POINT PROCESS MODELLING OF A LIMIT ORDER BOOK

Oden Petersen

Supervisor: Dr. Tom Stindl

School of Mathematics and Statistics
UNSW Sydney

November 2024

Submitted in partial fulfillment of the requirements of the degree of
Bachelor of Advanced Mathematics with Honours

Plagiarism statement

I declare that this thesis is my own work, except where acknowledged, and has not been submitted for academic credit elsewhere.

I acknowledge that the assessor of this thesis may, for the purpose of assessing it:

- Reproduce it and provide a copy to another member of the University; and/or,
- Communicate a copy of it to a plagiarism checking service (which may then retain a copy of it on its database for the purpose of future plagiarism checking).

I certify that I have read and understood the University Rules in respect of Student Academic Misconduct, and am aware of any potential plagiarism penalties which may apply.

By signing this declaration I am agreeing to the statements and conditions above.

Signed: _____ Date: _____

Acknowledgements

Abstract

This is the abstract

Contents

Chapter 1 Introduction 1
1.1 Limit Order Books . . . 2
1.2 The Matching Algorithm . . . 3
1.2.1 The Bid and Ask . . . 3
1.2.2 Liquidity . . . 4
1.2.3 Queue Priority . . . 5
1.3 Queueing Models of the Order Book . . . 5

Chapter 2 Point Process Models 7
2.1 Overview of Point Processes . . . 7
2.1.1 The Counting Measure . . . 8
2.1.2 The Expectation Measure . . . 8
2.1.3 Intensity of a Point Process . . . 9
2.1.4 Residuals and the Compensator Process . . . 10
2.1.5 Marked and Multivariate Point Processes . . . 10
2.1.6 Composite Point Processes . . . 11
2.2 Point Process Models of Book Updates . . . 12
2.2.1 Poisson Processes . . . 12
2.2.2 Inhomogeneous Poisson Process . . . 13
2.2.3 Hawkes Processes . . . 13
2.2.4 State Dependence . . . 15
2.2.5 Raising the Intensity to a Power . . . 17

Chapter 3 Estimation 19
3.1 Empirical Likelihood . . . 19
3.1.1 Empirical Intensity and Loglikelihood . . . 20
3.2 Maximum Likelihood Estimation . . . 20
3.2.1 Poisson Processes . . . 21
3.3 Maximum Likelihood Estimation for Composite Processes . . . 22
3.3.1 Expectation Maximisation . . . 23
3.4 General Point Process Model . . . 24
3.4.1 Differentiating with respect to the Parameters . . . 25
3.5 Uncertainty Quantification . . . 31
3.6 Computational Concerns . . . 31
3.7 Model Selection . . . 31

Chapter 4 Simulation 32
4.1 Point Process Simulation . . . 32

4.2 . . . 33
4.3 Simulation Study of Estimation Methods . . . 33
4.4 Impulse Response Function . . . 33

Chapter 5 Application to KOSPI/SPY Data 34
5.1 Dataset . . . 34
5.2 Point Process Modelling . . . 34
5.2.1 Santa Fe Model . . . 34
5.2.2 Inhomogeneity . . . 34
5.2.3 Hawkes Kernel . . . 35
5.2.4 State Dependence . . . 35
5.2.5 Regression . . . 35
5.2.6 Power Hawkes . . . 35
5.2.7 Idiosyncratic Daily Behaviour . . . 35
5.3 notes . . . 35
5.3.1 Clustering Ratio . . . 37
5.4 Summary of datasets . . . 37
5.5 Inhomogeneity . . . 37
5.6 Multidimensionality . . . 37
5.7 Self-Excitation and Mutual Excitation . . . 38
5.8 Queue Size Dependence . . . 38

Chapter 6 Conclusion 39
6.0.1 Hidden States . . . 39
6.0.2 Daily Variation . . . 39

Chapter 7 Appendix: Foundations of Probabilistic Models 40
7.1 Measure Theory Fundamentals . . . 40
7.2 Probability Spaces . . . 41
7.2.1 Conditionalisation . . . 42
7.3 Density of a Measure . . . 43
7.4 Stochastic Process Fundamentals . . . 43

References 44

Chapter 1

Introduction

On an average day in 2023, the US stock market saw around $500 billion dollars worth of shares traded on various exchanges and other venues, almost 2% of the annual GDP [7]. Modern securities markets facilitate the exchange of shares and other financial assets at extremely high frequency, as a result of aggressive investment in specialised networking hardware, custom-made computer architectures, and high-throughput machine learning systems. In facilitating capital flows for the global economy, financial markets have at the same time managed to claim an increasing fraction of resources and attention, with economic consequences that are not yet fully understood [1].

Despite many mysteries and open questions about the origins and dynamics of market phenomena, the financial sector itself has readily adapted to the increasing scale and complexity of the markets in which it operates. The practical design of exchange rules and trading systems has in large part been an empirical endeavour on the part of market participants, operators, and regulators. Many phenomena have been observed to emerge in an apparently decentralised fashion from the application of exploitative heuristics and predictive algorithms that interact with exchanges and aggregate to produce desirable outcomes. While users of trading strategies aim to maintain acceptable risk levels while generating profits over the long term, market operators and regulators are tasked with the design of incentive mechanisms that exploit this self-interested behaviour to improve market outcomes, including reduced transaction costs, fast and accurate incorporation of external information (such as economic news or earnings reports), and adherence to various concepts of fairness, propriety, and legality.

In this thesis, I describe and extend prior work from the empirical market microstructure literature, making extensive use of the state-dependent Hawkes process model for event arrivals. I begin with a conceptual overview of the trading mechanism, and formalise the mathematical tools that will be used to construct and describe variations on the basic Hawkes process model. I continue with a summary of existing literature on point processes as applied to market data, including both mathematical foundations and empirical findings. Next, I explore techniques to reduce the computational burden of parametric inference for point processes on large datasets. Finally, empirical applications and findings are discussed, including applications of generative modeling to a variety of open problems in market microstructure.

1.1 Limit Order Books

Intraday trading allows participants to respond to exogenous news and endogenous market events in a manner that maintains acceptable levels of risk and generates profits over the long term. This activity has a significant influence on the formation of market prices and plays a key role in reducing transaction costs while increasing the speed at which large institutions can control their exposure to various financial risks and opportunities.

Securities exchanges facilitate automated matching of buyers and sellers at prices favourable to both. Understanding the dynamics of this exchange process at a high degree of resolution can provide insights into the design of automated *matching engines* that produce desirable market behaviour, as well as insights into the design of *trading strategies* that exploit the dynamics of the exchange process to generate profits.

A matching engine is tasked with receiving and acting on various messages from market participants indicating their intent to buy or sell a particular security with particular conditions. As a result of this process, trades may be formed that match buyers and sellers at a mutually agreeable price and quantity. Trade reports are broadcast to relevant participants and may be used for the purposes of risk management and forecasting, as well as for the ultimate transfer of the assets that have been traded (which often occurs after trading hours).

A typical matching engine permits two kinds of incoming messages, known as *order insertion* and *order cancellation*, and maintains an internal state consisting of a single data structure, known as a *limit order book*. An order insertion message indicates a participant's willingness to buy (or sell) some quantity of a particular asset at or below (respectively, above) a particular price, and results in the addition of an *order* to the limit order book $\mathcal{L}$. Conversely, an order cancellation message results in the removal of a particular order from the limit order book, either in part (by reducing the remaining volume associated with the order) or in full (by removing the order entirely from the book). It might be good to include a statistic about roughly how large cancellation rates are, to highlight the importance of this message type.

Formally, a limit order book can be defined as a set of tuples ("orders") of the form

$$(\text{side}, \text{price}, \text{time}, \text{size}) = (s, p, t, q) \in \{-1, 1\} \times P \times T \times Q.$$

Each published order represents an intention to buy (or sell) some quantity of an asset at a maximum (respectively, minimum) price.

For example, Figure 1.1 shows the structure of a limit order submitted at 9:52am, expressing an intention to sell up to two units of a particular security for a price at least as favourable as \$84.10 per unit.

Describe tick sizes, lot sizes.

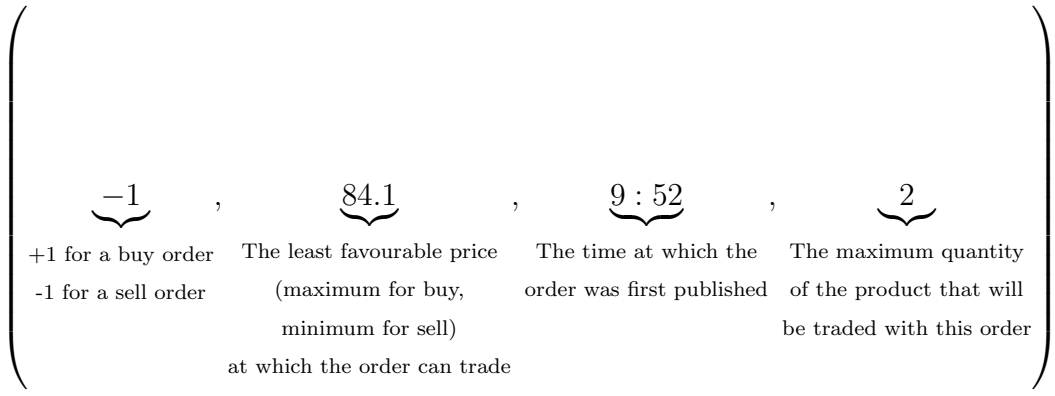


Figure 1.1: Components of an example limit order

1.2 The Matching Algorithm

1.2.1 The Bid and Ask

In order to describe the process by which changes to $\mathcal{L}$ results in trades, I will consider partitioning $\mathcal{L}$ into a *bid side*,

$$\mathcal{L}^+ = \{(1, p, t, q) \in \mathcal{L}\},$$

and an *ask side*,

$$\mathcal{L}^- = \{(-1, p, t, q) \in \mathcal{L}\}.$$

The most competitive prices on each side of the book are known as the *bid price* and *ask price*, given by

$$\text{bid}_{\mathcal{L}} = \max_{\{p:(1,p,t,q)\in\mathcal{L}\}} p,$$

$$\text{ask}_{\mathcal{L}} = \min_{\{p:(-1,p,t,q)\in\mathcal{L}\}} p.$$

These prices are known as the *top levels* of the order book.

Whenever an order insertion message is received whose addition would result in $\text{bid}_{\mathcal{L}} \geq \text{ask}_{\mathcal{L}}$, the exchange will attempt to form trades with the existing orders on the opposing side of the book, such that as much volume as possible is matched. Existing orders on the opposing side of the book will be removed or depleted to equal the volume of the incoming order, and volume that cannot be matched at a price agreeable to the incoming order will finally be added to L. As a result of this process, the invariant

$$\text{bid}_{\mathcal{L}} < \text{ask}_{\mathcal{L}}$$

is maintained after the processing of each book event.

Because existing orders in $\mathcal{L}$ may be matched against incoming orders, the total volume posted to the book may be depleted over time, even in the absence of cancellations.

A book whose bid price exceeds the fair value of the product will be depleted on the bid side by traders seeking to sell the product at a premium to its true value. Conversely, an ask price below the fair value will be forced up by participants hoping

to buy at a discount. It is therefore common to regard the bid and ask as lower and upper bounds respectively on the consensus fair price of the product.

Motivated by this, the *midprice* is a naive point estimate for the consensus fair price, defined by

$$\text{mid}_{\mathcal{L}} = \frac{1}{2}\left(\text{bid}_{\mathcal{L}} + \text{ask}_{\mathcal{L}}\right).$$

Many other proxies for the consensus fair price exist that make greater use of information contained in $\mathcal{L}$, and it is common to use these as prediction targets in the construction of trading signals. I might discuss them later

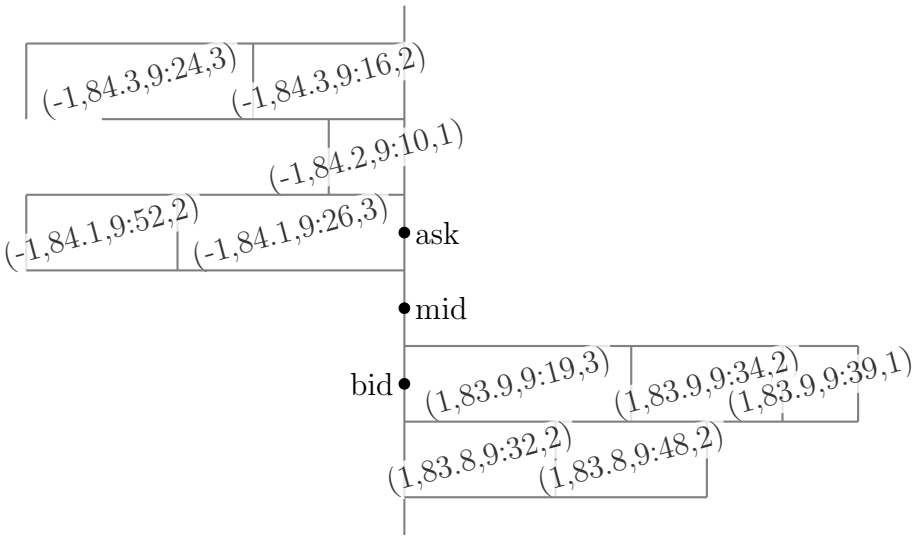


Figure 1.2: An example order book, arranged by order price and time

1.2.2 Liquidity

At any point in time, the contents of the limit order book represent trading opportunities presented to all market participants. The abundance of these opportunities, also known as *liquidity*, represents a positive externality insofar as it allows impatient traders (*liquidity takers*) to buy or sell products precisely under those circumstances where it is favourable to them. Conversely, order publishers (*liquidity providers*) must adhere to the terms of trades formed against a posted order, regardless of whether it is in their interests at the time the trade occurs.

One common measure for market liquidity is the *bid-ask spread*, defined as

$$\text{spread}_{\mathcal{L}} = \text{ask}_{\mathcal{L}} - \text{bid}_{\mathcal{L}}.$$

This can be described in units of price, but may also be described as a percentage of the midprice or as a multiple of the tick size (in which case we say the market is n ticks wide at some point in time). Notably, if $\text{mid}_{\mathcal{L}}$ is taken to represent the fair value of the product, then the *half-spread*, i.e. $\frac{1}{2}\text{spread}_{\mathcal{L}}$, represents the premium paid to liquidity providers by participants trading against the posted bid or ask. There are many other aspects of liquidity, including the quantity posted at each level, and the impact (instantaneous and permanent) of trades on the order book, which I may discuss later.

Adverse selection: large orders predict price moves

1.2.3 Queue Priority

When an order is submitted to the matching engine resulting in a trade, there may be some ambiguity about which existing orders it is to be matched against. For instance, consider the book displayed in figure 1.2, with a bid price of 83.9 and an ask price of 84.1, and suppose that a new buy order $(1, 84.2, 9:53, 2)$ arrives. Because the buy order has a price of 84.2, its addition to the book would raise the bid price to 84.2, exceeding the ask price of 84.1. Consequently, it needs to be matched against any of the sell orders with a price at least as favourable as 84.2, i.e. $(-1, 84.1, 9:26, 3)$, $(-1, 84.1, 9:52, 2)$, and $(-1, 84.2, 9:10, 1)$. It is not initially clear whether we should match two units against the first order, or against the second order, or match one unit against each of any two orders.

To eliminate this ambiguity, *queue priority* rules are set out that describe how to match incoming orders against opposing orders. In almost all cases, these rules obey *price priority*, where an existing order can only participate in a match if every order on the same side with a more competitive price has already been matched in full. This incentivises participants to submit orders with maximally competitive prices to increase the probability of a match, decreasing the bid-ask spread.

