, M., Fenn, D.J., Howison, S.D., 2013. Limit order books. *Quantitative Finance* 13, 1709–1742.
- Grimshaw, S.D., 1993. Computing maximum likelihood estimates for the generalized Pareto distribution. *Technometrics* 35, 185–191.
- Gu, G.F., Chen, W., Zhou, W.X., 2008. Empirical shape function of limit-order books in the Chinese stock market. *Physica A: Statistical Mechanics and its Applications* 387, 5182–5188.
- Gu, W., Peters, G.W., Clavier, L., Septier, F., Nevat, I., 2012. Receiver study of cooperative communications in convolved and additive alpha-stable interference plus gaussian thermal noise. 2012 International Symposium on Wireless Communication System (ISWCS) .
- Guan, Y., 2006. Tests for independence between marks and points of a marked point process. *Biometrics* 62, 126–134.
- Hall, A.D., Hautsch, N., 2006. Order aggressiveness and order book dynamics. *Empirical Economics* 30, 973–1005.
- Hardiman, S.J., Bercot, N., Bouchaud, J.P., 2013. Critical reflexivity in financial markets: a Hawkes process analysis. *The European Physical Journal B* 86.
- Hardiman, S.J., Bouchaud, J.P., 2014. Branching ratio approximation for the self-exciting Hawkes process. *Physical Review E* 90.
- Harris, L., 2003. *Trading and Exchanges: Market microstructure for practitioners*. Oxford University Press.

- Hasbrouck, J., 1988. Trades, quotes, inventories, and information. *Journal of Financial Economics* 22, 229–252.
- Hautsch, N., 2012. *Econometric of Financial High-Frequency Data*. Springer-Verlag Berlin Heidelberg.
- Hawkes, A.G., 1971. Spectra of some self-exciting and mutually exciting point processes. *Biometrika* 58, 83–90.
- Hawkes, A.G., 2018. Hawkes processes and their applications to finance: a review. *Quantitative Finance* 18, 193–198.
- Hewlett, P., 2006. Clustering of order arrivals, price impact and trade path optimization. *Workshop on Financial Modeling with Jump Processes, Ecole Polytechnique*, 6–8.
- Hosking, J., 1990. L-moments: analysis and estimation of distributions using linear combinations of order statistics. *Journal of the Royal Statistical Society. Series B (Methodological)* 52, 105–124.
- Hu, K., Ivanov, P.C., Chen, Z., Carpena, P., Stanley, H.E., 2001. Effect of trends on detrended fluctuation analysis. *Physical Review E* 64.
- Huang, H.C., Su, Y.C., Liu, Y.C., 2014. The performance of imbalance-based strategy on tender offer announcement day. *Investment Management and Financial Innovations* 11.
- Huang, W., Lehalle, C.A., Rosenbaum, M., 2015. Simulating and analyzing order book data: The queue-reactive mode. *Journal of the American Statistical Association* 110, 107–122.
- Jacod, J., Shiryaev, A., 2013. *Limit Theorems for Stochastic Processes*. volume 288. Springer Science & Business Media.
- Jaisson, T., 2015. Market impact as anticipation of the order flow imbalance. *Quantitative Finance* 15, 1123–1135.
- Kang, S.H., Cheong, C., Yoon, S.M., 2013. Intraday volatility spillovers between spot and futures indices: Evidence from the Korean stock market. *Physica A: Statistical Mechanics and its Applications* 392, 1795–1802.
- Kantelhardt, J.W., Koscielny-Bunde, E., Rego, H.H.A., Havlin, S., Bunde, A., 2001. Detecting long-range correlations with detrended fluctuations analysis. *Physica A: Statistical Mechanics and its Applications* 295, 441–454.
- Kettell, B., 2002. *Economics for financial markets*. Oxford; Boston: Butterworth-Heinemann.
- Khashanah, K., Chen, J., Hawkes, A.G., 2018. A slightly depressing jump model: intraday volatility pattern simulation. *Quantitative Finance* 18, 213–224.

- Kirchner, M., 2016. Hawkes and INAR(∞) processes. *Stochastic Processes and their Application* 126, 2494–2525.
- Kirchner, M., 2017a. An estimation procedure for the Hawkes process. *Quantitative Finance* 17, 571–595.
- Kirchner, M., 2017b. Perspective on Hawkes Processes. Ph.D. thesis. ETH Zürich.
- Kirchner, M., Bercher, A., 2018. A nonparametric estimation procedure for the Hawkes process: comparison with maximum likelihood estimation. *Journal of Statistical Computation and Simulation* 88, 1106–1116.
- Koning, A., Peng, L., 2008. Goodness-of-fit tests for a heavy tailed distribution. *Journal of Statistical Planning and Inference* 138, 3960–3981.
- Kratz, M., Resnick, S., 1996. The QQ-estimator and heavy tails. *Stochastic Models* 12, 699–724.
- Ladley, D., 2012. Zero intelligence in economics and finance. *The Knowledge Engineering Review* 27, 273–286.
- Lallouache, M., Challet, D., 2016. The limits of statistical significance of Hawkes processes fitted to financial data. *Quantitative Finance* 16, 1–11.
- Large, J., 2007. Measuring the resiliency of an electronic limit order book. *Journal of Financial Markets* 10, 1–25.
- Laub, P.J., Taimre, T., Pollet, P.K., 2015. Hawkes processes. Working Paper arXiv: 1507.02822 .
- Lee, C.M.C., Ready, M.J., 1991. Inferring trade direction from intraday data. *The Journal of Finance* 46, 733–746.
- Levine, R., 1997. Financial development and economic growth: Views and agenda. *Journal of Economic Literature* 35, 688–726.
- Lévy, P., 1924. Theorie des erreurs. La Loi de Gauss et les Lois Exceptionnelles, *Bull. Soc. Math.* 52, 587–615.
- Liniger, T., 2009. Multivariate Hawkes Processes. Ph.D. thesis. ETH Zürich.
- Liu, R.Y., 1988. Bootstrap procedures under some non-i.i.d. models. *The Annals of Statistics* 16, 1696–1708.
- Ljung, G.M., Box, G.E.P., 1978. On a measure of lack of fit in time series models. *Biometrika* 65, 297–303.
- Lo, I., Sapp, S.G., 2010. Order aggressiveness and quantity: How are they determined in a limit order market? *Journal of International Financial Markets, Institutions and Money* 20, 213–237.

- Lobato, I., Velasco, C., 2000. Long memory in stock-market trading volume. *Journal of Business & Economic Statistics* 18, 410–427.
- Longstaff, F.A., 2010. The subprime credit crisis and contagion in financial markets. *Journal of Financial Econometrics* 97, 436–450.
- Lu, X., Abergel, F., 2017. Limit order book modelling with high dimensional Hawkes processes. Working Paper hal-01512430 .
- Mandelbrot, B., 1960. The Pareto-Levy law and the distribution of income. *International Economic Review* 1, 79–106.
- Maslov, S., Mills, M., 2001. Price fluctuations from the order book perspective - empirical facts and a simple model. *Physica A: Statistical Mechanics and its Applications* 299, 234–246.
- Massey, F.J., 1951. The Kolmogorov-Smirnov test for goodness-of-fit. *Journal of the American Statistical Association* 46, 68–78.
- MathWorks, 2018. MATLAB Optimization Toolbox User’s Guide. The MathWorks, Inc. 3 Apple Hill Drive Natick, MA 01760-2098. r2018a edition.
- Mayntz, R., 2012. Crisis and control: Institutional change in financial market regulation. volume 75. Publication Series of the Max Planck Institute for the Study of Societies, Cologne, Germany.
- McCulloch, J.H., 1986. Simple consistent estimators of stable distribution parameters. *Communications in Statistics - Simulation and Computation* 15, 1109–1136.
- McCulloch, J.H., 1997. Numerical approximation of the symmetric stable distribution and density. *Communications in Statistics. Stochastic Models* 13, 759–774.
- McNeil, A., Frey, R., Embrechts, P., 2005. Quantitative risk management: concepts, techniques, and tools. Princeton University Press.
- Melchiori, M.R., 2006. Tools for sampling multivariate Archimedean copulas. Working Paper SSRN.1124682 .
- Moloney, N., 2010. EU financial market regulation after the global financial crisis: “more europe” or more risk? *Common Market Law Reivew* 47, 1317–1383.
- Morrison, J., Smith, J., 2002. Stochastic modeling of flood peaks using the generalized extreme value distribution. *Water resources research* 38, 41–1–41–12.
- Münnix, M.C., Schäfer, R., Guhr, T., 2011. Statistical causes for the Epps effect in microstructure noise. *International Journal of Theoretical and Applied Finance* 14, 1231–1249.
- Nelsen, R.B., 1999. An introduction to copulas. Springer, New York.

- Nikias, C., Shao, M., 1995. Signal processing with alpha stable distributions and applications. Wiley, New York.
- Nolan, J.P., 2007. Stable distributions: Models for heavy-tailed data. Birkhäuser, Boston.
- Ogata, Y., 1978. The asymptotic behaviour of maximum likelihood estimators for stationary point processes. *Annals of the Institute of Statistical Mathematics* 30, 243–261.
- Ogata, Y., 1981. On Lewis’ simulation method for point processes. *IEEE Transactions on Information Theory* 27, 23–31.
- Ogata, Y., 1988. Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical Association* 83, 9–27.
- O’Hara, M., 1997. Market Microstructure Theory. Wiley-Blackwell UK.
- Omi, T., Hirata, Y., Aihara, K., 2017. Hawkes process model with a time-dependent background rate and its application to high-frequency financial data. *Physical Review E* 96.
- Ozaki, T., 1979. Maximum likelihood estimation of Hawkes self-exciting point processes. *Annals of the Institute of Statistical Mathematics* 31, 145–155.
- Panayi, E., 2015. Modelling empirical features and liquidity resilience in the limit order book. Ph.D. thesis. University College of London.
- Panayi, E., Peters, G.W., Kosmidis, I., 2015. Liquidity commonality does not imply liquidity resilience commonality: A functional characterisation for ultra-high frequency cross-sectional LOB data. *Quantitative Finance* 15, 1737–1758.
- Pedersen, L.H., 2013. When everyone runs for the exit. NBER Working Paper No.15297 .
- Peters, G., Shevchenko, P., Young, M., Yip, W., 2011. Analytic loss distributional approach models for operational risk from the α -stable doubly stochastic compound processes and implications for capital allocation. *Insurance: Mathematics and Economics* 49, 565–579.
- Peters, G.W., Nevat, I., Septier, F., Clavier, L., 2012. Generalized inference models in doubly stochastic Poisson random fields for wideband communications: the PNSC (alpha) model. Working Paper arXiv:1207.1531 .
- Peters, G.W., Vishnia, G.R., 2017. Blockchain architectures for electronic exchange reporting requirements: EMIR, Dodd Frank, MiFID I/II, MiFIR, REMIT, reg NMS and T2S, in: Chuen, D.L.K., Deng, R.H. (Eds.), *Handbook of Blockchain, Digital Finance, and Inclusion*. Academic Press. volume 2, pp. 271–329.
- Peters, G.W., Wüthrich, M.V., Shevchenko, P.V., 2010. Chain ladder method: Bayesian bootstrap versus classical bootstrap. *Insurance: Mathematics and Economics* 47, 36–51.

- Pickands III, J., 1975. Statistical inference using extreme order statistics. *The Annals of Statistics* 3, 119–131.
- Politis, D.N., Romano, J.P., 1994. The stationary bootstrap. *Journal of the American Statistical Association* 89, 1303–1313.
- Pomponio, F., Abergel, F., 2013. Multiple-limit trades: empirical facts and applications to lead-lag measures. *Quantitative Finance* 13, 783–793.
- Potters, M., Bouchaud, J.P., 2003. More statistical properties of order books and price impact. *Physica A: Statistical Mechanics and its Applications* 324, 133–140.
- Prescott, P., Walden, A., 1980. Maximum likelihood estimation of the parameters of the generalized extreme-value distribution. *Biometrika* 67, 723–724.
- Rambaldi, M., Bacry, E., Lillo, F., 2017. The role of volume in order book dynamics: a multivariate Hawkes process analysis. *Quantitative Finance* 17, 999–1020.
- Rambaldi, M., Bacry, E., Muzy, J.F., 2018. Disentangling and quantifying market participant volatility contributions. Working Paper arXiv:1807.07036 .
- Ranaldo, A., 2004. Order aggressiveness in limit order book markets. *Journal of Financial Markets* 7, 53–74.
- Rasmussen, J.G., 2011. Temporal point processes: the conditional intensity function.
- Ravi, R., Sha, Y., 2014. Autocorrelated order-imbalance and price momentum in the stock market. *International Journal of Economics and Finance* 6, 39–54.
- Richards, K.L., Peters, G.W., Dunsmuir, W.T., 2015. Heavy-tailed features and dependence in limit order book volume profiles in futures markets. *International Journal of Financial Engineering* 2, 1–56.
- Rousseau, P., Wachtel, P., 2011. What is happening to the impact of financial deepening on economic growth? *Economic Inquiry* 49, 276–288.
- Royston, P., 1992. Which measures of skewness and kurtosis are best? *Statistics in Medicine* 11, 333–343.
- Russell, J.R., 1999. Econometric modelling of multivariate irregularly-spaced high-frequency data. Mimeo, University of Chicago, Graduate School of Business .
- Samorodnitsky, G., Taqqu, M.S., 1994. Stable non-Gaussian random processes: stochastic models with infinite variance. Chapman and Hall/CRC.
- Schittkowski, K., 1986. NLPQL: A fortran subroutine solving constrained nonlinear programming problems. *Annals of Operations Research* 5, 485–500.
- Schlather, M., Ribeiro, P.J., Diggle, P.J., 2004. Detecting dependence between marks and locations of marked point processes. *Journal of the Royal Statistical Society* 66, 79–93.