Beyond this, the two most common matching rules are known as *time priority* and *pro-rata*, which govern how matches are assigned in the case where orders have identical prices. Time priority requires that an order cannot participate in a match until all orders at the same price that have an earlier submission time have been fully matched. This incentivises early submission of orders and disincentivises cancellations. Pro-rata matching, on the other hand, seeks to allocate matches among orders with the same price in approximate proportion to the number of units in each order.

The ES and MES contracts traded on the Chicago Mercantile Exchange, which I will consider in which chapter, follow a pro-rata matching rule set out in citation. In order to determine a match from orders at a given price level, the size of each existing order is divided by the sum of all order sizes at that level, and this fraction is multiplied by the size of the incoming order and rounded down to the nearest integer. If the allocated trade quantity is less than two, it is rounded down to zero. After matching has taken place in this fashion, there may still be unmatched units in the incoming order, in which case they are matched against the remaining orders on the level according to time priority, and then if there are still unmatched units in the incoming order they may be matched against the next best price level or added to the order book.

1.3 Queueing Models of the Order Book

Changes in the price of a product over time may be viewed as emergent from individual order book events. For instance, changes to either the bid or ask price must result either from the addition of orders to the book or the depletion of a level due to cancellations or trades. From this perspective, the midprice is seen to follow a jump process governed by the evolution of the order book, which in turn is governed by the arrival of insertions or cancellations.

Analysing price changes from the perspective of event arrivals was originally (double check originality) proposed in cite the sante fe paper , where the arrival of order insertion and cancellation are taken to be governed by a Poisson process model. Despite the simplicity of their approach, they show that it replicates known empirical findings apparently unrelated to the model specification, such as the concavity of price impact (the expected change in midprice) as a function of trade size. This has come to be known as the Santa Fe model for limit order book evolution.

Chapter 5 of Bouchaud et. al. (2018) proposes queueing models for the top levels of the order book in a one-tick wide market. Under the assumption of Poisson processes for arrivals of insertion and cancellation requests at these levels (assuming identical rate parameters for the bid and ask), they Describe Bouchaud et al work [3]

Describe queue imbalance. It will be referred to later on

Evolution of the limit order book over time → motivate the emphasis on modeling arrival times

Price evolution as a jump process (what is a jump process) "Swishchuk and Huffman (2020) construct a compound Hawkes process" modeling price changes with a jump process where jump sizes are a markov chain. Jump process is controlled by one-dimensional point process. "Coinciding with the first preprint version of the present paper, Wu et al. (2019) develop a queue-reactive Hawkes process based on (2.4). In their model, X is endogenous and carries information about queue lengths in the LOB, while the multi-dimensional counting process driven by the intensity (2.4) models events pertaining to these queues. Wu et al. (2019) estimate their model on German bond (Bund) and index (DAX) futures LOB data." "Subsequently, Mounjid et al. (2019) generalise the queue-reactive Hawkes process to a more general point process framework that allows for non-linearity and quadratic Hawkes structure. Mounjid et al. (2019) additionally establish ergodicity for the model and also derive functional limit theorems for its long-term behaviour. They apply the model to evaluate and rank equities market makers on Euronext Paris."

Importance of arrival time modeling

`https://www.amazon.com/Point-Processes-Queues-Martingale-Statistics/dp/0387905`

Chapter 2

Point Process Models

With the goal of modelling the arrival time process, I will now provide an introductory overview of point processes, and then highlight some key variants that are relevant to order book modeling.

This section will make extensive use of theoretical concepts described in the appendix Make sure to update this if there ends up being more than one appendix. For an overview of measure theory, probability spaces, and stochastic processes, please refer to the relevant sections.

2.1 Overview of Point Processes

Given a probability space $(\Omega, \Sigma, \mathbb{P})$, and a measurable space (T, Σ_T) representing times, a point process is any increasing sequence of random times

$$\mathcal{T} : \Omega \to T^{\mathbb{N}},$$

meaning that $\mathcal{T}(\omega)_n$ is increasing in n for any $\omega \in \Omega$.

Concretely, T may be chosen to be $\mathbb{R}_{\geq 0}$, and equipped with the Borel σ-algebra $B(\mathbb{R}_{\geq 0})$.

The times in the point process are often referred to as *event times*, with the implication that the sequence represents the times at which some event of interest occurs (for instance, the arrival of messages sent to a matching engine).

For any point process, there exists a corresponding càdlàg stochastic process $N_{\mathcal{T}} : \Omega \times T \to \mathbb{N}$ known as the *counting process*, that gives the number of events having occurred before or at a given time. This is defined as

$$N_{\mathcal{T}}(\omega, t) = |\{i \in \mathbb{N} | \mathcal{T}(\omega)_i \leq t\}|.$$

Overlay barcode plot for trade times with corresponding counting process

All (or perhaps many) of the point processes considered in this thesis will additionally be adapted with respect to some filtration $\mathcal{F}$ indexed by T. In what sense?

Furthermore, I will only consider *nonexplosive point processes*, defined as those point processes $\mathcal{T}$ for which

$$\lim_{n \to \infty} \mathcal{T}(\omega)_n = \infty, \mathbb{P}\text{-a.s.}$$

For any bounded interval of time $(t_{\min}, t_{\max})$, a nonexplosive point process will almost surely contain only a finite set of times in that interval, i.e.

$$(t_{\min}, t_{\max}) \cap \mathcal{T}(\omega) \text{ finite}, \quad \mathbb{P}\text{-a.s.}$$

2.1.1 The Counting Measure

For any $\omega \in \Omega$, we can define the *counting measure* Λ_ω of the point process as

$$\Lambda_\omega : \Sigma_T \to \mathbb{N}$$

$$\Lambda_\omega(S) \coloneqq |S \cap \mathcal{T}(\omega)| .$$

For any finite or countable collection of disjoint sets $A_n \in \Sigma$, we have

$$\Lambda_\omega \left(\bigcup_n A_n \right) = \left| \bigcup_n (A_n \cap \mathcal{T}(\omega)) \right| .$$

Since subsets of disjoint sets are also disjoint, the terms in the union on the right-hand side will be disjoint, and so

$$\Lambda_\omega \left(\bigcup_n A_n \right) = \sum_n |A_n \cap \mathcal{T}(\omega)| = \sum_n \Lambda_\omega(A_n).$$

Thus Λ_ω is a measure on (T, Σ_T).

2.1.2 The Expectation Measure

Taking the expectation of the counting measure with respect to some measure $\mathbb{P}$ on a measurable space (Ω, Σ) gives the *expectation measure*,

$$\Lambda_{\mathbb{P}}(S) \coloneqq \mathbb{E}_{\mathbb{P}} \left[\Lambda_\omega(S) \right] = \int_\Omega \Lambda_\omega(S) d\mathbb{P}(\omega).$$

Since $\Lambda_\omega(S)$ is a non-negative function of ω, it follows that

$$\Lambda_{\mathbb{P}} : \Sigma_T \to \mathbb{R}_{\geq 0} \cap \{\infty\}.$$

Furthermore, for any finite or countable sequence of $A_n \in B(\mathbb{R})$, we have

$$\Lambda_{\mathbb{P}} \left(\bigcup_n A_n \right) = \int_\Omega \Lambda_\omega \left(\bigcup_n A_n \right) d\mathbb{P}(\omega) = \int_\Omega \sum_n \Lambda_\omega(A_n) d\mathbb{P}(\omega),$$

by the countable additivity of Λ_ω. Thus $\Lambda_{\mathbb{P}}$ is a measure on (T, Σ_T).

To handle the sum inside the integral, we can write each term as a sum of indicator functions:

$$\Lambda_\omega(A_n) = |A_n \cap \mathcal{T}(\omega)| = \sum_i 1_{A_n} \left(\mathcal{T}(\omega)_i \right) .$$

Then, since A_n are measurable sets, the indicator functions are each measurable functions of ω. Because finite sums and pointwise limits of measurable functions

are measurable, we have that $\Lambda_\omega(A_n)$ is measurable. Finally, since each term in the sum over n is a nonnegative measurable function, the integral commutes with the sum, and hence $\Lambda_{\mathbb{P}}$ is countably additive. Therefore the expectation measure is also a valid measure.

2.1.3 Intensity of a Point Process

For a point process $\mathcal{T}$, the counting measure of a half-open interval $(t, t+\epsilon]$ can be written in terms of the counting process $N_{\mathcal{T}}$ in the form

$$\Lambda_\omega\left((t, t+\epsilon]\right) = N_{\mathcal{T}}(\omega, t+\epsilon) - N_{\mathcal{T}}(\omega, t).$$

Integrating out ω with respect to $\mathbb{P}$ then gives us the identity

$$\Lambda_{\mathbb{P}}\left((t, t+\epsilon]\right) = \mathbb{E}_{\mathbb{P}}\left[N_{\mathcal{T}}(t+\epsilon)\right] - \mathbb{E}_{\mathbb{P}}\left[N_{\mathcal{T}}(t)\right],$$

allowing us to write the expectation measure in terms of a finite difference of the first moment of $N_{\mathcal{T}}$.

If $\mathbb{E}_{\mathbb{P}}\left[N_{\mathcal{T}}(t)\right]$ is differentiable at t, we can further say that

$$\lim_{\epsilon\to 0} \frac{\Lambda_{\mathbb{P}}\left((t, t+\epsilon]\right)}{\epsilon} = \lim_{\epsilon\to 0} \frac{\mathbb{E}_{\mathbb{P}}\left[N_{\mathcal{T}}(t+\epsilon)\right] - \mathbb{E}_{\mathbb{P}}\left[N_{\mathcal{T}}(t)\right]}{\epsilon} = \frac{d\mathbb{E}_{\mathbb{P}}\left[N_{\mathcal{T}}(t)\right]}{dt}.$$

For brevity, I will write

$$\lambda_{\mathbb{P}}(t) \coloneqq \frac{d\mathbb{E}_{\mathbb{P}}\left[N_{\mathcal{T}}(t)\right]}{dt}.$$

Assuming further that $\mathbb{E}_{\mathbb{P}}\left[N_{\mathcal{T}}(t)\right]$ is differentiable on some open interval $(t, t+\epsilon')$, Prove this is Radon-Nikodym derivative of $\Lambda_{\mathbb{P}}$ in the open interval. I will therefore refer to this quantity as the *expectation density*.

It then follows from the fundamental theorem of calculus that

$$\int_t^{t+\epsilon} \lambda_{\mathbb{P}}(s)ds = \lambda_{\mathbb{P}}(t)\cdot\epsilon + o(\epsilon), \quad \epsilon \to 0.$$

So since $\lambda_{\mathbb{P}}$ is a radon-nikodym derivative we can write Why?

$$\Lambda_{\mathbb{P}}\left((t, t+\epsilon]\right) = \lambda_{\mathbb{P}}(t)\cdot\epsilon + o(\epsilon), \quad \epsilon \to 0$$

Therefore the expected number of events occurring in a small interval $(t, t+\epsilon]$ is approximately proportional to its length.

If we instead take the expectations above conditioned on $\mathcal{F}_t$, we can define a stochastic process

$$\lambda(\omega, t) \coloneqq \lim_{s\to t^-} \lambda_{\mathbb{P}|\mathcal{F}_s}(\omega, t)$$

as the left limit of the expectation density as the filtration index approaches t. This is commonly known as the *intensity* or *arrival rate* of the point process, and represents the expected number of events that will arrive in the next ϵ units of time, for very small ϵ, based on the information contained in the filtration up to time t.

Show that $\mathbb{E}_{\mathbb{P}|\mathcal{F}_s}[\lambda(\omega, t)] = \lambda_{\mathbb{P}|\mathcal{F}_s}(t)$ for $s < t$

Add a note somewhere to check the whole document for correct usage of left vs right limit of λ. This is very important.

Existence and uniqueness shown in https://projecteuclid.org/journals/annals-of-probability/volume-24/issue-3/Stability-of-nonlinear-Hawkes-processes/10.1214/aop/1065725175 under restrictions on the kernel

Establish that the intensity of a point process uniquely determines its distribution, so if two have the same then theyre distributionally identical

2.1.4 Residuals and the Compensator Process

Given some point process $\mathcal{T}$ with intensity λ, we may define the stochastic process

$$\bar{\Lambda}(\omega, t) = \int_{t_{\min}}^{t} \lambda(\omega, s) ds,$$

known as the *compensator process* of $\mathcal{T}$. Since λ is adapted to some filtration $\mathcal{F}$, it follows why that $\bar{\Lambda}$ will be similarly adapted to $\mathcal{F}$.

Since λ is positive what if its zero?, any realisation of $\bar{\Lambda}$ must be increasing in t. The random sequence $\tau(\omega) = \left\{\bar{\Lambda}(\omega, \mathcal{T}(\omega)_i)\right\}_{i \in \mathbb{N}}$ will therefore be increasing for each ω, making it a point process. Theorem 7.4.1 from Daley and Vere-Jones (2003) [5] demonstrates that regardless of the initial point process $\mathcal{T}$, the transformed process τ is a Poisson process with unit rate, implying that the quantities

$$r_i(\omega) \coloneqq \int_{\mathcal{T}(\omega)_i}^{\mathcal{T}(\omega)_{i+1}} \lambda(\omega, t) dt = \tau(\omega)_{i+1} - \tau(\omega)_i$$

for each $i \in \mathbb{N}$ are independent and follow an Exp(1) distribution.

These are commonly known as the *residuals* of the point process, and serve as a common diagnostic tool for evaluating model quality. Deviation from Exp(1) or serial dependence between residuals are common signs of model misspecification or misestimation, and may give qualitative hints about how to improve the fit of a model to empirical data.

2.1.5 Marked and Multivariate Point Processes

A point process T may be *marked*, in which case it is associated with one or more random sequences of marks

$$\mathcal{X} : \Omega \to X^{\mathbb{N}},$$

drawn from a set X, that represent additional information about each event. These may be adapted to the filtration $\mathcal{F}$ In what sense?

A common special case of this is when X is partitioned into a set E of *event types*, and define a random sequence $\mathcal{E} \in E^{\mathbb{N}}$ whose ith entry is the event type of $\mathcal{X}_i$. In this case we may refer to $(\mathcal{T}, \mathcal{E})$ as a *multivariate point process*.

Each part e of the partition then has an associated point process

$$\mathcal{T}_e = \mathcal{T}_{\{n \in \mathbb{N} : \mathcal{X}_n \in e\}},$$

formed by the subsequence of times where the corresponding element of $\mathcal{X}$ is in e. These point processes will then have their own counting functions, intensities, and

other characteristics. It is common to arrange these quantities in vector form, with one entry for each event type. For instance, the counting function of a multivariate point process will be a function $N_{\mathcal{T}} : \Omega \times T \to \mathbb{N}^{|E|}$.

On the other hand, given a collection of point processes adapted with respect to a common filtration, we can form a single marked point process by interleaving the sequences of event times. The counting process, counting measure, expectation measure, and intensity of the combined process will be the sum of those for the individual point processes. Can i relate this to composite point processes

Barcode plot and counting process for buys and sells, and the combined process

2.1.6 Composite Point Processes

Suppose that some multivariate point process $\mathcal{T}$ has an intensity $\lambda(\omega, t)$ that can be written in the form

$$\lambda(\omega, t) = \sum_i \lambda_i(\omega, t)$$

for some collection of vector-valued stochastic processes λ_i, and define the *component probabilities* as vector-valued stochastic processes P_i having elements

$$P_i^e(\omega, t) = \frac{\lambda_i^e(\omega, t)}{\lambda_e(\omega, t)},$$

Note that if we were to choose some of the λ_i to be occasionally negative, the P_i^e for any particular event type e would not necessarily form a true probability distribution over the i. Nonetheless, we will always have

$$\sum_i P_i^e(\omega, t) = 1.$$

If we further suppose that the $\lambda_i(\omega, t)$ happen to be the intensity functions of a collection of point processes $\mathcal{T}_i$, then we will have that the set of marked event times resulting from a union of all the $\mathcal{T}_i$ is distributionally identical to the original point process $\mathcal{T}$. In particular, if we assume that the $\mathcal{T}_i$ are pairwise disjoint, we will have

$$\mathbb{P}((t, e) \in \mathcal{T}_i(\omega) | \mathcal{F}_t, (t, e) \in \mathcal{T}(\omega)) = \mathbb{E}_{\mathbb{P}} \left[P_i^e(\omega, \mathcal{T}(\omega)_k) | \mathcal{F}_t \right].$$

is this definitely true in general?