- Schneider, M., Lillo, F., Pelizzon, L., 2018. Modelling illiquidity spillovers with Hawkes processes: an application to the sovereign bond market. *Quantitative Finance* 18.
- Schoenberg, F.P., 2004. Testing separability in spatial-temporal marked point processes. *Biometrics* 60, 471–481.
- Shanno, D., 1970. Conditioning of quasi-newton methods for function minimization. *Mathematics of Computation* 24, 647–656.
- Silvennoinen, A., Thorp, S., 2013. Financialization, crisis and commodity correlation dynamics. *Journal of International Financial Markets, Institutions and Money* 24, 42–65.
- Simonsen, I., Hansen, A., 1998. Determination of the Hurst exponent by use of wavelet transforms. *Physical Review E* 58.
- Sirignano, J., Cont, R., 2018. Universal features of price formation in financial markets: perspective from deep learning. Working Paper SSRN 3141294 .
- Smith, E., Farmer, J., Gillemot, L., Krishnamurthy, S., 2003. Statistical theory of the continuous double auction. *Quantitative Finance* 3, 481–514.
- Smith, R., 1985. Maximum likelihood estimation in a class of nonregular cases. *Biometrika* 72, 67–90.
- Solo, V., Pasha, A., 2012. A test for independence between a point process and an analogue signal. *Journal of Time Series Analysis* 33, 824–840.
- Soros, G., 1987. *The Alchemy of Finance: Reading the Mind of the Market*. Wiley, New York.
- Stindl, T., Chen, F., 2018. Likelihood based inference for the multivariate renewal Hawkes process. *Computational Statistics and Data Analysis* 123, 131–145.
- Tang, K., Xiong, W., 2012. Index investment and the financialization of commodities. *Financial Analysts Journal* 68, 54–74.
- Toke, I.M., 2011. “Market making” in an order book model and its impact on the spread, in: Abergel, F., Chakrabarti, B., A., A.C., Mitra, M. (Eds.), *Econophysics of Order-driven Markets*. New Economic Windows. Springer, Milano, pp. 49–64.
- Toke, I.M., 2015. The order book as a queue system: average depth and influence of the size of limit orders. *Quantitative Finance* 15, 795–808.
- Toke, I.M., 2016. Reconstruction of order flows using aggregated data. *Market Microstructure and Liquidity* 2.
- Toke, I.M., Pomponio, F., 2012. Modelling trades-through in a limit order book using Hawkes processes. Working Paper hal-00745554 .

- Tomaskovic-Devey, D., Lin, K.H., 2013. Financialization: causes, inequality consequences, and policy implications. *N.C. Banking Institute* 18, 167–194.
- Tóth, B., Kértész, J., 2009. The Epps effect revisited. *Quantitative Finance* 9, 793–802.
- Trebbi, F., Xiao, K., 2017. Regulation and market liquidity. Working Paper mmsc.2017.2876 .
- Valiante, D., 2013. Commodities Price Formation: Financialization and beyond. Technical Report. CEPS-ECMI Task Force Report, Centre for European Policy Studies, Brussels.
- Viney, C., 2002. Financial Institutions, Instruments and Markets. McGraw-Hill Book Company Australia Pty Limited.
- Vinkovskaya, E., 2014. A point process model for the dynamics of limit order books. Ph.D. thesis.
- Vogel, R., Fennessey, N., 1993. L moment diagrams should replace product moment diagrams. *Water Resources Research* 29, 1745–1752.
- Yang, S.Y., Liu, A., Chen, J., Hawkes, A.G., 2018. Applications of a multivariate Hawkes process to joint modeling of sentiment and market return events. *Quantitative Finance* 18, 295–310.
- Zheng, B., Roueff, F., Abergel, F., 2014. Modelling bid and ask prices using constrained Hawkes processes: Ergodicity and scaling limit. *SIAM Journal on Financial Mathematics* 5, 99–136.
- Zhu, L., 2013. Nonlinear Hawkes Processes. Ph.D. thesis. New York University.
- Zolotarev, V.M., 1983. Univariate stable distributions. *Translations of Mathematical Monographs*, American Math. Soc. 65.
- Zolotarev, V.M., 1986. One-Dimensional Stable Distributions. *Translations of Mathematical Monographs*, American Mathematical Society.
- Zovko, I., Farmer, J.D., 2002. The power of patience: a behavioural regularity in limit order placement. *Quantitative Finance* 2, 387–392.

Appendix A

Chapter 4

A.1 Copula models for distributions of marks

In this section we outline some important classes of models that are relevant to the application of modelling the dynamics of the LOB. All material on copula models that follows, is an extract of relevant work by Cruz et al. (2015). We will consider the joint distribution of the marks constructed according to a meta-copula framework, see Balkema et al. (2010) and Cruz et al. (2015). Under such a framework we consider developing the joint distribution of the marks random vector based on the second component of Sklars theorem. The joint multivariate distribution follows

$$F_{\mathbf{X}}(x_1, \dots, x_d) = C(F_1(x_1), F_2(x_2), \dots, F_d(x_d); \Upsilon), \quad (\text{A.1})$$

for some choice of model for the copula C , parametrized by Υ and marginal distributions.

In this section we briefly present the two classes of copula model we consider in this work, Gaussian and Archimedean.

A.1.1 Gaussian Copula

The d -dimensional Gaussian copula is obtained by the transformation of the multivariate Normal distribution

$$C(u_1, \dots, u_d) = F_N^\Sigma(F_N^{-1}(u_1), \dots, F_N^{-1}(u_d)) \quad (\text{A.2})$$

and its density is

$$c(u_1, \dots, u_d) = \frac{f_N^\Sigma(F_N^{-1}(u_1), \dots, F_N^{-1}(u_d))}{\prod_{k=1}^d f_N(F_N^{-1}(u_k))}. \quad (\text{A.3})$$

$F_N(\cdot)$ is the standard Normal distribution with density $f_N(\cdot)$. The standard multivariate Normal distribution is $F_N^\Sigma(\cdot)$ with density $f_N^\Sigma(\cdot)$. They have zero means, unit variances, and correlation matrix Σ .

A.1.2 Archimedean Copula

Generally, Archimedean copulae are not derived from a well-known parametric multivariate distribution, nevertheless they can be stated explicitly in a simple form. Many Archimedean copulae have been proposed in the literature, see Nelsen (1999), with many further copulae available as extensions and combinations of these base copulae. Archimedean copulae are attractive to researchers and practitioners, due to their directly interpretable tail dependence features and parsimonious representations.

We begin this section with a basic definition of the bivariate Archimedean copula and then this is generalized to the d -variate copula case. This is followed by a detailed account of the required properties of the generator function of this family of parametric dependence models. Further detail can be found in Cruz et al. (2015), with only key results related directly to this research presented here.

According to Denuit et al. (2005, definition 4.7.6), the extension of the Archimedean copula family to d -dimensions is achieved by considering the strictly monotone generator function ψ , such that $\psi : (0, 1] \rightarrow \mathbb{R}^+$ with $\psi(1) = 0$, then the resulting Archimedean copula can be expressed as detailed next.

Definition 7 (*d -Dimensional Archimedean Copula*). *A d -dimensional copula C is called Archimedean if for some generator ψ it can be represented as*

$$C(\mathbf{u}) = \psi\{\psi^{-1}(u_1) + \cdots + \psi^{-1}(u_d)\} = \psi\{t(\mathbf{u})\} \quad \forall \mathbf{u} \in [0, 1]^d, \quad (\text{A.4})$$

where $\psi^{-1}(0) = \inf\{t : \psi(t) = 0\}$.

The family of Archimedean copula models has the following useful properties (as detailed in a simple bivariate setting) presented in Lemma 1. It should be noted that the result generalizes to d dimensions, however for the purposes of this research we consider only the bivariate case.

Lemma 1. *Let C be an Archimedean copula with generator ψ . Then according to Nelsen (1999, lemma 4.1.2 and theorem 4.1.5), the following properties hold:*

1. *C is an Archimedean copula if it can be represented by*

$$C(u, v) = \psi^{[-1]}(\psi(u) + \psi(v)),$$

where ψ is the generator of this copula and is a continuous, strictly decreasing function from $[0, 1]$ to $[0, \infty]$ such that $\psi(1) = 0$ and $\psi^{[-1]}$ is the pseudo inverse of ψ ;

2. *C is symmetric, $C(u, v) = C(v, u) \forall (u, v) \in [0, 1] \times [0, 1]$;*
3. *C is associative, $C(C(u, v), w) = C(u, C(v, w)) \forall (u, v, w) \in [0, 1]^3$;*
4. *If $c > 0$ is any constant, then $c\psi$ is a generator of C .*

One-Parameter Archimedean Members

Cruz et al. (2015) provide some explicit distribution and density representations for some widely utilized subfamilies of Archimedean copulae families. Lemma 2 presents the one parameter versions of the Archimedean copulae.

Lemma 2. *From the results in Nelsen (1999, section 4.3, table 4.1), the distribution and density functions of the Clayton, Gumbel, and Frank copulae subfamilies follow.*

1. **Clayton Copula.** *The distribution and density are given respectively as*

$$C^c(u_1, \dots, u_d) = \left(1 - n + \sum_{k=1}^n u_k^{-\rho^c}\right)^{-1/\rho^c}, \quad (\text{A.5})$$

$$c^c(u_1, \dots, u_d) = \left(1 - n + \sum_{k=1}^n (u_k)^{-\rho^c}\right)^{-n-\frac{1}{\rho^c}} \prod_{k=1}^d \left((u_k)^{-\rho^c-1} ((k-1)\rho^c + 1)\right), \quad (\text{A.6})$$

where $\rho^c \in [0, \infty)$ is the dependence parameter. The generator and inverse generator for the Clayton copula are given by

$$\psi_C(t) = (t^{-\rho} - 1); \quad \psi_C^{-1}(s) = (1 + s)^{-\frac{1}{\rho}}. \quad (\text{A.7})$$

The Clayton copula does not have upper tail dependence. Its lower tail dependence is $\lambda_L = 2^{-1/\rho^c}$.

2. **Gumbel Copula.** *The distribution function is given by*

$$C^g(u_1, \dots, u_d) = \exp\left(-\left[\sum_{k=1}^d (-\ln(u_k))^{\rho^g}\right]^{\frac{1}{\rho^g}}\right), \quad (\text{A.8})$$

where $\rho^g \in [1, \infty)$ is the dependence parameter. The generator and inverse generator for the Gumbel copula are given by

$$\psi_G(t) = (-\ln t)^\rho; \quad \psi_G^{-1}(s) = \exp\left(-s^{\frac{1}{\rho}}\right). \quad (\text{A.9})$$

The Gumbel copula does not have lower tail dependence. The upper tail dependence of the Gumbel copula is $\lambda_U = 2 - 2^{1/\rho^g}$. In the bivariate case, the explicit expression for the Gumbel copula density is given by

$$\begin{aligned} c(u_1, u_2) &= \frac{\partial^2}{\partial u_1 \partial u_2} C(u_1, u_2) \\ &= C(u_1, u_2) u_1^{-1} u_2^{-1} \left[\sum_{k=1}^2 (-\ln u_k)^\rho\right]^{2\left(\frac{1}{\rho}-1\right)} (\ln u_1 \ln u_2)^{\rho-1} \\ &\quad \times \left[1 + (\rho - 1) \left[\sum_{k=1}^2 (-\ln u_k)^\rho\right]^{-\frac{1}{\rho}}\right]. \end{aligned}$$

3. **Frank Copula.** The distribution function is given by

$$C^F(u_1, \dots, u_n) = \frac{1}{\rho} \ln \left(1 + \frac{\prod_{k=1}^d (e^{\rho^F u_k} - 1)}{(e^{\rho^F} - 1)^{n-1}} \right), \quad (\text{A.10})$$

where $\rho^F \in \mathbb{R}/\{0\}$ is the dependence parameter. The Frank copula does not have upper or lower tail dependence. In the bivariate case, one can represent the copula density for the Frank distribution as

$$\begin{aligned} c(u_1, u_2) &= \frac{\partial^2}{\partial u_1 \partial u_2} C(u_1, u_2) \\ &= \frac{\rho [1 - \exp(-\rho)] \exp(-\rho(u_1 + u_2))}{([1 - \exp(-\rho)] - (1 - \exp(-\rho u_1))(1 - \exp(-\rho u_2)))^2}. \end{aligned}$$

In general, it will be of practical use to be able to evaluate the copula density pointwise and it has already been demonstrated that this will in general require up to d -th order derivatives for a d -variate Archimedean copula of mixed types. Hence, it is possible to combine the following derivative results for the different Archimedean copula models discussed and their generators with the formula for composite differentiation in Cruz et al. (2015, definition 11.8) (Table A.1).

Table A.1: Archimedean copula generator functions, inverse generator functions, and generator function d -th derivatives, $\psi^{(d)}$.

Family	ψ	ψ^{-1}	$(-1)^d \psi^{(d)}$
Clayton	$(1+t)^{-1/\rho}$	$(s^{-\rho} - 1)$	$\frac{\Gamma(d+1/\rho)}{\Gamma(1/\rho)} (1+t)^{-(d+1/\rho)}$
Frank	$-\frac{1}{\rho} \ln [1 - e^{-t}(1 - e^{-\rho})]$	$-\ln \frac{e^{-s\rho} - 1}{e^{-\rho} - 1}$	$\frac{1}{\rho} Li_{-(d-1)} \{ (1 - e^{-\rho}) e^{-t} \}$
Gumbel	$\exp(-t^{1/\rho})$	$(-\ln s)^\rho$	$\frac{\psi_\rho(t)}{t^d} P_{d,1/\rho}^g(t^{1/\rho})$

The densities for the one-parameter copulae in this table can be calculated using Cruz et al. (2015, equation 11.74).