In this case, we refer to $\mathcal{T}$ as a *composite point process*, and to $\mathcal{T}_i$ as the *components*. Any dataset $\mathcal{D}$ consisting of event times and marks may be augmented by attributing each event time to the component containing it, with this augmented dataset referred to as the *complete data*. This term is commonly used in an expectation-maximisation context to denote the augmentation of data by some choice of values for the latent variables, and I will describe the use of an EM-style algorithm for estimation of composite point processes in the chapter on estimation. In this case, the unique component containing each event time is a latent variable.

2.2 Point Process Models of Book Updates

For each extension: - show some example realisations for various parameter choices: barcode and intensity plot

2.2.1 Poisson Processes

The simplest modeling assumption we can make is to assume that the intensity of the process is a constant $\nu \in \mathbb{R}$. A point process T whose intensity λ is given by some positive constant ν is known as a Poisson process. Equivalently, if $\mathcal{T}$ has counting measures Λ_ω, we say that it is a Poisson process with rate ν if and only if the following two properties hold:

1. For any finite collection of disjoint measurable sets $A_n \in \Sigma_T$, their counting measures $\Lambda_\omega(A_n)$ are independent random variables.
2. For any measurable set $A \in \Sigma_T$, we have

$$\Lambda_{\mathbb{P}}(A) = \nu \mu_{\text{Lebesgue}}(A) \propto \mu_{\text{Lebesgue}}(A).$$

Why is this the same as saying that the intensity is constant? The random variables $\mathcal{T}(\omega)_{i+1} - \mathcal{T}(\omega)_i$ for each $i \in \mathbb{N}$ will then be independently $\text{Exp}(\nu)$ distributed, with a mean of $\frac{1}{\nu}$. Is there a quick proof of this

By Markov's inequality, we know that

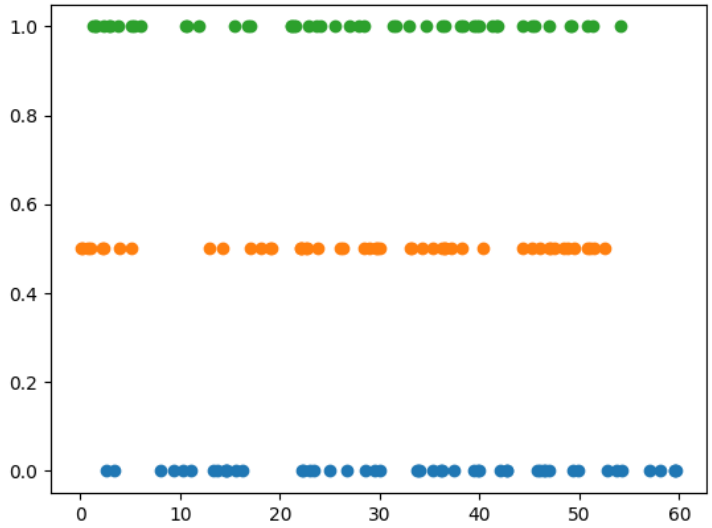
$$\begin{aligned}\mathbb{P}\left(\mathcal{T}(\omega)_k \leq k^2\right) &= \mathbb{P}\left(-\mathcal{T}(\omega)_k \geq -k^2\right) \\ &\leq \frac{\mathbb{E}_{\mathbb{P}}[\mathcal{T}(\omega)_k]}{k^2} = \frac{\mathbb{E}_{\mathbb{P}}[\sum_{j=0}^{k-1} \mathcal{T}(\omega)_{j+1} - \mathcal{T}(\omega)_j]}{k^2} = \frac{k}{k^2\nu} = \frac{1}{k\nu},\end{aligned}$$

from which it follows that

$$\lim_{n\to\infty} \mathcal{T}(\omega)_n = \infty, \mathbb{P}\text{-a.s.},$$

and hence the Poisson process is non-explosive.

In the multivariate case with n event types, we have a vector $\nu \in \mathbb{R}^n_{>0}$ containing the constant intensities of each individual point process.



I think I should change this to be separate plots. Can show the intensity function for each; e.g. for inhomogenous it's the same for every realisation, but for hawkes process the intensity function is itself random.

2.2.2 Inhomogeneous Poisson Process

An inhomogeneous Poisson process is a generalisation of the Poisson process for which the second condition is replaced by the weaker requirement that every measurable set $A \in \Sigma_T$ with finite Lebesgue measure $\mu_{\text{Lebesgue}}(A)$ has a finite expectation measure, and every A satisfying $\mu_{\text{Lebesgue}}(A) = 0$ also satisfies $\Lambda_{\mathbb{P}}(A) = 0$.

By the Radon-Nikodym theorem, since $\Lambda_{\mathbb{P}}$ is absolutely continuous with respect to μ_{Lebesgue}, the expectation density $\lambda_{\mathbb{P}}$ exists. Furthermore, because the counting measure on $(t, t+\epsilon]$ is independent from the counting measure on any disjoint set, we have that $\lambda_{\mathbb{P}}(\omega, t)$ is adapted to the same filtration $\mathcal{F}$ as the counting process, from which it follows that the intensity process is simply $\lambda = \lambda_{\mathbb{P}}$ and is therefore constant with respect to ω.

Since λ is deterministic, the compensator process $\bar{\Lambda}$ will also be deterministic (i.e. a function of t only), which allows us to simulate an inhomogenous point process by taking the random sequence $\left\{\bar{\Lambda}^{-1}(\tau(\omega)_i)\right\}_{i\in\mathbb{N}}$ for some Poisson process τ, so long as the inverse function $\bar{\Lambda}^{-1}$ exists (i.e. $\bar{\Lambda}$ is strictly increasing).

Deterministic variation in λ as a function of t is useful for modeling deterministic seasonality in the point process, such as having more events near the start and end of each trading session, or near known events such as news releases.

Show U-shape from empirical data Show U-shape from quadratic IPP

2.2.3 Hawkes Processes

In the case of the homogenous and inhomogenous Poisson processes, the intensity function is deterministic, i.e. constant in ω. While such an assumption is appropriate for modeling the arrival of independent events, it is common in financial markets for events to "cluster" together in a non-deterministic fashion, with the intensity dependent on the recent historical behaviour of the point process. One

explicit test for this phenomenon is to examine the autocorrelation of residuals of an inhomogenous Poisson process model, which will be covered in some section of the applications chapter. Show ACF of inter-arrival times

The increase in intensity seen after a series of related events in quick succession is known as *self-excitation*[1], and requires us to introduce a stochastic component to the intensity of our model that will depend on the realised history of the point process up until t. By analogy with time series literature, the dependence of forecasts on recent history is referred to as *autoregressive*.

A popular class of autoregressive models are known as *Hawkes Processes*, first introduced by Original hawkes process paper https://academic.oup.com/biomet/article-abstract/58/1/83/224809?redirectedFrom=fulltext Discuss history of their application to financial datasets

Univariate Hawkes process models require that the intensity process have the functional form

$$\lambda(\omega, t) = \nu + \int_{t_{\min}}^{t} k(t-s) d\Lambda_\omega(s),$$

where $\nu \in \mathbb{R}_{>0}$ represents a constant arrival rate, k is a kernel function encoding the impact of each event on the future intensity, and $t_{\min}$ is the earliest time events can occur, such as the opening time of a trading session. I will assume that k takes positive values for nonnegative arguments, and is zero for negative arguments. Existence and nonexplosiveness of such processes is established... where? Integral is open on the right bound of the interval of integration

In the case of a multivariate point process $(\mathcal{T}, \mathcal{E})$, we can generalise this functional form to

$$\lambda_e(\omega, t) = \nu_e + \sum_{e'} \int_{t_{\min}}^{t} k_{e',e}(t-s) d\Lambda_\omega^{e'}(s),$$

where $k_{e',e}$ is the kernel associated with the pair of event types $(e', e) \in E^2$. For convenience, we can arrange the $\lambda_e(\omega, t)$ and Λ_ω into $|E|$-dimensional vector-valued functions $\lambda(\omega, t)$ and Λ_ω. Similarly, the base rates ν_e can be arranged into a vector ν, and the functions $k_{e',e}$ into a matrix-valued function $k(t-s)$. Then finally we have

$$\lambda(\omega, t) = \nu + \int_{t_{\min}}^{t} k(t-s)^T d\Lambda_\omega(s).$$

Let $K(t) = \int_0^t k(u) du$.

Since λ can't take on negative values, the kernel function must be positive. Show also that it must have finite L1 norm

It is possible to generate samples from a multivariate Hawkes process $(\mathcal{T}, \mathcal{E})$ in the following manner. Firstly, draw samples from a multivariate Poisson process $\mathcal{T}'$ with rate ν, and let τ_0 be the resulting sequence of pairs $(t, e) \in T \times E$, ordered by time. For each e, we have the function $K_{e_i,e}$ and its inverse function $K_{e_i,e}^{-1}$. Draw

[1] In the case of multivariate point processes, interactions between different event types are often called *cross-exciting*. The opposite behaviour, where recent occurrences temporarily decrease the intensity, is known as *self-inhibition* or *cross-inhibition*.

samples from an independent Poisson process π_i^e with unit rate until one of the times exceeds $K_{e_i,e}(\infty)$. Following this, let

$$g_i^e = \{(K_{e_i,e}^{-1}(t) + \tau'_{i,i}, e) | t \in \pi_i^e\},$$

where $\tau'_{i,i}$ is the time component of the pair $\tau_{i,i}$. Lastly, let

$$\tau_{i+1} = \tau_i \cup \bigcup_{e \in E} g_i^e.$$

Then the sequence of pairs $\{\tau_{i,i}\}_{i \in \mathbb{N}}$ will be distributed in accordance with the desired multivariate Hawkes process. Proof? or source? Immigration-birth interpretation [13] (aka Watson-Galton models). Introduce the term *parent event* Stability criterion. I think it's det $K < 1$. but how to prove. have $\bar{\lambda} = (I - K)^{-1}\nu$.

For computational simplicity, it is often convenient to use a kernel of the form

$$k_{e',e}(u) = \sum_i \alpha_{e',e,i} \exp(-\beta_{e',e,i} u),$$

for some real numbers $\alpha_{e',e,i}$ and positive real numbers $\beta_{e',e,i}$. Suppose that $k_{e',e}$ has some different functional form, but is still continuous. The Weierstrass approximation theorem guarantees that every continuous function on some closed interval can be uniformly approximated as closely desired by polynomials. Since polynomials in $\exp(-u)$ with no constant term can be written in the above form, and this is a continuous bijective function of u, it follows that any continuous function on a closed interval can be approximated arbitrarily well by a kernel of this form and a constant term. Since the constant term is arbitrarily well approximated by very small values of $\beta_{e',e,i}$, we can approximate any other continuous kernel function arbitrarily well by a linear combination of decreasing exponentials. This proof needs more detail. Use triangle inequality

Explain why it is better computationally

According to `https://ieeexplore.ieee.org/document/7416001` `https://arxiv.org/pdf/`
hawkes processes are fully determined by the first two moments of the intensity function. Proof in the paper.

2.2.4 State Dependence

Just as the arrivals of events in the limit order book governs the evolution of its state over time, we might reasonably expect that the arrival rate of certain kinds of events is dependent on the present state of the order book, as captured by a variety of features. Vinkovskaya (2014) [**?**] proposed a regime-switching variant of the Hawkes process model for which the kernel function additionally depends on whether the market was one tick wide at the time of the parent event, finding a greater self-excitation effect for the various event types in the one-tick-wide regime.

Moriaru-Patrichi and Pakkanen (2022) [**?**] find similarly that a regime-switching variant learns different parameters for the one-tick-wide regime, as well as in a model using five different states based on queue imbalance. Their work is distinguished from prior regime-switching approaches in the sense that they also model the evolution of the order book state over time as a finite state-switching process,

with transition probabilities at each event time also dependent on the event type. While this coupling of a point process model with state-switching behaviour admits broader application than simple modeling of arrival times, this is not the primary focus of the work, and the simple nature of their state model leaves room for further extension.

The model of [**?**] takes the form

$$\lambda_e(\omega, t) = \nu_{\mathcal{X}(\omega,t),e} + \sum_{e' \in E} \int_{t_{\min}}^{t} k_{\mathcal{X}(\omega,s),e',e}(t-s) d\Lambda_{\omega}^{e'}(s), \tag{2.2.1}$$

where $\mathcal{X}(\omega, t)$ is a discrete-valued stochastic process representing the order book state, with jumps only at the event times of the point process.

Notably, the focus of prior work has been primarily on the case with a finite number of discrete states, possibly obtained by discretisation of a continuous quantity such as queue imbalance. Consideration of multiple state variables quickly becomes difficult, as the number of states grows exponentially in the number of independent state variables. No prior information is taken into account regarding the monotonicity of excitation with respect to continuous state variables, and no extrapolation can take place on this basis.

One deficiency of this particular model is the lack of dependence on the order book state at time t. Information about states is only taken into account at the times of the parent events, making the long-term effects of previous events independent of the present state. To address this, I propose a more general model of the form

$$\lambda_e(\omega, t) = \nu_{\mathcal{X}(\omega,t),e} + \sum_{e' \in E} \int_{t_{\min}}^{t} k_{\mathcal{X}(\omega,s),\mathcal{X}(\omega,t),e',e}(t-s) d\Lambda_{\omega}^{e'}(s).$$

Assuming that for each (e', e) the kernel function $k_{e',e}$ is a sum of terms $k_{e',e,i}$ that are separable in $\mathcal{X}(s)$, $\mathcal{X}(t)$, and $t - s$, explain why this assumption still lets us approximate any true situation with enough component kernels, we obtain

$$\lambda_e(\omega, t) = \nu_{\mathcal{X}(\omega,t),e} + \sum_{i} \sum_{e' \in E} \int_{t_{\min}}^{t} A_{e',e,i}(\mathcal{X}(\omega, s)) B_{e',e,i}(\mathcal{X}(\omega, t)) k_{e',e,i}(t-s) d\Lambda_{\omega}^{e'}(s). \tag{2.2.2}$$

In the case where A and B are finite-valued, the state space X is partitioned into a variety of discrete states. If we allow $k_{e',e,i}$ to take the same functional form as the state-dependent kernel of equation 2.2.1, then a model of the form 2.2.2 with one component for every state nests a model of the form 2.2.1 by making $B_{e',e,i}$ a positive constant, and setting $A_{e',e,i}$ to 1 if its argument is in the ith state or 0 otherwise.

A further generalisation would write A and B as continuous functions of the order book state. A convenient choice is the form $\exp(c \cdot \chi(\mathcal{X}(\omega, t)))$, where $\chi(\mathcal{X}(\omega, t))$ is a feature vector derived from the full order book state $\mathcal{X}(t)$ and c is a coefficient vector of the same size. This ensures that A and B are necessarily positive and captures monotonic effects of the state variables on the intensity, while still being flexibly parameterised. If $k_{e',e,i}$ is allowed to be zero, then introducing one component per

state, we can replicate the finite-state model by letting χ be a vector with one element for each state, that is one at the entry corresponding to the present state and zero elsewhere.

Weierstrass approximation thm again applies: with enough components, any continuous functional form for A and B can be uniformly well approximated on an arbitrarily large interval

Emphasise that this is novel

2.2.5 Raising the Intensity to a Power

Though the Hawkes process intensity manages to capture self-exciting feedback loops in the behaviour of the point process, it may be the case that the response to a series of events is better captured by some nonlinear function. For instance, it may be the case that two identical events in close succession should contribute more or less than double to the intensity what a single event might contribute on its own, a situation that is not captured by the model as described so far.

One way to induce nonlinearity in the response function is to raise it to some positive real power κ, obtaining

$$\lambda_e(\omega, t) = \nu_e + \left(\sum_{e' \in E} \int_{t_{\min}}^{t} A_{e',e}(\mathcal{X}(\omega, s)) B_{e',e}(\mathcal{X}(\omega, t)) k_{e',e}(t-s) d\Lambda_\omega^{e'}(s) \right)^{\kappa}.$$

More generally, we might have a linear combination of such terms, writing

$$\begin{aligned} \lambda_{e',e}^{i}(\omega, t) &= \int_{t_{\min}}^{t} A_{e',e,i}(\mathcal{X}(\omega, s)) B_{e',e,i}(\mathcal{X}(\omega, t)) k_{e',e,i,i}(t-s) d\Lambda_\omega^{e'}(s), \\ \lambda_e^i(\omega, t) &= \sum_{e' \in E} \lambda_{e',e}^{i}(\omega, t), \\ \lambda_e(\omega, t) &= \nu_e + \sum_i \pm_i \lambda_e^i(\omega, t)^{\kappa_i}. \end{aligned} \tag{2.2.3}$$

I made the background rate state-independent. Explain why I can choose very small κ to get the same result as state-dependence. By taking κ_j to be integers, and supposing that the excitation components λ_e^j are scalar multiples of λ_e^0, we see that this nests a more restricted model where the intensity is given by some some polynomial function of λ_e^0. By the Weierstrass approximation theorem, we are therefore able to uniformly approximate an intensity formed by any continuous function of λ_e^0, so long as λ_e^0 remains within some arbitrarily large interval. In order to apply gradient-based methods in estimating κ_j, however, it is more convenient to let it be any positive real number. but what to do if the base is negative? well, just dont let it be, force alpha to be nonnegative

Explain why this nests the quadratic hawkes process "We introduce and establish the main properties of QHawkes ("Quadratic" Hawkes) models. QHawkes models generalize the Hawkes price models introduced in E. Bacry et al. (2014), by allowing all feedback effects in the jump intensity that are linear and quadratic in past returns. A non-parametric fit on NYSE stock data shows that the off-diagonal component of the quadratic kernel indeed has a structure that standard Hawkes

models fail to reproduce. Our model exhibits two main properties, that we believe are crucial in the modelling and the understanding of the volatility process: first, the model is time-reversal asymmetric, similar to financial markets whose time evolution has a preferred direction. Second, it generates a multiplicative, fat-tailed volatility process, that we characterize in detail in the case of exponentially decaying kernels, and which is linked to Pearson diffusions in the continuous limit. Several other interesting properties of QHawkes processes are discussed, in particular the fact that they can generate long memory without necessarily be at the critical point. Finally, we provide numerical simulations of our calibrated QHawkes model, which is indeed seen to reproduce, with only a small amount of quadratic non-linearity, the correct magnitude of fat-tails and time reversal asymmetry seen in empirical time series."

computational details covered in estimation section

Chapter 3

Estimation

Selecting a model from some parametric family of point process models based on empirical order book data serves two primary purposes. Firstly, estimating the parameters of a sufficiently parsimonious and interpretable model serves to provide direct insight into the relationship between various order book events and states, such as by identifying the sign of various parameters, comparing the magnitude of parameters to one another, and performing hypothesis tests to see if a parameter is significantly different from some default value. From this we may tentatively draw qualitative conclusions about the true behaviour of the data generating process governing market behaviour. Secondly, by estimating both a point process model and a model for state transitions at each event time, we can identify a model that will allow us to simulate synthetic order book data in a manner reasonably consistent with actual market behaviour. Repeated samples from a simulation of true market behaviour may be used to predict the evolution of the order book from some initial market state, or to answer counterfactual questions such as the effect of placing a trade with some quantity at some point in time in the same way that the Santa Fe model implies a price impact model in the paper that proposed it.

In this thesis I will focus on maximum likelihood estimation from parametric families, using a quasi-Newton variant of the EM algorithm that updates parameters iteratively based on the gradient of the loglikelihood and the diagonal of its hessian. is this accurate

3.1 Empirical Likelihood

I will work with a dataset of the form $\mathcal{D} = (\mathcal{T}_{\text{obs}}, \mathcal{X}_{\text{obs}})$, consisting of a sequence of event times and event marks observed in the time interval

$$[t_{\min}, t_{\max}].$$

We can define a counting measure $\Lambda_{\mathcal{D}}$ on the space of events times (T, Σ_T) by

$$\Lambda_{\mathcal{D}} : \Sigma_T \to \mathbb{R}_{\geq 0}$$

$$\Lambda_{\mathcal{D}}(A) \coloneqq |A \cap \mathcal{T}_{\text{obs}}| \,.$$

Similarly, we can define a càdlàg counting function

$$N_{\mathcal{D}} : T \to \mathbb{N}$$

$$N_{\mathcal{D}}(t) \coloneqq |\mathcal{T}_{\text{obs}} \cap [t_{\min}, t]| = \int_{[t_{\min},t]} d\Lambda_{\mathcal{D}}.$$

In the multivariate case we will further have $\Lambda^e_{\mathcal{D}}$ and $N^e_{\mathcal{D}}$ for each event type e.

Given a subset of the data $\mathcal{D}_t$ containing only the event times and marks up until some time t, it is reasonable to ask whether we can predict the remainder of the dataset using some model of the point process that generated the data.

3.1.1 Empirical Intensity and Loglikelihood

If we consider a point process $\mathcal{T}$ defined with respect to $\mathbb{P}$ and having natural filtration $\mathcal{F}$, we can construct an *empirical intensity function*

$$\lambda_{\mathcal{D}}(t) \coloneqq \lambda_{\mathbb{P}|\mathcal{F}_t,\mathcal{D}_t}(t)$$

as the intensity conditional on the σ-algebra $\mathcal{F}_t$ and the contents of $\mathcal{D}_t$. For the point processes considered in this thesis, we have formulae representing $\lambda(\omega, t)$ as an expression involving t, $\mathcal{X}$, and the counting measure Λ_ω restricted to $[t_{\min}, t)$. In this setting, it suffices to replace Λ_ω with $\Lambda_{\mathcal{D}}$ in order to compute the empirical intensity function, from which we can proceed to compute the loglikelihood of some model with reference to the event times (conditional on the marks), given in equation 7.1.2 of [5] as

$$\ell_{\mathcal{D}} = \int_{t_{\min}}^{t_{\min}} \log\left(\lambda_{\mathcal{D}}(t)\right) d\Lambda_{\mathcal{D}}(t) - \int_{t_{\min}}^{t_{\min}} \lambda_{\mathcal{D}}(t) dt. \tag{3.1.1}$$

In the case of a multivariate point process, the empirical intensity function will be a vector with elements $\lambda^e_{\mathcal{D}}(t)$, and the loglikelihood will be given by

$$\ell_{\mathcal{D}} = \sum_e \left(\int_{t_{\min}}^{t_{\min}} \log\left(\lambda^e_{\mathcal{D}}(t)\right) d\Lambda_{\mathcal{D}}(t) - \int_{t_{\min}}^{t_{\min}} \lambda^e_{\mathcal{D}}(t) dt \right). \tag{3.1.2}$$

maybe point out that even defining what loglikelihood means for a point process is nontrivial and refer to the relevant section of [5] for more detail If we have access to multiple empirical datasets $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_n$ representing independent realisations of some underlying data generating process, the loglikelihood of the combined dataset $\mathcal{D}_{\text{all}}$ is simply $\sum_{i=1}^n \ell_{\mathcal{D}_i}$.

3.2 Maximum Likelihood Estimation

Given a parametric family of marked multivariate point processes,

$$\{(\mathcal{T}_\theta, \mathcal{X}) : \theta \in \Theta\},$$

we wish to select one model from the family that best describes the empirical data.

Because the likelihood function tells us the probability density of the observed data under the corresponding model, a higher likelihood function corresponds to a better fit of model to data. Assuming that the likelihood has a unique global maxima

$$\hat{\theta}_{\text{MLE}}(\mathcal{D}_{\text{all}}) \coloneqq \operatorname*{argmax}_\theta L(\theta; \mathcal{D}),$$

this quantity is known as the *maximum likelihood estimate* for θ, and is a popular method of selecting a single "best" model from a parametric family.

For any point process $\mathcal{T}_\theta$ in the family, as well as some model for the marks $\mathcal{X}$, we can construct a synthetic dataset $\mathcal{D}'$. Then $\hat{\theta}_{\text{MLE}}(\mathcal{D}_{\text{all}})$ is a random variable.

As the number of independent realisations n grows, we have the classical result that

$$\sqrt{n}\left(\hat{\theta}_{\text{MLE}}(\mathcal{D}_{\text{all}}) - \theta\right) \to_{\mathbb{P}} \mathcal{N}(0, \mathcal{I}(\theta)^{-1}),$$

where $\mathcal{I}(\theta)$ is the Fisher information matrix. Furthermore, it is demonstrated in [19], that even for a single realisation of a point process model, the MLE is consistent and asymptotically normal as the size of the observation interval $[t_{\min}, t_{\max}]$ grows, in the sense that

$$\sqrt{t_{\max} - t_{\min}}\left(\hat{\theta}_{\text{MLE}}(\mathcal{D}_{\text{all}}) - \theta\right) \to_{\mathbb{P}} \mathcal{N}(0, \mathcal{I}(\theta)^{-1}),$$

maybe double check this under a permissive set of assumptions satisfied by all the models considered in this thesis.

Because finding global maxima is in general a hard problem, it is common to first weaken our search to finding stationary points of $\ell_{\mathcal{D}_{\text{all}}}(\theta)$ using the gradient vector $\nabla_\theta \ell_{\mathcal{D}_{\text{all}}}(\theta)$. We can then check whether the hessian matrix $H_\theta \ell_{\mathcal{D}_{\text{all}}}(\theta)$ satisfies the requirement of negative definiteness, eliminating local minima and saddle points. Finally, we can select the best local maxima among those found in the hope that either it is the global maxima, or that it has a similar enough likelihood to the true $\hat{\theta}_{\text{MLE}}$ to be a good model choice.

Note that the gradient and hesssian of $\ell_{\mathcal{D}_{\text{all}}}$ will be the sum of the gradients (respectively, hessians) for each $\ell_{\mathcal{D}_i}$, $i = 1, 2, \ldots n$. So it suffices to consider a single realisation $\mathcal{D}$, for which each point process $\mathcal{T}_\theta$ will have an associated empirical intensity $\lambda^\theta_{\mathcal{D}}(t)$ with entries $\lambda^{\theta,e}_{\mathcal{D}}(t)$ at each time t. Then we will have

$$\begin{aligned}
\nabla_\theta \ell_{\mathcal{D}}(\theta) &= \sum_e \left(\int_{t_{\min}}^{t_{\max}} \nabla_\theta \log(\lambda^{\theta,e}_{\mathcal{D}}(t)) d\Lambda_{\mathcal{D}}(t) - \int_{t_{\min}}^{t_{\max}} \nabla_\theta \lambda^{\theta,e}_{\mathcal{D}}(t) dt \right) \\
&= \sum_e \left(\int_{t_{\min}}^{t_{\max}} \frac{\nabla_\theta \lambda^{\theta,e}_{\mathcal{D}}(t)}{\lambda^{\theta,e}_{\mathcal{D}}(t)} d\Lambda_{\mathcal{D}}(t) - \int_{t_{\min}}^{t_{\max}} \nabla_\theta \lambda^{\theta,e}_{\mathcal{D}}(t) dt \right), \\
H_\theta \ell_{\mathcal{D}}(\theta) &= \sum_e \left(\int_{t_{\min}}^{t_{\max}} H_\theta \log(\lambda^{\theta,e}_{\mathcal{D}}(t)) d\Lambda_{\mathcal{D}}(t) - \int_{t_{\min}}^{t_{\max}} H_\theta \lambda^{\theta,e}_{\mathcal{D}}(t) dt \right) \\
&= \sum_e \left(\int_{t_{\min}}^{t_{\max}} \left(\frac{H_\theta \lambda^{\theta,e}_{\mathcal{D}}(t)}{\lambda^{\theta,e}_{\mathcal{D}}(t)} - \left\| \frac{\nabla_\theta \lambda^{\theta,e}_{\mathcal{D}}(t)}{\lambda^{\theta,e}_{\mathcal{D}}(t)} \right\|^2 \right) d\Lambda_{\mathcal{D}}(t) - \int_{t_{\min}}^{t_{\max}} H_\theta \lambda^{\theta,e}_{\mathcal{D}}(t) dt \right).
\end{aligned} \tag{3.2.1}$$

explain that I will use a quasi newton method to maximise loglikelihood

3.2.1 Poisson Processes

Suppose we have a family of multivariate Poisson process models parameterised by their constant intensity vector ν with components ν_e. The log-likelihood of each

model is then given by

$$\begin{aligned}\ell_{\mathcal{D}}(\nu) &= \sum_e \left(\int_{t_{\min}}^{t_{\max}} \log(\nu_e) d\Lambda_{\mathcal{D}} - \int_{t_{\min}}^{t_{\max}} \nu_e dt \right) \\ &= \sum_e \left(\log(\nu) \int_{t_{\min}}^{t_{\max}} d\Lambda_{\mathcal{D}} - \nu \int_{t_{\min}}^{t_{\max}} dt \right) \\ &= \sum_e \left(\log(\nu) N_{\mathcal{D}}\left(t_{\max}\right) - \nu \left(t_{\max} - t_{\min}\right) \right).\end{aligned} \tag{3.2.2}$$

This implies that

$$\begin{aligned}\frac{\partial}{\partial \nu_e} \ell_{\mathcal{D}}(\nu) &= \frac{1}{\nu_e} N_{\mathcal{D}}\left(t_{\max}\right) - \left(t_{\max} - t_{\min}\right) \\ \frac{\partial}{\partial \nu_e \partial \nu_{e'}} \ell_{\mathcal{D}}(\nu) &= -\delta_{e,e'} \frac{N_{\mathcal{D}}\left(t_{\max}\right)}{\nu_e^2} \leq 0,\end{aligned} \tag{3.2.3}$$

where $\delta_{e,e'}$ is 1 if and only if $e = e'$ and 0 otherwise.
Since the hessian is a diagonal matrix with negative diagonal entries, it is negative definite, and so a unique local maxima exists and is a global maxima. Using the fact that $\nabla_\nu \ell_{\mathcal{D}}(\nu) = 0$, we obtain

$$\hat{\nu}_{\text{MLE}} = \frac{N_{\mathcal{D}}(t_{\max})}{t_{\max} - t_{\min}}. \tag{3.2.4}$$

Intuitively, the estimate for the arrival rate ν is the average number of events per unit time in the observed data. Note also that for the maximum likelihood model, the sample average of the residuals r will be exactly equal to one, since we have

$$\sum_{i=1}^{N_{\mathcal{D}}(t_{\max})} r_i = \bar{\Lambda}(t_{\max}) = \hat{\nu}_{\text{MLE}}(t_{\max} - t_{\min}) = N_{\mathcal{D}}(t_{\max}).$$

3.3 Maximum Likelihood Estimation for Composite Processes

Suppose that for each point process $\mathcal{T}_\theta$ in a parametric family we have

$$\lambda^\theta(\omega, t) = \sum_i \lambda^{i,\theta^i}(\omega, t),$$

where the parameter vector θ may be written as a concatenation of the parameter vectors θ^i governing the terms $\lambda^{i,\theta^i}(\omega, t)$. Then we will have empirical component probabilities $P_{\mathcal{D},i}$ with elements

$$P^e_{\mathcal{D},i}(t) = \frac{\lambda^{i,\theta^i,e}_{\mathcal{D}}(t)}{\lambda^e_{\mathcal{D}}(t)}.$$