We note the following definitions are utilized

$$\begin{aligned} a_{dk}^g \left(\frac{1}{\rho} \right) &= \frac{d!}{k!} \sum_{i=1}^k \binom{k}{i} \binom{i/\rho}{d} (-1)^{d-i}, \quad k \in \{1, \dots, d\}, \\ Li_s(z) &= \sum_{k=1}^{\infty} \frac{z^k}{k^s}, \\ P_{d, \frac{1}{\rho}}^g \left(t^{\frac{1}{\rho}} \right) &= \sum_{k=1}^d a_{dk}^g \left(\frac{1}{\rho} \right) \left(t^{\frac{1}{\rho}} \right)^k. \end{aligned} \quad (\text{A.11})$$

Hence, using these closed form results, combined with the knowledge of composite function differentiation via Fa di Bruno, the required multivariate densities can be determined for implementation of likelihood and Bayesian estimation methods using the

definition given by

$$\begin{aligned}
c(u_1, u_2, \dots, u_n, \dots, u_d) &= \frac{\partial C(u_1, u_2, \dots, u_n, \dots, u_d)}{\partial u_1 \partial u_2 \cdots \partial u_n \cdots \partial u_d} \\
&= \frac{\psi^{(d)}(\psi^{-1}(u_1) + \cdots + \psi^{-1}(u_d))}{\psi^{(1)}(\psi^{-1}(u_1)) \psi^{(1)}(\psi^{-1}(u_2)) \cdots \psi^{(1)}(\psi^{-1}(u_d))}.
\end{aligned} \tag{A.12}$$

Remark 3. *It is also worth noting that in the case of the bivariate Clayton copula model, it can be shown that this subfamily is comprehensive in its coverage in the sense that its dependence properties can range from the Frechet-Hoeffding lower bound of perfect negative dependence, through to the Frechet-Hoeffding upper bound corresponding to perfect positive dependence.*

It is also worth noting that occasionally an alternative parametrization of the multivariate Clayton copula is presented according to the distribution

$$C_\theta\{u_1, u_2, \dots, u_d\} = \max\left\{u_1^{1-\theta} + u_2^{1-\theta} + \cdots + u_d^{1-\theta} - d + 1, 0\right\}^{\frac{1}{1-\theta}}, \tag{A.13}$$

with $\theta \in [0, \infty) \setminus \{1\}$. In addition, in the case that $\theta \in [0, 1)$, the condition on the dimension given by $d \leq \lfloor (1 - \theta)^{-1} \rfloor + 1$ is required to ensure that the resulting function is a valid distribution, see discussions in Nelsen (1999).

A.2 Moments

For the boost normalization required in (4.4), there is a common component that will be required, given by the following identities linking marginal central, $\mu_{j,k}$ and non-central (raw) moments, $\mu'_{j,k}$. Due to the choice in boost function and dimensions of marks, we consider up to the order four, $j \in \{1 \dots 4\}$ for the moments and up to the bivariate case, $k \in \{1, 2\}$ for marks. We define the non-central moments in terms of central moments as:

$$\begin{aligned}
\mu'_{2,k} &= \mu_{2,k} + \mu_{1,k}^2; \\
\mu'_{3,k} &= \mu_{3,k} + 3\mu_{2,k}\mu'_{1,k} + \mu_{1,k}^3; \\
\mu'_{4,k} &= \mu_{4,k} + 4\mu_{3,k}\mu'_{1,k} + 6\mu_{2,k}\mu_{1,k}^2 + \mu_{1,k}^4.
\end{aligned} \tag{A.14}$$

A.3 Simulation algorithm for the univariate Hawkes process for both recursive and non-recursive procedures

To simulate a series of random events according to the specification of a given Hawkes SEPP, we utilize Ogata's modified thinning algorithm (Ogata, 1981). It can be applied to a Hawkes SEPP with any specification of decay function. For the general framework and notation, we rely on the algorithm description by Liniger (2009, Algorithm 1.21), however extensions are made for inclusion of joint dependence structures for the marks.

The running time of calculations for a Hawkes SEPP with a power-law decay will have a quadratic complexity $O(N_T^2)$. There are various ways to optimize the code. Firstly, Liniger (2009) proposes an algorithm that involves truncating the intensity and using the ϵ parameter to improve the approximation precision. We have not implemented this. As we will discuss, we have utilized array programming (vectorization) to transform loops to vector operations for speed of calculation for the non-recursive intensity function. For the non-recursive process, this is especially important as each iteration in the inner loop requires a cumulative calculation of the history of the process, substantially slowing computation within a traditional programming structure.

The intensity process λ , is left continuous with right-hand side limits. For the purposes of simulation, we define λ^+ as the right-hand side limit after time t , that is, after the intensity is boosted. We denote λ^- as the left-hand-side limit, also known as the *hazard rate* (Liniger, 2009).

The algorithm is comprised of an outer and inner loop. The outer loop iteration variable is $i \in \{2, \dots, n\}$ and the inner loop is indexed by $r \geq 1$. This procedure applies to the recursive method. We present both the recursive method and the non-recursive method in the algorithm below.

Algorithm 13 (Hawkes SEPP simulation: recursive and non-recursive). *Initialization:*

$$i = 2; \quad \lambda^+(1) = \eta/(1 - \vartheta).$$

Monte-Carlo method for the case of jointly dependent marks

Examples of some functions in MATLAB are provided below.

1. For $Y_{i,j} \sim F_{i,j}$, where $j \in \{2, \dots, d\}$ are jointly dependent random variables with copula W and marginal distributions $F_{i,1}, \dots, F_{i,d}$, let $U_{i,1}, \dots, U_{i,d}$ have joint distribution C . For a very long time series, $i = 1 \dots n$ (E.g. $n = 20,000$):

(a) Simulate a random variable $u_{i,1} \sim U[0, 1]$

- example code in MATLAB: $\rho = \text{copulaparam}(\text{'Gaussian'}, \rho_s, \text{'type'}, \text{'spearman'}); [u_1, u_2] = \text{copularnd}(\text{'Gaussian'}, \rho_s, 1);$

(b) Simulate a random variable $u_{i,2}$ from $C_{i,2}(\cdot|u_{i,1})$. Continue simulating, such that you simulate a random variable $u_{i,d}$ from $C_{i,d}(\cdot|u_{i,d-1})$;

(c) Sample $(Y_{i,1}, \dots, Y_{i,d})$ from $F^{-1}(u_{i,1}), \dots, F^{-1}(u_{i,d})$

- example code in MATLAB: $y_{i,j} = \text{gpinv}(u_{i,j}, \zeta_j, \delta_j, 0).$

2. Define $y_{i,j} := Y_{i,j}$ and calculate the long run empirical moments and correlation in the event of joint dependence, which we will consider below and as required for the specification of boost function, for example, $\mathbb{E}[Y_j]$, $\text{Var}[Y_j]$ and $\rho(Y_1, \dots, Y_d)$.