It follows that

$$\begin{aligned}
\nabla_{\theta^i}\ell_{\mathcal{D}}(\theta) &= \sum_e \left(\int_{t_{\min}}^{t_{\max}} \frac{\nabla_{\theta^i}\lambda_{\mathcal{D}}^{i,\theta^i,e}(t)}{\lambda_{\mathcal{D}}^{\theta,e}(t)} d\Lambda_{\mathcal{D}}(t) - \int_{t_{\min}}^{t_{\max}} \nabla_{\theta^i}\lambda_{\mathcal{D}}^{\theta,e}(t)dt \right) \\
&= \sum_e \left(\int_{t_{\min}}^{t_{\max}} \frac{\nabla_{\theta^i}\lambda_{\mathcal{D}}^{i,\theta^i,e}(t)}{\lambda_{\mathcal{D}}^{i,\theta^i,e}(t)} \cdot \frac{\lambda_{\mathcal{D}}^{i,\theta^i,e}(t)}{\lambda_{\mathcal{D}}^{\theta,e}(t)} d\Lambda_{\mathcal{D}}(t) - \int_{t_{\min}}^{t_{\max}} \nabla_{\theta^i}\lambda_{\mathcal{D}}^{\theta,e}(t)dt \right) \\
&= \sum_e \left(\int_{t_{\min}}^{t_{\max}} \nabla_{\theta^i} \log\left(\left| \lambda_{\mathcal{D}}^{i,\theta^i,e}(t) \right| \right) \cdot P_{\mathcal{D},i}^e(t) d\Lambda_{\mathcal{D}}(t) - \int_{t_{\min}}^{t_{\max}} \nabla_{\theta^i}\lambda_{\mathcal{D}}^{\theta,e}(t)dt \right) \\
H_{\theta^i}\ell_{\mathcal{D}}(\theta) &= \sum_e \int_{t_{\min}}^{t_{\max}} \left(\frac{H_{\theta^i}\lambda_{\mathcal{D}}^{i,\theta^i,e}(t)}{\lambda_{\mathcal{D}}^{i,\theta^i,e}(t)} - \frac{\left(\nabla_{\theta^i}\lambda_{\mathcal{D}}^{i,\theta^i,e}(t)\right)\left(\nabla_{\theta^i}\lambda_{\mathcal{D}}^{i,\theta^i,e}(t)\right)^T}{\lambda_{\mathcal{D}}^{i,\theta^i,e}(t)^2} \cdot P_{\mathcal{D},i}^e(t) \right) \cdot P_{\mathcal{D},i}^e(t) d\Lambda_{\mathcal{D}}(t) \\
&\quad - \sum_e \int_{t_{\min}}^{t_{\max}} H_\theta \lambda_{\mathcal{D}}^{i,\theta^i,e}(t)dt,
\end{aligned} \tag{3.3.1}$$

where $\lambda^{i,\theta^i,e}(\omega,t)$ are the elements of the vector $\lambda^{i,\theta^i}(\omega,t)$. The full gradient $\nabla_\theta \ell_{\mathcal{D}}(\theta)$ is then the concatenation of the individual gradients $\nabla_{\theta^i}\ell_{\mathcal{D}}(\theta)$, and similarly the hessian $H_\theta \ell_{\mathcal{D}}(\theta)$ will be a block-diagonal matrix with $H_{\theta^i}\ell_{\mathcal{D}}(\theta)$ placed in the block whose rows and columns correspond to θ^i, and zeros in the remaining entries.

3.3.1 Expectation Maximisation

Not sure if I will include this! If $\left(\lambda_i^{\theta_i}\right)_{\mathcal{D}}$ are positive, we have

$$\log\left(\sum_i \left(\lambda_i^{\theta_i}\right)_{\mathcal{D}}(t) \right) = \sum_i \log\left(\left(\lambda_i^{\theta_i}\right)_{\mathcal{D}}(t) \right) B_{i,t}^{\theta_1,\theta_2,\cdots} - \sum_i \log\left(B_{i,t}^{\theta_1,\theta_2,\cdots} \right) B_{i,t}^{\theta_2,\theta_2,\cdots},$$

By Gibb's inequality, it follows that

$$-\sum_i \log\left(B_{i,t}^{\theta_1,\theta_2,\cdots} \right) B_{i,t}^{\theta_1,\theta_2,\cdots} \leq -\sum_i \log\left(B_{i,t}^{\theta_1',\theta_2',\cdots} \right) B_{i,t}^{\theta_1,\theta_2,\cdots}$$

EM algorithm and proof of convergence.

Can cite original dempster paper? As well as the one that corrects the errors in that paper

- Visualisation of branching matrix. How does it evolve throughout the fitting procedure?
- Quasi EM approximation
- Closed form constant time M step for multivariate state dependent hawkes processes
- Method of scoring (Newton's method)
- Negative probabilities. Conditions for kernel nonnegativity? (Probably not tractable)

3.4 General Point Process Model

Recall that the model with component intensities raised to powers κ_i was so far the most general model, nesting all others described. I will therefore focus my attention on this model with the intention that results for the simpler models will follow from this analysis. Recall the intensity as described in (2.2.3),

$$\begin{aligned}
\lambda_e(\omega,t) &= \nu_e + \sum_i \pm_i \lambda_e^i(\omega,t)^{\kappa_i},\\
\lambda_e^i(\omega,t) &= \sum_{e'\in E},\\
\lambda_{e',e}^i(\omega,t) &= \int_{t_{\min}}^t A_{e',e,i}(\mathcal{X}(\omega,s))B_{e',e,i}(\mathcal{X}(\omega,t))k_{e',e,i,i}(t-s)d\Lambda_\omega^{e'}(s).
\end{aligned}$$

To be more concrete, I will consider an intensity of the form

$$\begin{aligned}
\lambda_e(\omega,t) &= \nu_e + \sum_i \pm_i \lambda_e^i(\omega,t)^{\kappa_i},\\
\lambda_e^j(\omega,t) &= \sum_{e'\in E} \lambda_{e',e}^j(\omega,t),\\
\lambda_{e',e}^j(\omega,t) &= \int_{t_{\min}}^t \alpha_{e',e,i}A_{e',e,i}(\omega,s)B_{e',e,i}(\omega,t)\exp(-\beta_i(t-s))d\Lambda_\omega^{e'}(s),\\
A_{e',e,i}(\omega,s) &= \exp(a_{e',e,i}\cdot\chi(\mathcal{X}(\omega,s))),\\
B_{e',e,i}(\omega,t) &= \exp(b_{e',e,i}\cdot\chi(\mathcal{X}(\omega,t))).
\end{aligned} \tag{3.4.1}$$

This is parameterised by $\nu,\alpha,\beta,a,b,\kappa$.
On some dataset $\mathcal{D}$, the empirical intensity will then have the form

$$\begin{aligned}
\lambda_{\mathcal{D}}^e(t) &= \nu_e + \sum_i \pm_i \lambda_{\mathcal{D}}^{i,e}(t)^{\kappa_i},\\
\lambda_{\mathcal{D}}^{e,i}(t) &= \sum_{e'\in E} \lambda_{\mathcal{D}}^{e',e,i}(t)\\
\lambda_{\mathcal{D}}^{e',e,i}(t) &= \int_{t_{\min}}^t \alpha_{e',e,i}A_{\mathcal{D}}^{e',e,i}(s)B_{\mathcal{D}}^{e',e,i}(t)\exp(-\beta_i(t-s))d\Lambda_{\mathcal{D}}^{e'}(s),\\
A_{\mathcal{D}}^{e',e,i}(s) &= \exp(a_{e',e,i}\cdot\chi(\mathcal{X}_{\mathcal{D}}(s))),\\
B_{\mathcal{D}}^{e',e,i}(t) &= \exp(b_{e',e,i}\cdot\chi(\mathcal{X}_{\mathcal{D}}(t))),
\end{aligned} \tag{3.4.2}$$

where $\mathcal{X}_{\mathcal{D}}(t)$ represents the most recent recorded state of the order book at time t.

3.4.1 *Differentiating with respect to the Parameters*

Treating ν_e as the 0th component intensities and $\pm_i \lambda_{\mathcal{D}}^{i,e\kappa_i}$ as the ith, we can write down component probabilities

$$\begin{aligned} P_{\mathcal{D},0}^{e}(t) &= \frac{\nu_e}{\lambda_{\mathcal{D}}^{e}(t)}, \\ P_{\mathcal{D},i}^{e}(t) &= \frac{\pm_i \lambda_{\mathcal{D}}^{e,i}(t)^{\kappa_i}}{\lambda_{\mathcal{D}}^{e}(t)}. \end{aligned} \tag{3.4.3}$$

From here, we can compute derivatives of $\ell_{\mathcal{D}}$ with respect to each parameter, as in equation (3.3.1), in order to determine the newton step?.

$$\begin{aligned} \frac{\partial}{\partial \nu_e}\ell_{\mathcal{D}} &= \frac{1}{\nu_e}\int_{t_{\min}}^{t_{\max}} P_{\mathcal{D},0}^{e}(t) d\Lambda_{\mathcal{D}}^{e}(t) - \int_{t_{\min}}^{t_{\max}} dt \\ &= \frac{1}{\nu_e}\int_{t_{\min}}^{t_{\max}} P_{\mathcal{D},0}^{e}(t) d\Lambda_{\mathcal{D}}^{e}(t) - (t_{\max} - t_{\min}), \\ \frac{\partial^2}{(\partial \nu_e)^2}\ell_{\mathcal{D}} &= -\frac{1}{\nu_e^2}\int_{t_{\min}}^{t_{\max}} P_{\mathcal{D},0}^{e}(t) d\Lambda_{\mathcal{D}}^{e}(t), \end{aligned} \tag{3.4.4}$$

By writing $\lambda_{\mathcal{D}}^{e',e,i}(t)$ as

$$\exp(-\beta_i t)\alpha_{e',e,i} B_{\mathcal{D}}^{e',e,i}(t) \int_{t_{\min}}^{t} A_{\mathcal{D}}^{e',e,i}(s)\exp(\beta_i s) d\Lambda_{\mathcal{D}}^{e'}(s),$$

it becomes evident that

$$\begin{aligned} \lambda_{\mathcal{D}}^{e',e,i}(t+\epsilon) &= \exp(-\beta\epsilon)\lambda_{\mathcal{D}}^{e',e,i}(t), \\ \lambda_{\mathcal{D}}^{e,i}(t+\epsilon) &= \exp(-\beta\epsilon)\lambda_{\mathcal{D}}^{e,i}(t), \\ \pm_i \lambda_{\mathcal{D}}^{e,i}(t+\epsilon)^{\kappa_i} &= \pm_i \exp(-\beta\kappa_i\epsilon)\lambda_{\mathcal{D}}^{e,i}(t)^{\kappa_i}, \end{aligned}$$

for any $\epsilon > 0$ so long as $\sum_{e'} \Lambda_{\mathcal{D}}^{e'}((t, t+\epsilon)) = 0$. On the other hand, at some event time t with event type e, the right-limit of $\lambda_{\mathcal{D}}^{e',e,i}$ will be equal to

$$\lambda_{\mathcal{D}}^{e',e,i}(t^+) = \frac{B_{\mathcal{D}}^{e',e,i}(t)}{B_{\mathcal{D}}^{e',e,i}(t^-)}\lambda_{\mathcal{D}}^{e',e,i}(t) + \alpha_{e',e,i} A_{\mathcal{D}}^{e',e,i}(t) B_{\mathcal{D}}^{e',e,i}(t), \tag{3.4.5}$$

where $B_{\mathcal{D}}^{e',e,i}(t^-)$ is the left-limit of $B_{\mathcal{D}}^{e',e,i}$.

Partition the observation interval $(t_{\min}, t_{\max})$ into $|\mathcal{D}|+1$ different intervals T_j with endpoints in $\mathcal{D} \cup \{t_{\min}, t_{\max}\}$. Furthermore, define

$$\begin{aligned}
\Gamma_0^\gamma(T_j) &= \int_{\inf T_j}^{\sup T_j} \exp(-\gamma(t - \inf T_j))dt \\
&= \frac{1}{\gamma}(1 - \exp(-\gamma\mu_{\text{Lebesgue}}(T_j))), \\
\Gamma_1^\gamma(T_j) &= \int_{\inf T_j}^{\sup T_j} (t - \inf T_j)\exp(-\gamma(t - \inf T_j))dt \\
&= \frac{1}{\gamma^2}(1 - \exp(-\gamma\mu_{\text{Lebesgue}}(T_j))(\gamma\mu_{\text{Lebesgue}}(T_j) + 1)), \\
\Gamma_2^\gamma(T_j) &= \int_{\inf T_j}^{\sup T_j} (t - \inf T_j)^2\exp(-\gamma(t - \inf T_j))dt \\
&= \frac{1}{\gamma^3}(2 - \exp(-\gamma\mu_{\text{Lebesgue}}(T_j))(\gamma^2\mu_{\text{Lebesgue}}(T_j)^2 + 2\gamma\mu_{\text{Lebesgue}}(T_j) + 2)),
\end{aligned} \tag{3.4.6}$$

with $\mu_{\text{Lebesgue}}(T_j) = \sup T_j - \inf T_j$.
Then differentiating with respect to $\alpha_{e',e,i}$ yields

$$\begin{aligned}
\frac{\partial}{\partial\alpha_{e',e,i}}\lambda_{\mathcal{D}}^{e,i}(t)^{\kappa_i} &= \frac{\kappa_i}{\alpha_{e',e,i}}\lambda_{\mathcal{D}}^{e',e,i}(t)\lambda_{\mathcal{D}}^{e,i}(t)^{\kappa_i-1} \\
\frac{\partial}{\partial\alpha_{e',e,i}}\ell_{\mathcal{D}} &= \int_{t_{\min}}^{t_{\max}} \frac{\kappa_i\lambda_{\mathcal{D}}^{e',e,i}(t)}{\alpha_{e',e,i}\lambda_{\mathcal{D}}^{e,i}(t)} \cdot P_{\mathcal{D},i}^e(t)d\Lambda_{\mathcal{D}}^e(t) - \pm_i\frac{\kappa_i}{\alpha_{e',e,i}}\int_{t_{\min}}^{t_{\max}} \lambda_{\mathcal{D}}^{e',e,i}(t)\lambda_{\mathcal{D}}^{e,i}(t)^{\kappa_i-1}dt \\
&= \int_{t_{\min}}^{t_{\max}} \frac{\kappa_i\lambda_{\mathcal{D}}^{e',e,i}(t)}{\alpha_{e',e,i}\lambda_{\mathcal{D}}^{e,i}(t)} \cdot P_{\mathcal{D},i}^e(t)d\Lambda_{\mathcal{D}}^e(t) \\
&\quad - \pm_i\frac{\kappa_i}{\alpha_{e',e,i}}\sum_j \lambda_{\mathcal{D}}^{e',e,i}((\inf T_j)^+)\lambda_{\mathcal{D}}^{e,i}((\inf T_j)^+)^{\kappa_i-1}\Gamma_0^{\beta_i\kappa_i}(T_j), \\
\frac{\partial^2}{(\partial\alpha_{e',e,i})^2}\lambda_{\mathcal{D}}^{e,i}(t)^{\kappa_i} &= \frac{\kappa_i(\kappa_i-1)}{\alpha_{e',e,i}^2}\lambda_{\mathcal{D}}^{e',e,i}(t)^2\lambda_{\mathcal{D}}^{e,i}(t)^{\kappa_i-2} \\
\frac{\partial^2}{(\partial\alpha_{e',e,i})^2}\ell_{\mathcal{D}} &= \int_{t_{\min}}^{t_{\max}} \left(\frac{\kappa_i(\kappa_i-1)\lambda_{\mathcal{D}}^{e',e,i}(t)^2}{\alpha_{e',e,i}^2\lambda_{\mathcal{D}}^{e,i}(t)^2} - \left(\frac{\kappa_i\lambda_{\mathcal{D}}^{e',e,i}(t)}{\alpha_{e',e,i}\lambda_{\mathcal{D}}^{e,i}(t)}\right)^2 \cdot P_{\mathcal{D},i}^e(t)\right) \cdot P_{\mathcal{D},i}^e(t)d\Lambda_{\mathcal{D}}^e(t) \\
&\quad - \pm_i\int_{t_{\min}}^{t_{\max}} \frac{\kappa_i(\kappa_i-1)}{\alpha_{e',e,i}^2}\lambda_{\mathcal{D}}^{e',e,i}(t)^2\lambda_{\mathcal{D}}^{e,i}(t)^{\kappa_i-2}dt \\
&= \int_{t_{\min}}^{t_{\max}} \left(\frac{\kappa_i(\kappa_i-1)\lambda_{\mathcal{D}}^{e',e,i}(t)^2}{\alpha_{e',e,i}^2\lambda_{\mathcal{D}}^{e,i}(t)^2} - \left(\frac{\kappa_i\lambda_{\mathcal{D}}^{e',e,i}(t)}{\alpha_{e',e,i}\lambda_{\mathcal{D}}^{e,i}(t)}\right)^2 \cdot P_{\mathcal{D},i}^e(t)\right) \cdot P_{\mathcal{D},i}^e(t)d\Lambda_{\mathcal{D}}^e(t) \\
&\quad - \pm_i\sum_j \frac{\kappa_i(\kappa_i-1)}{\alpha_{e',e,i}^2}\lambda_{\mathcal{D}}^{e',e,i}((\inf T_j)^+)^2\lambda_{\mathcal{D}}^{e,i}(\inf T_j)^+)^{\kappa_i-2}\Gamma_0^{\beta_i\kappa_i}(T_j).
\end{aligned}$$

Next, we will have

$$\frac{\partial}{\partial\beta_i}\lambda_{\mathcal{D}}^{e',e,i}(t) = -\int_{t_{\min}}^{t}(t-s)\alpha_{e',e,i}A_{\mathcal{D}}^{e',e,i}(s)B_{\mathcal{D}}^{e',e,i}(t)\exp(-\beta_i(t-s))d\Lambda_{\mathcal{D}}^{e'}(s),$$
$$\frac{\partial^2}{(\partial\beta_i)^2}\lambda_{\mathcal{D}}^{e',e,i}(t) = \int_{t_{\min}}^{t}(t-s)^2\alpha_{e',e,i}A_{\mathcal{D}}^{e',e,i}(t)B_{\mathcal{D}}^{e',e,i}(t)\exp(-\beta_i(t-s))d\Lambda_{\mathcal{D}}^{e'}(s),$$

and in the case where $\sum_{e'}\Lambda_{\mathcal{D}}^{e'}((t,t+\epsilon)) = 0$ we can express these recursively as

$$\frac{\partial}{\partial\beta_i}\lambda_{\mathcal{D}}^{e',e,i}(t+\epsilon) = \exp(-\beta_i\epsilon)\left(-\epsilon\lambda_{\mathcal{D}}^{e',e,i}(t) + \frac{\partial}{\partial\beta_i}\lambda_{\mathcal{D}}^{e',e,i}(t)\right),$$
$$\frac{\partial^2}{(\partial\beta_i)^2}\lambda_{\mathcal{D}}^{e',e,i}(t+\epsilon) = \exp(-\beta_i\epsilon)\left(\epsilon^2\lambda_{\mathcal{D}}^{e',e,i}(t) - 2\epsilon\frac{\partial}{\partial\beta_i}\lambda_{\mathcal{D}}^{e',e,i}(t) + \frac{\partial^2}{(\partial\beta_i)^2}\lambda_{\mathcal{D}}^{e',e,i}(t)\right).$$

Using these results, we can then express the derivatives of $\lambda_{\mathcal{D}}^{e,i}$ and $\ell_{\mathcal{D}}$ with respect to β_i as

$$\frac{\partial}{\partial\beta_i}\lambda_{\mathcal{D}}^{e,i}(t)^{\kappa_i} = \kappa_i\lambda_{\mathcal{D}}^{e,i}(t)^{\kappa_i-1}\sum_{e'}\frac{\partial}{\partial\beta_i}\lambda_{\mathcal{D}}^{e',e,i}(t),$$

$$\begin{aligned}\frac{\partial}{\partial\beta_i}\ell_{\mathcal{D}} &= \int_{t_{\min}}^{t_{\max}}\frac{\kappa_i\sum_{e'}\frac{\partial}{\partial\beta_i}\lambda_{\mathcal{D}}^{e',e,i}(t)}{\lambda_{\mathcal{D}}^{e,i}(t)}P_{\mathcal{D},i}^{e}(t)d\Lambda_{\mathcal{D}}^{e}(t)\\ &\quad - \pm_i\kappa_i\int_{t_{\min}}^{t_{\max}}\lambda_{\mathcal{D}}^{e,i}(t)^{\kappa_i-1}\sum_{e'}\frac{\partial}{\partial\beta_i}\lambda_{\mathcal{D}}^{e',e,i}(t)dt\\ &= \int_{t_{\min}}^{t_{\max}}\frac{\kappa_i\sum_{e'}\frac{\partial}{\partial\beta_i}\lambda_{\mathcal{D}}^{e',e,i}(t)}{\lambda_{\mathcal{D}}^{e,i}(t)}P_{\mathcal{D},i}^{e}(t)d\Lambda_{\mathcal{D}}^{e}(t)\\ &\quad + \pm_i\kappa_i\sum_{j}\lambda_{\mathcal{D}}^{e,i}((\inf T_j)^+)^{\kappa_i-1}\sum_{e'}\lambda_{\mathcal{D}}^{e',e,i}((\inf T_j)^+)\Gamma_1^{\beta_i\kappa_i}(T_j)\\ &\quad - \pm_i\kappa_i\sum_{j}\lambda_{\mathcal{D}}^{e,i}((\inf T_j)^+)^{\kappa_i-1}\sum_{e'}\frac{\partial}{\partial\beta_i}\lambda_{\mathcal{D}}^{e',e,i}(\inf T_j)^+)\Gamma_0^{\beta_i\kappa_i}(T_j),\end{aligned}$$

$$\frac{\partial^2}{(\partial\beta_i)^2}\lambda_{\mathcal{D}}^{e,i}(t)^{\kappa_i} = \kappa_i(\kappa_i-1)\lambda_{\mathcal{D}}^{e,i}(t)^{\kappa_i-2}\left(\sum_{e'}\frac{\partial}{\partial\beta_i}\lambda_{\mathcal{D}}^{e',e,i}(t)\right)^2 + \kappa_i\lambda_{\mathcal{D}}^{e,i}(t)^{\kappa_i-1}\sum_{e'}\frac{\partial^2}{(\partial\beta_i)^2}\lambda_{\mathcal{D}}^{e',e,i}(t),$$

$$\begin{aligned}\frac{\partial^2}{(\partial\beta_i)^2}\ell_{\mathcal{D}} &= \int_{t_{\min}}^{t_{\max}}\left(\kappa_i(\kappa_i-1)\left(\frac{\sum_{e'}\frac{\partial}{\partial\beta_i}\lambda_{\mathcal{D}}^{e',e,i}(t)}{\lambda_{\mathcal{D}}^{e,i}(t)}\right)^2 + \kappa_i\frac{\sum_{e'}\frac{\partial^2}{(\partial\beta_i)^2}}{\lambda_{\mathcal{D}}^{e,i}(t)}\right)P_{\mathcal{D},i}^{e}(t)d\Lambda_{\mathcal{D}}^{e}(t)\\ &\quad - \int_{t_{\min}}^{t_{\max}}\kappa_i^2\left(\frac{\sum_{e'}\frac{\partial}{\partial\beta_i}\lambda_{\mathcal{D}}^{e',e,i}(t)}{\lambda_{\mathcal{D}}^{e,i}(t)}\right)^2 P_{\mathcal{D},i}^{e}(t)^2d\Lambda_{\mathcal{D}}^{e}(t)\\ &\quad - \pm_i\kappa_i(\kappa_i-1)\int_{t_{\min}}^{t_{\max}}\lambda_{\mathcal{D}}^{e,i}(t)^{\kappa_i-2}\left(\sum_{e'}\frac{\partial}{\partial\beta_i}\lambda_{\mathcal{D}}^{e',e,i}(t)\right)^2 dt\\ &\quad + \pm_i\kappa_i\int_{t_{\min}}^{t_{\max}}\lambda_{\mathcal{D}}^{e,i}(t)^{\kappa_i-1}\sum_{e'}\frac{\partial^2}{(\partial\beta_i)^2}\lambda_{\mathcal{D}}^{e',e,i}(t)dt\end{aligned}$$

$$
\begin{aligned}
&= \int_{t_{\min}}^{t_{\max}} \left(\kappa_i(\kappa_i - 1) \left(\frac{\sum_{e'} \frac{\partial}{\partial \beta_i} \lambda_{\mathcal{D}}^{e',e,i}(t)}{\lambda_{\mathcal{D}}^{e,i}(t)} \right)^2 + \kappa_i \frac{\sum_{e'} \frac{\partial^2}{(\partial \beta_i)^2}}{\lambda_{\mathcal{D}}^{e,i}(t)} \right) P_{\mathcal{D},i}^{e}(t) d\Lambda_{\mathcal{D}}^{e}(t) \\
&- \int_{t_{\min}}^{t_{\max}} \kappa_i^2 \left(\frac{\sum_{e'} \frac{\partial}{\partial \beta_i} \lambda_{\mathcal{D}}^{e',e,i}(t)}{\lambda_{\mathcal{D}}^{e,i}(t)} \right)^2 P_{\mathcal{D},i}^{e}(t)^2 d\Lambda_{\mathcal{D}}^{e}(t) \\
&- \pm_i \kappa_i(\kappa_i - 1) \sum_j \lambda_{\mathcal{D}}^{e,i}((\inf T_j)^+)^{\kappa_i - 2} \Gamma_0^{\beta_i \kappa_i}(T_j) \left(\frac{\partial}{\partial \beta_i} \lambda_{\mathcal{D}}^{e,i}(\inf T_j)^+) \right)^2 \\
&+ \pm_i 2\kappa_i(\kappa_i - 1) \sum_j \lambda_{\mathcal{D}}^{e,i}((\inf T_j)^+)^{\kappa_i - 1} \Gamma_1^{\beta_i \kappa_i}(T_j) \frac{\partial}{\partial \beta_i} \lambda_{\mathcal{D}}^{e,i}(\inf T_j)^+) \\
&- \pm_i \kappa_i(\kappa_i - 1) \sum_j \lambda_{\mathcal{D}}^{e,i}((\inf T_j)^+)^{\kappa_i} \Gamma_2^{\beta_i \kappa_i}(T_j) \\
&+ \kappa_i \sum_j \lambda_{\mathcal{D}}^{e,i}((\inf T_j)^+)^{\kappa_i - 1} \Gamma_0^{\beta_i \kappa_i}(T_j) \frac{\partial^2}{(\partial \beta_i)^2} \lambda_{\mathcal{D}}^{e,i}((\inf T_j)^+) \\
&- \pm_i 2\kappa_i \sum_j \lambda_{\mathcal{D}}^{e,i}((\inf T_j)^+)^{\kappa_i - 1} \Gamma_1^{\beta_i \kappa_i}(T_j) \frac{\partial}{\partial \beta_i} \lambda_{\mathcal{D}}^{e,i}((\inf T_j)^+) \\
&+ \pm_i \kappa_i \sum_j \lambda_{\mathcal{D}}^{e,i}((\inf T_j)^+)^{\kappa_i} \Gamma_2^{\beta_i \kappa_i}(T_j) \\
&= \int_{t_{\min}}^{t_{\max}} \left(\kappa_i(\kappa_i - 1) \left(\frac{\sum_{e'} \frac{\partial}{\partial \beta_i} \lambda_{\mathcal{D}}^{e',e,i}(t)}{\lambda_{\mathcal{D}}^{e,i}(t)} \right)^2 + \kappa_i \frac{\sum_{e'} \frac{\partial^2}{(\partial \beta_i)^2}}{\lambda_{\mathcal{D}}^{e,i}(t)} \right) P_{\mathcal{D},i}^{e}(t) d\Lambda_{\mathcal{D}}^{e}(t) \\
&- \int_{t_{\min}}^{t_{\max}} \kappa_i^2 \left(\frac{\sum_{e'} \frac{\partial}{\partial \beta_i} \lambda_{\mathcal{D}}^{e',e,i}(t)}{\lambda_{\mathcal{D}}^{e,i}(t)} \right)^2 P_{\mathcal{D},i}^{e}(t)^2 d\Lambda_{\mathcal{D}}^{e}(t) \\
&- \pm_i \kappa_i(\kappa_i - 1) \sum_j \lambda_{\mathcal{D}}^{e,i}((\inf T_j)^+)^{\kappa_i - 2} \Gamma_0^{\beta_i \kappa_i}(T_j) \left(\frac{\partial}{\partial \beta_i} \lambda_{\mathcal{D}}^{e,i}(\inf T_j)^+) \right)^2 \\
&+ \pm_i 2\kappa_i(\kappa_i - 2) \sum_j \lambda_{\mathcal{D}}^{e,i}((\inf T_j)^+)^{\kappa_i - 1} \Gamma_1^{\beta_i \kappa_i}(T_j) \frac{\partial}{\partial \beta_i} \lambda_{\mathcal{D}}^{e,i}(\inf T_j)^+) \\
&- \pm_i \kappa_i(\kappa_i - 2) \sum_j \lambda_{\mathcal{D}}^{e,i}((\inf T_j)^+)^{\kappa_i} \Gamma_2^{\beta_i \kappa_i}(T_j) \\
&+ \pm_i \kappa_i \sum_j \lambda_{\mathcal{D}}^{e,i}((\inf T_j)^+)^{\kappa_i - 1} \Gamma_0^{\beta_i \kappa_i}(T_j) \frac{\partial^2}{(\partial \beta_i)^2} \lambda_{\mathcal{D}}^{e,i}((\inf T_j)^+).
\end{aligned}
$$