Outer loop

1. Initialization:

$$\tau_1 := t_{i-1}; \quad \lambda_r^- := \lambda_{i-1}^+.$$

Inner loop

- (a) Calculate a new time τ_r

$$\tau_r = \begin{cases} t(i-1) + E/\lambda^+(i-1), & \text{if } r = 1, \\ \tau_{r-1} + E/\lambda^+(i-1), & \text{otherwise,} \end{cases}$$

where $E \sim \text{Exp}(1)$

- (b) Calculate the left hand limit intensity λ_r^- . The calculation of the intensity function λ_r^- is dependent on the choice of recursive versus non-recursive procedure. Note that power-law decay function can only be evaluated via the non-recursive procedure. We outline both below.

Recursive: For the case of an exponential decay function in (4.21), we can utilize the more efficient recursive expression for the simulation.

$$\lambda_r^- = \eta + e^{-\alpha(\tau_r - t_{i-1})} [\lambda_{i-1}^+ - \eta].$$

Non-recursive: The non-recursive procedure requires evaluation of the whole history at each new time point. Following Algorithm 5, we present a procedure utilizing array programming (vectorization) to transform loops to vector operations for speed of calculation for the non-recursive intensity function.

- i. Create a time difference matrix with row i and column j ,

$$TD_{i,r}^{kj} = \begin{bmatrix} \mathbf{t_2 - t_1} & 0 & \dots & \dots & \dots & t_2 - \tau_r \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \mathbf{t_{i-1} - t_1} & \dots & \dots & \mathbf{t_{i-1} - t_{i-2}} & 0 & t_{i-1} - \tau_r \\ \tau_r - \mathbf{t_1} & \dots & \dots & \dots & \tau_r - \mathbf{t_{i-1}} & 0 \end{bmatrix} \quad (\text{A.15})$$

For the purposes of each iteration in the simulation and in contrast to the non-recursive calculation, we are only taking the sum of the $i = r$ th row, up to the $j = i - 1$ th column at each iteration. The **bolded** text represents the time differences that are utilized. From Algorithm 5 and for each decay function specification we evaluate the following.

- A. For the exponential kernel in (4.21)

$$w_{i,r} = \alpha \sum_{j=1}^{k-1} \exp(-\alpha TD^{kj}).$$

$$\lambda_{i,r}^- = \begin{cases} \eta + \vartheta\alpha w_{i,r}, & \text{if } i = 2, \\ \eta + \vartheta\alpha w_{i,r}g(\mathbf{x}; \phi, \psi), & \text{otherwise.} \end{cases}$$

B. For the power-law kernel in (4.23)

$$w_{i,r} = \sum_{j=1}^{i-1} \frac{(\alpha - 1)\beta}{(1 + \beta T D^{kj})^\alpha}.$$

$$\lambda_{i,r}^- = \begin{cases} \eta + \vartheta w_{i,r}, & \text{if } i = 2, \\ \eta + \vartheta w_{i,r}g(\mathbf{x}; \phi, \psi). & \text{otherwise.} \end{cases}$$

(c) Sample a standard uniform random variable $U_r \sim U(0, 1)$ and let

$$u_r := U_r \lambda_{i-1}^+.$$

Check the condition

$$u_r \leq \hat{\lambda}_r^-.$$

If this condition is true, exit the inner loop. Failing this condition being met, we start the inner loop again, sampling a new τ_r , where $r = r + 1$.

Exit the inner loop.

2. Define $t_i := \tau_r$.
3. Define $\lambda(i) := \lambda^-(r)$
4. In the case of univariate marks or multi-dimensional marks that are independent, sample a random variable from $X_{i,j} \sim F_{i,j}$, where $j \in \{1, \dots, d\}$ and define $x_{i,j} := X_{i,j}$.
5. In the case of $d \geq 2$ jointly dependent marks with copula W and marginal mark distributions $F_{i,1}, \dots, F_{i,d}$, let $U_{i,1}, \dots, U_{i,d}$ have joint distribution C . Follow the method outlined in ‘Monte-Carlo method for the case of jointly dependent marks [step (a)]’, to obtain samples $(X_{i,1}, \dots, X_{i,d})$ from $F^{-1}(u_{i,1}), \dots, F^{-1}(u_{i,d})$.
6. Calculate the boost function $g(\mathbf{x}; \phi, \psi)$, using the normalization adjustment in Section 4.2.1 in the event of jointly dependent marks. For the evaluation of $\mathbb{E}[X_1 X_2] \neq 0$, we require the estimation of the linear correlation. As described in Section 4.2.1, if this cannot be calculated explicitly, we utilize the long run empirical linear correlation from the ‘Monte-Carlo method for the case of jointly dependent marks’ above, $\rho(Y_1, \dots, Y_d)$.
7. **Recursive:** Define $\lambda^+(i) = \lambda_r^- + \vartheta\alpha g(\mathbf{x}; \phi; \psi)$.
8. **Non-recursive:** Define $\lambda^+(i) = \lambda_r^- + \vartheta g(\mathbf{x}; \phi; \psi)$.
9. Let, $i=i+1$.

For the implementation of the algorithm, we need to define an end point for the iteration variable r , which we set to 200.

A.3.1 Simulation study to verify that the recursive method matches the non-recursive method in the case of exponential decay

For the full joint likelihood method, we propose two algorithms, a non-recursive process (Algorithm 5) and a recursive process (Algorithm 7) for the estimation of the intensity function. For the case of an intensity function with an exponential decay, which is a Markov process, the recursive procedure will work and it is a far more efficient algorithm for calculation. For the case of an intensity function with a power-law decay, the non-recursive process is required. In the exponential decay case, where either method can be chosen, they are mathematically equivalent and aside from the speed of calculation, they will result in equivalent simulation results and parameter estimations from the full joint likelihood method.

Verification of the implementation of both methods and their mathematical equivalence in likelihood estimation, can be done by estimation with both methods for the same replicates, using the exponential decay function. Table A.2 shows that both methods produce identical results to four decimal places. Figure A.1 presents the difference between the parameter estimates of each method and we can see that the difference is negligible.

Table A.2: Recursive method and the non-recursive method for the full joint likelihood parameter estimation of a Hawkes SEPP with linear boost and an Exponentially distributed mark, $X \sim \text{Exp}(\lambda = 1)$. The simulation has 1000 replicates of sample size $n = 1000$ each.