Differentiating with respect to $a_{e',e,i}$, we obtain

$$\begin{aligned}
\nabla_{a_{e',e,i}}\lambda_{\mathcal{D}}^{e',e,i}(t) &= \int_{t_{\min}}^{t} \chi(\mathcal{X}_{\mathcal{D}}(s))\alpha_{e',e,i}A_{\mathcal{D}}^{e',e,i}(s)B_{\mathcal{D}}^{e',e,i}(t)\exp(-\beta_i(t-s))d\Lambda_{\mathcal{D}}^{e'}(s),\\
\nabla_{a_{e',e,i}}\lambda_{\mathcal{D}}^{e,i}(t)^{\kappa_i} &= \kappa_i\lambda_{\mathcal{D}}^{e,i}(t)^{\kappa_i-1}\nabla_{a_{e',e,i}}\lambda_{\mathcal{D}}^{e',e,i}(t),\\
\nabla_{a_{e',e,i}}\ell_{\mathcal{D}} &= \int_{t_{\min}}^{t_{\max}} \frac{\kappa_i\nabla_{a_{e',e,i}}\lambda_{\mathcal{D}}^{e',e,i}(t)}{\lambda_{\mathcal{D}}^{e,i}(t)}P_{\mathcal{D},i}^{e}(t)d\Lambda_{\mathcal{D}}^{e}(t) - \pm_i\int_{t_{\min}}^{t_{\max}} \kappa_i\lambda_{\mathcal{D}}^{e,i}(t)^{\kappa_i-1}\nabla_{a_{e',e,i}}\lambda_{\mathcal{D}}^{e',e,i}(t)dt\\
&= \int_{t_{\min}}^{t_{\max}} \frac{\kappa_i\nabla_{a_{e',e,i}}\lambda_{\mathcal{D}}^{e',e,i}(t)}{\lambda_{\mathcal{D}}^{e,i}(t)}P_{\mathcal{D},i}^{e}(t)d\Lambda_{\mathcal{D}}^{e}(t)\\
&\quad - \pm_i\kappa_i\sum_j \Gamma_0^{\beta_i\kappa_i}(T_j)\lambda_{\mathcal{D}}^{e,i}((\inf T_j)^+)^{\kappa_i-1}\nabla_{a_{e',e,i}}\lambda_{\mathcal{D}}^{e',e,i}((\inf T_j)^+),\\
H_{a_{e',e,i}}\lambda_{\mathcal{D}}^{e',e,i}(t) &= \int_{t_{\min}}^{t} \chi(\mathcal{X}_{\mathcal{D}}(s))\chi(\mathcal{X}_{\mathcal{D}}(s))^T\alpha_{e',e,i}A_{\mathcal{D}}^{e',e,i}(s)B_{\mathcal{D}}^{e',e,i}(t)\exp(-\beta_i(t-s))d\Lambda_{\mathcal{D}}^{e'}(s),\\
H_{a_{e',e,i}}\lambda_{\mathcal{D}}^{e,i}(t)^{\kappa_i} &= \kappa_i(\kappa_i-1)\lambda_{\mathcal{D}}^{e,i}(t)^{\kappa_i-2}\left(\nabla_{a_{e',e,i}}\lambda_{\mathcal{D}}^{e',e,i}(t)\right)\left(\nabla_{a_{e',e,i}}\lambda_{\mathcal{D}}^{e',e,i}(t)\right)^T,\\
H_{a_{e',e,i}}\ell_{\mathcal{D}} &= \int_{t_{\min}}^{t_{\max}} \frac{\kappa_i(\kappa_i-1)(\nabla_{a_{e',e,i}}\lambda_{\mathcal{D}}^{e',e,i}(t))(\nabla_{a_{e',e,i}}\lambda_{\mathcal{D}}^{e',e,i}(t))^T}{\lambda_{\mathcal{D}}^{e,i}(t)^2}P_{\mathcal{D},i}^{e}(t)d\Lambda_{\mathcal{D}}^{e}(t)\\
&\quad + \int_{t_{\min}}^{t_{\max}} \frac{\kappa_i H_{a_{a',e,i}}\lambda_{\mathcal{D}}^{e',e,i}(t)}{\lambda_{\mathcal{D}}^{e,i}(t)}P_{\mathcal{D},i}^{e}(t)d\Lambda_{\mathcal{D}}^{e}(t)\\
&\quad - \int_{t_{\min}}^{t_{\max}} \frac{\kappa_i^2(\nabla_{a_{e',e,i}}\lambda_{\mathcal{D}}^{e',e,i}(t))(\nabla_{a_{e',e,i}}\lambda_{\mathcal{D}}^{e',e,i}(t))^T}{\lambda_{\mathcal{D}}^{e,i}(t)^2}P_{\mathcal{D},i}^{e}(t)^2d\Lambda_{\mathcal{D}}^{e}(t)\\
&\quad - \pm_i\int_{t_{\min}}^{t_{\max}} \kappa_i(\kappa_i-1)\lambda_{\mathcal{D}}^{e,i}(t)^{\kappa_i-2}\left(\nabla_{a_{e',e,i}}\lambda_{\mathcal{D}}^{e',e,i}(t)\right)\left(\nabla_{a_{e',e,i}}\lambda_{\mathcal{D}}^{e',e,i}(t)\right)^T dt\\
&= \int_{t_{\min}}^{t_{\max}} \frac{\kappa_i(\kappa_i-1)(\nabla_{a_{e',e,i}}\lambda_{\mathcal{D}}^{e',e,i}(t))(\nabla_{a_{e',e,i}}\lambda_{\mathcal{D}}^{e',e,i}(t))^T}{\lambda_{\mathcal{D}}^{e,i}(t)^2}P_{\mathcal{D},i}^{e}(t)d\Lambda_{\mathcal{D}}^{e}(t)\\
&\quad + \int_{t_{\min}}^{t_{\max}} \frac{\kappa_i H_{a_{a',e,i}}\lambda_{\mathcal{D}}^{e',e,i}(t)}{\lambda_{\mathcal{D}}^{e,i}(t)}P_{\mathcal{D},i}^{e}(t)d\Lambda_{\mathcal{D}}^{e}(t)\\
&\quad - \int_{t_{\min}}^{t_{\max}} \frac{\kappa_i^2(\nabla_{a_{e',e,i}}\lambda_{\mathcal{D}}^{e',e,i}(t))(\nabla_{a_{e',e,i}}\lambda_{\mathcal{D}}^{e',e,i}(t))^T}{\lambda_{\mathcal{D}}^{e,i}(t)^2}P_{\mathcal{D},i}^{e}(t)^2d\Lambda_{\mathcal{D}}^{e}(t)\\
&\quad - \pm_i\kappa_i(\kappa_i-1)\sum_j \Gamma_0^{\beta_i\kappa_i}(T_j)\lambda_{\mathcal{D}}^{e,i}((\inf T_j)^+)^{\kappa_i-2}\\
&\quad \cdot\left(\nabla_{a_{e',e,i}}\lambda_{\mathcal{D}}^{e',e,i}((\inf T_j)^+)\right)\left(\nabla_{a_{e',e,i}}\lambda_{\mathcal{D}}^{e',e,i}((\inf T_j)^+)\right)^T.
\end{aligned}$$

Similarly, differentiating with respect to $b_{e',e,i}$ yields

$$
\begin{aligned}
\nabla_{b_{e',e,i}} \lambda_{\mathcal{D}}^{e',e,i}(t) &= \chi(\mathcal{X}_{\mathcal{D}}(t^-))\lambda_{\mathcal{D}}^{e',e,i}(t), \\
\nabla_{b_{e',e,i}} \lambda_{\mathcal{D}}^{e,i}(t)^{\kappa_i} &= \kappa_i \lambda_{\mathcal{D}}^{e,i}(t)^{\kappa_i - 1}\chi(\mathcal{X}_{\mathcal{D}}(t^-))\lambda_{\mathcal{D}}^{e',e,i}(t), \\
\nabla_{b_{e',e,i}} \ell_{\mathcal{D}} &= \int_{t_{\min}}^{t_{\max}} \frac{\kappa_i \chi(\mathcal{X}_{\mathcal{D}}(t^-))\lambda_{\mathcal{D}}^{e',e,i}(t)}{\lambda_{\mathcal{D}}^{e,i}(t)} P_{\mathcal{D},i}^{e}(t) d\Lambda_{\mathcal{D}}^{e}(t) \\
&\quad - \pm_i \int_{t_{\min}}^{t_{\max}} \kappa_i \lambda_{\mathcal{D}}^{e,i}(t)^{\kappa_i - 1}\chi(\mathcal{X}_{\mathcal{D}}(t^-))\lambda_{\mathcal{D}}^{e',e,i}(t) dt \\
&= \int_{t_{\min}}^{t_{\max}} \frac{\kappa_i \chi(\mathcal{X}_{\mathcal{D}}(t^-))\lambda_{\mathcal{D}}^{e',e,i}(t)}{\lambda_{\mathcal{D}}^{e,i}(t)} P_{\mathcal{D},i}^{e}(t) d\Lambda_{\mathcal{D}}^{e}(t) \\
&\quad - \pm_i \kappa_i \sum_j \Gamma_0^{\beta_i \kappa_i}(T_j) \lambda_{\mathcal{D}}^{e,i}((\inf T_j)^+)^{\kappa_i - 1}\chi(\mathcal{X}_{\mathcal{D}}((\inf T_j)))\lambda_{\mathcal{D}}^{e',e,i}((\inf T_j)^+), \\
H_{b_{e',e,i}} \lambda_{\mathcal{D}}^{e',e,i}(t) &= \chi(\mathcal{X}_{\mathcal{D}}(t^-))\chi(\mathcal{X}_{\mathcal{D}}(t^-))^T \lambda_{\mathcal{D}}^{e',e,i}(t), \\
H_{b_{e',e,i}} \lambda_{\mathcal{D}}^{e,i}(t)^{\kappa_i} &= \kappa_i(\kappa_i - 1)\chi(\mathcal{X}_{\mathcal{D}}(t^-))\chi(\mathcal{X}_{\mathcal{D}}(t^-))^T \lambda_{\mathcal{D}}^{e,i}(t)^{\kappa_i - 2}\lambda_{\mathcal{D}}^{e',e,i}(t)^2, \\
H_{b_{e',e,i}} \ell_{\mathcal{D}} &= \int_{t_{\min}}^{t_{\max}} \kappa_i(\kappa_i - 1)\chi(\mathcal{X}_{\mathcal{D}}(t^-))\chi(\mathcal{X}_{\mathcal{D}}(t^-))^T \frac{\lambda_{\mathcal{D}}^{e',e,i}(t)^2}{\lambda_{\mathcal{D}}^{e,i}(t)^2} P_{\mathcal{D},i}^{e}(t) d\Lambda_{\mathcal{D}}^{e}(t) \\
&\quad + \int_{t_{\min}}^{t_{\max}} \kappa_i \chi(\mathcal{X}_{\mathcal{D}}(t^-))\chi(\mathcal{X}_{\mathcal{D}}(t^-))^T \frac{\lambda_{\mathcal{D}}^{e',e,i}(t)}{\lambda_{\mathcal{D}}^{e,i}(t)} P_{\mathcal{D},i}^{e}(t) d\Lambda_{\mathcal{D}}^{e}(t) \\
&\quad - \int_{t_{\min}}^{t_{\max}} \kappa_i^2 \chi(\mathcal{X}_{\mathcal{D}}(t^-))\chi(\mathcal{X}_{\mathcal{D}}(t^-))^T \frac{\lambda_{\mathcal{D}}^{e',e,i}(t)^2}{\lambda_{\mathcal{D}}^{e,i}(t)^2} P_{\mathcal{D},i}^{e}(t)^2 d\Lambda_{\mathcal{D}}^{e}(t) \\
&\quad - \pm_i \int_{t_{\min}}^{t_{\max}} \kappa_i(\kappa_i - 1)\chi(\mathcal{X}_{\mathcal{D}}(t^-))\chi(\mathcal{X}_{\mathcal{D}}(t^-))^T \lambda_{\mathcal{D}}^{e,i}(t)^{\kappa_i - 2}\lambda_{\mathcal{D}}^{e',e,i}(t)^2 dt \\
&= - \int_{t_{\min}}^{t_{\max}} \kappa_i \chi(\mathcal{X}_{\mathcal{D}}(t^-))\chi(\mathcal{X}_{\mathcal{D}}(t^-))^T \frac{\lambda_{\mathcal{D}}^{e',e,i}(t)^2}{\lambda_{\mathcal{D}}^{e,i}(t)^2} P_{\mathcal{D},i}^{e}(t) d\Lambda_{\mathcal{D}}^{e}(t) \\
&\quad + \int_{t_{\min}}^{t_{\max}} \kappa_i \chi(\mathcal{X}_{\mathcal{D}}(t^-))\chi(\mathcal{X}_{\mathcal{D}}(t^-))^T \frac{\lambda_{\mathcal{D}}^{e',e,i}(t)}{\lambda_{\mathcal{D}}^{e,i}(t)} P_{\mathcal{D},i}^{e}(t) d\Lambda_{\mathcal{D}}^{e}(t) \\
&\quad - \pm_i \kappa_i(\kappa_i - 1) \sum_j \Gamma_0^{\beta_i \kappa_i}(T_j) \lambda_{\mathcal{D}}^{e,i}((\inf T_j)^+)^{\kappa_i - 2} \\
&\quad \cdot \chi(\mathcal{X}_{\mathcal{D}}((\inf T_j)))\chi(\mathcal{X}_{\mathcal{D}}((\inf T_j)))^T \lambda_{\mathcal{D}}^{e',e,i}((\inf T_j)^+)^2.
\end{aligned}
$$

Finally, differentiating with respect to κ_i yields

$$\begin{aligned}
\frac{\partial}{\partial \kappa_i} \lambda_{\mathcal{D}}^{e,i}(t)^{\kappa_i} &= \lambda_{\mathcal{D}}^{e,i}(t)^{\kappa_i} \log \lambda_{\mathcal{D}}^{e,i}(t) \\
\frac{\partial}{\partial \kappa_i} \ell_{\mathcal{D}} &= \int_{t_{\min}}^{t_{\max}} \log \lambda_{\mathcal{D}}^{e,i}(t) P_{\mathcal{D},i}^{e}(t) d\Lambda_{\mathcal{D}}^{e}(t) - \pm_i \int_{t_{\min}}^{t_{\max}} \lambda_{\mathcal{D}}^{e,i}(t)^{\kappa_i} \log \lambda_{\mathcal{D}}^{e,i}(t) dt \\
&= \int_{t_{\min}}^{t_{\max}} \log \lambda_{\mathcal{D}}^{e,i}(t) P_{\mathcal{D},i}^{e}(t) d\Lambda_{\mathcal{D}}^{e}(t) \\
&\quad - \pm_i \sum_j \lambda_{\mathcal{D}}^{e,i}((\inf T_j)^+)^{\kappa_i} \left(\log \lambda_{\mathcal{D}}^{e,i}((\inf T_j)^+) \Gamma_0^{\beta_i \kappa_i}(T_j) - \beta_i \Gamma_1^{\beta_i \kappa_i}(T_j) \right).
\end{aligned}$$

3.5 Uncertainty Quantification

Fisher information calculations (different to observed information)

Asymptotic normality

Parametric Bootstrap

3.6 Computational Concerns

Sensor fusion for parallelisation

Momentum (analyse autocorrelation of parameter changes throughout the learning process)

Exploiting sparsity [14]

3.7 Model Selection

Performance of information criteria for selection of Hawkes process models of financial data https://www.tandfonline.com/doi/full/10.1080/14697688.2017.1403140

Chapter 4

Simulation

Simulation of the limit order book in accordance with some model of its behaviour serves at least three different purposes. Firstly, in order to perform uncertainty quantification without appealing to asymptotic behaviour of the MLE, we can perform a parametric bootstrap: simulate multiple realisations of the point process using the MLE parameters, and find the MLE on each of these new synthetic datasets to compare the distribution of these new estimates with the "true" value used for simulation. Secondly, access to a generative model for order book data lets us predict the distribution of future prices and other characteristics of the order book conditional on the information observed so far. Finally, such a model allows us to answer counterfactual questions, such as quantifying the effect that making a trade of some size will have on market prices, in a way that a purely predictive model might not be able to.

4.1 Point Process Simulation

Ogata (1981) [**?**] describes an algorithm to iteratively sample from an arbitrary point process model for which we can obtain an upper bound on the intensity between the present time t and the next (unknown) event time. Suppose that a multivariate point process $\mathcal{T}$ has empirical intensities $\overset{e}{\mathcal{D}}(t)$ for any dataset $\mathcal{D}$. To simulate $\mathcal{T}$ on an interval $(t_{\min}, t_{\max})$, begin with an empty dataset $\mathcal{D}_0$ and $t_0 = t_{\min}$, and then for each $i = 0, 1, \ldots$ do the following:

1. Compute the upper bound $\bar{\lambda}_{\mathcal{D}_i}(t_i)$ such that we are guaranteed $\bar{\lambda}_{\mathcal{D}_i}(t_i) \geq \sum_e \lambda^e_{\mathcal{D}_i}(s)$ for $s > t_i$.
2. Sample U from an exponential distribution with rate parameter $\bar{\lambda}_{\mathcal{D}_i}(t_i)$. Then with probability $\frac{\lambda_{\mathcal{D}_i}(t_i+U)}{\bar{\lambda}_{\mathcal{D}_i}(t_i)}$, let $\mathcal{D}_{i+1} = \mathcal{D}_i \cup \{t_i + U\}$ and go to the next step; otherwise, add another exponential random variable with rate parameter $\bar{\lambda}_{\mathcal{D}_i}(t_i + U)$ to U and repeat.
3. Let $t_{i+1} = \max \mathcal{D}_{i+1}$.
4. If $t_{i+1} \geq t_{\max}$, stop the simulation. The final dataset is $\mathcal{D}_i$.