Parameter	Recursive				Non-recursive			
	mean	std	min	max	mean	std	min	max
Immigration, η	0.0020	0.0002	0.0015	0.0028	0.0020	0.0002	0.0015	0.0028
Branching, ϑ	0.6980	0.0325	0.5866	0.7981	0.6980	0.0325	0.5866	0.7981
Decay, α	0.0101	0.0009	0.0072	0.0129	0.0101	0.0009	0.0072	0.0129
Scale, λ	1.0001	0.0315	0.9167	1.0975	1.0001	0.0315	0.9167	1.0975
Boost, ψ	0.5257	0.2149	0.0280	1.9221	0.5257	0.2149	0.0280	1.9221

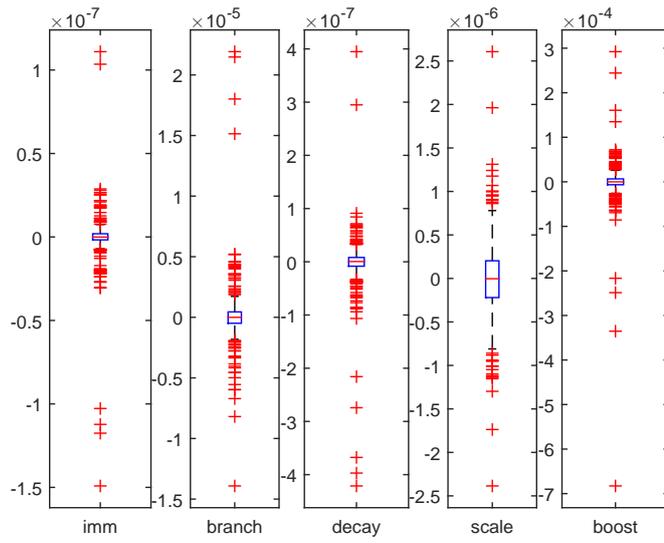


Figure A.1: Recursive method parameter estimate minus the non-recursive method parameter estimates in the full joint likelihood of a Hawkes SEPP with linear boost and an Exponentially distributed mark, $X \sim \text{Exp}(\lambda = 1)$. The simulation has 1000 replicates of sample size $n = 1000$ each.

Appendix B

Chapter 6

B.1 Score Test Derivation Details

Recall the notation for parameters is $\nu = (\theta, \phi, \psi)^\top$ and the true value under H_0 is $\nu^* = (\theta^*, \phi^*, 0)$. The derivatives of the log-likelihood (4.15) with respect to ν are

$$\partial_\nu l_g(\nu) = \int_{[0,T] \times \mathbb{X}} \frac{1}{\lambda_g(t; \nu)} \partial_\nu \lambda_g(t; \nu) N_g(dt \times d\mathbf{x}) - \partial_\nu \Lambda(T; \nu) + \frac{1}{f(x; \phi)} \partial_\nu f(x; \phi). \quad (\text{B.1})$$

Consider the components of these partial derivatives. First with respect to $\theta = (\eta, \vartheta, \alpha)$. Since $\partial_\theta f(x; \phi) = 0$ and $\partial_\psi f(x; \phi) = 0$ the last term in (B.1) does not contribute to $\partial_\theta l_g(\theta)$ or $\partial_\psi l_g(\theta)$. Also it is straightforward to show that $\partial_\theta l_g(\nu)|_{\nu^*}$ does not depend on the marks. Now

$$\partial_\phi \lambda_g(t; \nu)|_{\nu^*} = \vartheta^* \int_{[0,t) \times \mathbb{X}} w(t-s; \alpha^*) \frac{-h(x; \phi^*, 0)}{\mathbb{E}_{\phi^*}[h(X; \phi, 0)]} \partial_\phi \mathbb{E}_{\phi^*}[h(X; \phi, \psi)]|_{\nu^*} N_g^0(ds \times d\mathbf{x}) = 0,$$

since, at ν^* , $\psi = 0$ and $h(x_m; \phi^*, 0) = 1$, $\mathbb{E}_{\phi^*}([h(X; \phi^*, 0)]) = 1$ and we have additionally assumed that $\partial_\phi \mathbb{E}_\phi[h(X; \phi, \psi)]|_{\nu^*} = 0$ in Section 1. Hence $\partial_\phi \Lambda_g(T; \nu)|_{\psi=0} = 0$ also. Consequently

$$\partial_\phi l_g(\nu^*) = \partial_\phi \ln f(\mathbf{x}; \phi^*).$$

Now

$$\partial_\psi \lambda_g(t; \nu)|_{\nu^*} = \vartheta^* \int_{[0,t) \times \mathbb{X}} w(t-s; \alpha^*) G(\mathbf{x}; \phi^*) N_g^0(ds \times d\mathbf{x}),$$

and inserting the above expressions into the vector of derivatives of the log likelihood (B.1) gives (6.11) which we repeat here for reference:

$$\partial_\psi l_g(\nu^*) = \int_{[0,T]} \lambda(t; \theta^*)^{-1} \partial_\psi \lambda_g(t; \nu^*) \tilde{N}(dt).$$

It is clear that, since $\partial_\theta l_g(\nu^*)$ does not contain the marks x_m , then

$$\mathbb{E}_{\nu^*}[\partial_\theta l_g(\nu^*) \partial_\psi l_g(\nu^*)^\top] = 0.$$

Also $\partial_\phi l_g(\nu^*) = \partial_\phi \ln f(\mathbf{x}; \phi^*)$ results in

$$\mathbb{E}_{\nu^*}[\partial_\theta l_g(\nu^*) \partial_\phi l_g(\nu^*)^\top] = 0.$$

To complete the proof that the information matrix is block diagonal with blocks corresponding to the partition of ν into θ , ϕ and ψ we require to show that

$$\mathbb{E}_{\nu^*}[\partial_\theta l_g(\nu^*) \partial_\psi l_g(\nu^*)^\top] = 0.$$

Now $\partial_\phi l_g(\nu_0)$ and $\partial_\psi l_g(\nu_0)$ both involve functions of the marks X_m , but using Condition 1 part (v) the zero expected value follows.

To put all this together to derive the score statistic (6.17) we note that, using the properties of the first partial derivative obtained above and the fact that the score vectors with respect to θ and ϕ are zero at the maximum likelihood estimates obtained under H_0 , we have $\partial_\nu l_g(\hat{\nu}_0) = (0, 0, \partial_\psi l_g(\hat{\nu}_0))^\top$. Hence the score statistic is

$$S_n = \partial_\psi l_g(\hat{\nu}_0)^\top \mathcal{I}_{\psi, \psi}(\nu_0)^{-1} \partial_\psi l_g(\hat{\nu}_0), \quad (\text{B.2})$$

where $\mathcal{I}_{\psi, \psi}(\nu_0) = \mathbb{E}_{\nu_0}[\partial_\psi l_g(\nu_0) \partial_\psi l_g(\nu_0)^\top]$. In practice the information matrix for ψ is evaluated empirically using estimates of θ and ϕ obtained under H_0 , that is $\hat{\theta}_T$ is obtained from the likelihood using the intensity process without marks and $\hat{\phi}_T$ maximises the likelihood for the marks density. Alternatives based on sample moments for the marks process are discussed in the main text.

The second derivative matrix of the log-likelihood with respect to ψ is

$$\begin{aligned} \partial_{\psi\psi}^2 l_g(\nu) &= \int_{[0, T] \times \mathbb{X}} \partial_\psi (\lambda_g(t; \theta)^{-1} \partial_\psi \lambda_g(t; \nu)) N_g(ds \times d\mathbf{x}) - \int_{[0, T]} \partial_{\psi\psi}^2 \lambda_g(t; \nu) dt \\ &= \int_{[0, T] \times \mathbb{X}} \partial_\psi (\lambda_g(t; \theta)^{-1} \partial_\psi \lambda_g(t; \nu)) (N_g(ds \times d\mathbf{x}) - \lambda_g(t; \nu^*) dt) \\ &\quad - \int_{[0, T]} \lambda_g(t; \nu)^{-1} \partial_{\psi\psi}^2 \lambda_g(t; \nu) \{\lambda_g(t; \nu) - \lambda_g(t; \nu^*)\} dt \\ &\quad - \int_{[0, T]} \lambda_g(t; \nu)^{-2} \partial_\psi \lambda_g(t; \nu)^{\otimes 2} \lambda_g(t; \nu^*) dt, \end{aligned} \quad (\text{B.3})$$

where

$$\partial_{\psi\psi}^2 \lambda_g(t; \nu) = \vartheta \int_{[0, t] \times \mathbb{X}} w(t-s; \alpha) \partial_{\psi\psi}^2 g(\mathbf{x}; \phi, \psi) N_g(ds \times d\mathbf{x}). \quad (\text{B.4})$$

Let $\nu_0 = (\theta, \phi, 0)$ (that is at any values of θ and ϕ when $\psi = 0$) then (B.3) and (B.4)

evaluate at ν_0 , as

$$\begin{aligned}\partial_{\psi\psi}^2 l_g(\nu_0) &= \int_{[0,T]} \{\lambda(t;\theta)^{-1} \partial_{\psi\psi}^2 \lambda_g(t;\nu_0) - \lambda(t;\theta)^{-2} \partial_{\psi} \lambda_g(t;\nu_0)^{\otimes 2}\} \tilde{N}(ds) \\ &\quad - \int_{[0,T]} \lambda(t;\theta)^{-1} \partial_{\psi\psi}^2 \lambda_g(t;\nu_0) \{\lambda(t;\theta) - \lambda(t;\theta^*)\} dt \\ &\quad - \int_{[0,T]} \lambda(t;\theta)^{-2} \partial_{\psi} \lambda_g(t;\nu_0)^{\otimes 2} \lambda(t;\theta^*) dt,\end{aligned}\tag{B.5}$$

and

$$\partial_{\psi\psi}^2 \lambda_g(t;\nu) = \vartheta \int_{[0,t] \times \mathbb{X}} w(t-s; \alpha) \partial_{\psi\psi}^2 g(\mathbf{x}; \phi, \psi) N_g(ds \times d\mathbf{x}),\tag{B.6}$$

respectively.

To derive the information matrix at ν^* (the true parameter under H_0) we get, using (6.11),

$$\begin{aligned}\mathcal{I}_{\psi,\psi}(\nu^*) &= \mathbb{E}_{\nu^*} [\partial_{\psi} l_g(\nu^*) \partial_{\nu} l_g(\nu^*)^{\top}] \\ &= \mathbb{E} \left[\left(\int_{[0,T]} \lambda(t;\theta^*)^{-1} \partial_{\psi} \lambda_g(t;\nu^*) \tilde{N}(dt) \right)^{\otimes 2} \right] \\ &= \mathbb{E} \left[\int_{[0,T]} \lambda(t;\theta^*)^{-1} \partial_{\psi} \lambda_g(t;\nu^*)^{\otimes 2} dt \right],\end{aligned}\tag{B.7}$$

which is (6.20). Alternatively, using (B.5), we get

$$\begin{aligned}\mathcal{I}_{\psi,\psi}(\nu^*) &= -\mathbb{E}_{\nu^*} [\partial_{\psi\psi}^2 l_g(\nu^*)] \\ &= -\mathbb{E} \left[\int_{[0,T]} \lambda(t;\theta^*)^{-1} \partial_{\psi\psi}^2 \lambda_g(t;\nu^*) \tilde{N}(dt) \right] \\ &\quad + \mathbb{E} \left[\int_{[0,T]} \lambda(t;\theta^*)^{-2} \partial_{\psi} \lambda_g(t;\nu^*)^{\otimes 2} N(dt) \right].\end{aligned}\tag{B.8}$$

Now

$$\partial_{\psi\psi}^2 g(X; \phi, \psi)|_{\nu^*} = H'(X) - \mathbb{E}_{\phi^*} [H'(X)] - 2\mathbb{E}_{\phi^*} [H(X)](H(X) - \mathbb{E}_{\phi^*} [H(X)]),$$

which has expectation zero, hence

$$\mathbb{E}[\partial_{\psi\psi}^2 \lambda_g(t;\nu^*) | \mathcal{F}_{t-}] = 0,$$

so that the first term in (B.8) is zero. Hence

$$\mathcal{I}_{\psi,\psi}(\nu^*) = \mathbb{E} \left[\int_{[0,T]} \lambda(t;\theta^*)^{-2} (\partial_{\psi} \lambda_g(t;\nu^*))^{\otimes 2} N(dt) \right],\tag{B.9}$$

which is (6.22) and also obviously equal to (B.7) by using $N(dt) = \tilde{N}(dt) + \lambda(t;\theta^*)dt$ in (B.9).

B.2 Properties of decay functions

Two decay functions are often suggested and some others such as the mixture of exponential decay functions (Hardiman et al., 2013) can be covered with these examples.

Exponential Decay:

$$w(s; \alpha) = \alpha e^{-\alpha s}, \quad s > 0, \quad \alpha > 0$$

$$W(s; \alpha) = 1 - e^{-\alpha s}$$

$$\frac{\partial w(s; \alpha)}{\partial \alpha} = (1 - \alpha s)e^{-\alpha s}$$

$$\frac{\partial W(s; \alpha)}{\partial \alpha} = s e^{-\alpha s}$$

Power Law Decay:

$$w(s; \alpha, \beta) = \frac{(\alpha - 1)\beta}{(1 + \beta s)^\alpha}, \quad s > 0, \quad \alpha > 2, \quad \beta > 0$$

$$W(s; \alpha, \beta) = 1 - \frac{1}{(1 + \beta s)^{\alpha-1}}$$

$$\frac{\partial w(s; \alpha, \beta)}{\partial \alpha} = \frac{\beta}{(1 + \beta s)^\alpha} [1 - (\alpha - 1) \log(1 + \beta s)]$$

$$\frac{\partial w(s; \alpha, \beta)}{\partial \beta} = \frac{(\alpha - 1)(1 + \beta s - \alpha \beta^2)}{(1 + \beta s)^{\alpha+1}}$$

$$\frac{\partial W(s; \alpha, \beta)}{\partial \alpha} = \frac{\log(1 + \beta s)}{(1 + \beta s)^{\alpha-1}}$$

$$\frac{\partial W(s; \alpha, \beta)}{\partial \beta} = \frac{(1 - \alpha)s}{(1 + \beta s)^\alpha}$$

Requirements (4.2) clearly hold for both examples. Likewise uniform boundedness of $|w(s; \alpha)|$ is also clear. In (4.2) and (4.3) we also assumed that $|\partial w(s, \alpha)/\alpha_j|$ and $|\partial W(s, \alpha)/\alpha_j|$ are uniformly bounded for $s > 0$ and all α in a ball round the true parameter vector α . These conditions also clearly hold for the exponential and power law decay functions using the above expressions for the derivatives.