With different initial conditions for the iteration, we can also simulate the point process conditional on some fixed initial data. This allows us to predict the distribution of future outcomes (e.g. prices) following some observed history.

Prove correctness

4.2 Simulating Order Book Events

4.2.1 Order Insertions

4.2.2 Order Cancellations

4.2.3 Modifying an Order

4.2.4 Trades

4.3 Simulation Study of Estimation Methods

Convergence analysis for simple models (eg univariate)

4.4 Impulse Response Function

Causal analysis, price impact of orders.

What does inserting a single exogenous event do to the order book, price, etc? Are there analytical formulas for this?

[20] [21] sources that use hawkes process models to simulate an order book for analysis purposes

Chapter 5

Application to KOSPI/SPY Data

Introduce

5.1 Dataset

20240920 and before - ESU4 september expiry 20240922 and after - ESZ4 december expiry

5.2 Point Process Modelling

5.2.1 Santa Fe Model

Poisson Processes Add a mixture distribution to model the arrivals to the book Make base rate depend on volume in front and distance from mid

5.2.2 Inhomogeneity

Piecewise Linear Spline

Components have the form $\lambda^e(t) = a \max(t-k, 0)$

$$
\begin{aligned}
\ell &= \sum_e \int (\log(a_e) + \log(\max(t-k,0))\, d\Lambda_e - \sum_e \int a_e \max(t-k,0) dt \\
&= \sum_e \int (\log(a_e) + \log(\max(t-k,0))\, d\Lambda_e - \frac{1}{2}(t_{\max}-k)^2 \sum_e a_e \\
\frac{\partial \ell}{\partial a_e} &= \frac{1}{a_e}\Lambda_e(t_{\max}) - \frac{1}{2}(t_{\max}-k)^2 \\
\frac{\partial^2 \ell}{\partial a_e^2} &= -\frac{\Lambda_e(t_{\max})}{a_e^2} \\
\frac{\partial \ell}{\partial k} &= -\sum_e \int \frac{1_{t>k}}{\max(t-k,0)} d\Lambda_e - (k - t_{\max}) \sum_e a_e \\
\frac{\partial^2 \ell}{\partial k^2} &= \sum_e \int \frac{1_{t>k}}{\max(t-k,0)^2} d\Lambda_e - \sum_e a_e \\
\frac{\partial^2 \ell}{\partial k \partial a_e} &= t_{\max} - k
\end{aligned}
$$

5.2.3 Hawkes Kernel

5.2.4 State Dependence

discrete and regressive, as well as 'reverse' state dependence - dependence on state just before triggered event rather than just state at trigerring event. Quantify which of these is the best. Feature selection using hessian info

More event types with 'factor' model. Marks should play as little a role as possible

State variables in original MPP paper - spread, queue imbalance.

5.2.5 Regression

5.2.6 Power Hawkes

quadratic and beyond

5.2.7 Idiosyncratic Daily Behaviour

5.3 notes

Baseline time series model - Zero-inflated gaussian or tdist mixture with GARCH vol and autoregressive midprice changes

Move from complex mark models, simple point process models → simple mark models, complex point process models. Reduces behaviour to emergence.

Models for the point process

1. Poisson Process
2. Inhomogenous Poisson Process on multiple days of data, check for residual autocorrelation (Polynomial, spline, mixture of truncated gaussians - almost KDE but not quite. Information criteria model selection.)
3. Hawkes Process with Inhomogenous Background (+ sinusoid kernel)
4. IHP with state dependence. Which variables?
5. IHP with regression - generalisation of state dependence. Again, which variables?
6. IHP with 'reverse' state dependence - depends on state just before triggered event rather than just state at triggering event. Is this worth it?
7. More event types. With 'factor' model. Marks should play as little a role as possible.
8. Quadratic and higher order
9. Day-specific metaorders with some distribution of schedules (eg volume, time, price-sensitive, liquidity-sensitive). These can insert in response to various market phenomena. This then 'explains away' some of the activity, depressing the estimates for the regular kernels. They're basically ordinary kernels of their own. But the parameters are drawn from a distribution. And there can be multiple metaorders, drawn from some mixture distribution.

Models for the marks/state

1. State dependence & markov switching
2. Gaussian mixture. Examine residuals. May need to try tdist mixture.
3. Autoregression
4. Autoregression w/ GARCH

5. Kalman Filter (on marks only. Does this work, though? Does it mess with the point process somehow?); can I have KF with heteroskedasticity somehow?
6. Options factor model. How to weight strikes?

Diagnostics

- Likelihood, parameter count, information criteria, CV likelihood
- Residual autocorrelation (Ljung-Box pval), residual distribution (KS pval)
- Microprice quality: signature plot of realised vol against time horizon
- Predicting volatility. Baseline = GARCH w/ EMA features, Baseline = implied vol.
- Predicting volume traded
- Predicting liquidity
- Price impact of trades (function of size) and passive quotes
- Check whether things have a relationship with: spread, recent volume, predicted volatility, realised volatility, etc.; include plot if needed
- Filtered vs smoothed metaorders

Model Characteristics

- Shape of background rate
- Kernel function learned for each event type against each other
- How kernel function changes throughout the day based on time-varying coefficients
- How kernel function responds to various book features
- Metaorder duration/size/impact
- TWAP and VWAP (what DMA calls percent of volume) cost for various sizes,urgencies; can also test with foreknowledge of closing price. TWAP with altered schedule over the course of a day (DMA book calls this 'tilting'). What are the recommendations here?

Strategies - The ideal situation is that these look good wrt midprice, ok wrt bid/ask spread costs, bad when market impact is considered. Don't know fees

- Market making into a spread arb
- Microprice crossing spread arb - shouldn't happen often
- Vol factors stat arb - does it just sell vol or wings?

how to quantify compute time? number of EM steps, etc

Hawkes processes and their applications to finance: a review https://www.tandfonline.com/doi/
https://www.tandfonline.com/journals/rejf20/collections/Hawkes-Processes-in-Finance

- Replicate findings from [13]
- Hidden events (& events on different exchanges) - either Poisson distributed or more complex
- Modeling changes in the entire order book
- Market impact (are there any datasets on this? square-root law, other common findings. power law impact for hawkes processes is explicitly studied here `https://arxiv.org/pdf/1805.07134`)
- Optimal execution - VWAP, TWAP, Almgren-Chriss

- Midprice change prediction/explanation - explicit formula or simulation?
- Realised volatility prediction
- Correlated products (with low beta, preferably - or see what is done in literature studies of correlated products)
- Options (if I can get data) - would give lots of (nonlinearly) correlated products. Can estimate the correlation between products at any point in time using factor loadings & historical factor correlations. Here is one source: https://www.nber.org/system/files/working$_papers/w29369/w29369.pdf.Optionspri$ https://arxiv.org/abs/1602.03043$Closedformforoptimalexecutionofsignals(includ$ $impact$)
- This paper arxiv.org/pdf/2401.11495 shows a functional limit theorem for hawkes processes behaving as integrated CIR processes. These are a popular volatility model so this makes sense!

https://quant.stackexchange.com/questions/59593/what-are-some-currently-open-p
https://papers.ssrn.com/sol3/papers.cfm?abstract$_id = 4844711OptionPricingUsingH$

A slightly depressing jump model: intraday volatility pattern simulation https://www.tandfonline.com/doi/full/10.1080/14697688.2017.1403139 Implementat and evaluation of the Heston-Queue-Hawkes option pricing model https://uu.diva

5.3.1 Clustering Ratio

Variance of interevent time divided by expectation. Bouchaud p 165. Should this section be elsewhere? Does this mean something for residuals too?

Point process control https://pages.stern.nyu.edu/ rcaldent/courses/B60.4308$_file$ *Process.pdf*

Recent Paper - Limit Order Book Dynamics and Order Size Modelling Using Compound Hawkes Process https://arxiv.org/abs/2312.08927 https://www.sciencedi

Databento info: CME: ES MES OPRA: SPX - Index options SPY - Index ETF options https://databento.medium.com/getting-futures-tick-sizes-and-notional-t

https://databento.com/portal/datasets/XNAS.ITCH/ETF/SPY https://databento.com/

https://databento.com/portal/datasets/GLBX.MDP3/Futures/ES https://databento.c

5.4 Summary of datasets

- data structure and granularity - which assets are involved

5.5 Inhomogeneity

Two choices for handling inhomogeneity: - Reparameterise to 'business time' in preprocessing, as recommended by [3] in 9.3.1 - Splines/polynomials etc - Could also combine both - Can also focus on a small part of the data. counting process should look roughly linear

5.6 Multidimensionality

poisson model

correlations in trade sign (buys vs sells arrival rate nonstationary)

trades vs inserts vs cancels

different price levels have different arrival rates -> sante fe model is wrong
how to not need a separate event type for each price level

5.7 Self-Excitation and Mutual Excitation

i.e. poisson vs hawkes

5.8 Queue Size Dependence

queue size (esp. cancellations) and also book imbalance
state-dependent process can be used here
queue-reactive process is a little bit more fit for purpose

Chapter 6

Conclusion

This is the conclusion

6.0.1 Hidden States

I might mention that people tried this but it wont be in the thesis at all. Most likely delete ‘‘Cohen and Elliott (2013) introduce a one-dimensional Markov-modulated Hawkes process following (2.4), where X is an exogenous Markov process in a finite state space (exogenous in the sense that the counting process does not influence X). The key feature of their work is that they assume X to be unobservable, leading them to derive a filtering procedure for the estimation of the current state of X.’’ (hidden states)

6.0.2 Daily Variation

Metaorders

Chapter 7

Appendix: Foundations of Probabilistic Models

In order to describe precisely the various models explored in this thesis, it is necessary to introduce some key mathematical concepts that form the basis for model specifications and related derivations. In this appendix, I cover the basics of measure, probability, and stochastic processes.

7.1 Measure Theory Fundamentals

In order to formalise the concept of a stochastic process, it is necessary to introduce the concept of a measure space.

For any set X, we can construct the power set

$$\mathcal{P}(X) = \{S : S \subseteq X\}$$

which is a new set that contains as its elements every subset S of X.

We say that a subset Σ of $\mathcal{P}(X)$ is a *σ-algebra* on X if and only if it satisfies the following three properties:

1. Containment of the full space, i.e.

$$X \in \Sigma.$$

2. Closure under complements, i.e.

$$\forall S \in \Sigma, X \backslash S \in \Sigma.$$

3. Closure under countable union, i.e.

$$\forall \{A_n\}_{n=0}^{\infty} \in \Sigma^{\mathbb{N}}, \quad \bigcup_{n=0}^{\infty} A_n \in \Sigma.$$

Elements of Σ are known as *measurable sets*. A common example of a σ-algebra is the Borel σ-algebra $B(X)$ of a topological space X (e.g. $\mathbb{R}$), defined as the smallest σ-algebra such that every open set is measurable.

We then say that a *measure* on (X, Σ) is any function $\mu : \Sigma \to \mathbb{R}_{\geq 0} \cup \{\infty\}$ that is *countably additive*, meaning that for any finite or countable sequence of disjoint sets $A_n \in \Sigma$, we have

$$\mu\left(\bigcup_n A_n\right) = \sum_n \mu(A_n).$$

Assuming that at least one set $S \in \Sigma$ has finite measure, we then have

$$\mu(S) = \mu(S \cap \emptyset) = \mu(S) + \mu(\emptyset) \Rightarrow \mu(\emptyset) = 0.$$

We refer to the combined triple (X, Σ, μ) as a *measure space*.

Informally, a measure formalises intuitions about the size, mass, or significa of a set of points. For instance, a set $S \subseteq X$ with measure zero is known as a *null set*, and a property that holds only for points in a null set is said to be true *almost nowhere* in X. Conversely, a property that holds for every point except those in a null set is said to be true *almost everywhere* (a.e.) in X. In this way, a measure quantifies how significant or negligible the exceptions to a heuristic principle may be.

Similarly, familiar concepts of *length*, *area* and *volume* are all formalised by a family of translation-invariant measures on $\mathbb{R}^n$, known as the n-dimensiona *Lebesgue measures*.

The concept of a measure is foundational to the definition of the Lebesgue integral. Integration of a function f over a measurable set S with respect to a measure μ is written as

$$\int_{t \in S} f(t) d\mu(t),$$

while integration with respect to the Lebesgue measure will often be written simply as

$$\int_S f(t) dt.$$

I will not cover the definition of the Lebesgue measure or Lebesgue integral here, but they can be found in most textbooks on measure theory.

7.2 Probability Spaces

In the special case where $\mu(X) = 1$, we refer to μ as a *probability measure*, and to X as a *sample space*. Correspondingly, the term *almost everywhere* is replaced with *almost surely* (a.s.), indicating a property that holds on a set of points with probability measure one.

A function from one σ-algebra to another is called *measurable* if and only if the preimage of any measurable set in the codomain is a measurable set in the domain. A measurable function whose domain is a sample space Ω equipped with a probability measure $\mathbb{P}$ is known as a *random variable*.

Measurable sets in a sample space are often referred to as *events*, and the measure of such a set is called the probability of the event. For instance, the set S of points $\omega \in \Omega$ for which a random variable $X : \Omega \to \mathbb{R}$ satisfies a particular property $P : \mathbb{R} \to \{\text{True}, \text{False}\}$ will have measure equal to the probability of that property being true.[1]

[1] The predicate P must be measurable.

A more general concept is ***expectation***. For a real-valued random variable $X : \Omega \to \mathbb{R}$, we define the expectation of X with respect to $\mathbb{P}$ to be the linear functional

$$\mathbb{E}_{\mathbb{P}}[X(\omega)] \coloneqq \int_{\omega \in \Omega} X(\omega) d\mathbb{P}(\omega).$$

For binary-valued random variables $X : \Omega \to \{0, 1\}$, we have the identity

$$\mathbb{E}_{\mathbb{P}}[X(\omega)] = \mathbb{P}\left(\{\omega \in \Omega : X(\omega) = 1\}\right).$$

7.2.1 Conditionalisation

By defining ‘‘events" as Σ-measurable subsets of Ω, we allow knowledge of an event to inform us about which possible values of the hidden ω could have produced the result we actually observe. In this way, the knowledge that ω has produced one observed event allows us to draw probabilistic conclusions about the occurrence of a different, related event. For instance, learning that a die has rolled an even number tells us that it cannot possibly have rolled a 3, and makes the proposition that a 2 has been rolled a more reasonable guess than otherwise.

In order to capture the relationships between events, and to make systematic inferences about hidden events from observed ones, it is necessary to describe mathematical rules for ***conditionalisation***.

Given two σ-algebras Σ_1, Σ_2 on ω, we say that Σ_1 is a ***sub-σ-algebra*** of Σ_2 if and only if $\Sigma_1 \subseteq \Sigma_2$. In the case where $\Sigma_1 \neq \Sigma_2$, Σ_1 is said to be ***coarser*** than Σ_2, in the sense that Σ_2 contains events that cannot be expressed as Σ_1-measurable sets. Knowing which Σ_2-events the hidden ω falls under can therefore give us more information about the exact value of ω.

One common example of a sub-σ-algebra arises by considering the set

$$\{X^{-1}(A) : A \in \mathbb{R}\} \subseteq \mathcal{P}(\Omega),$$

consisting of all the preimages of Borel-measurable sets under a random variable $X : \Omega \to \mathbb{R}$. This is known as the σ-algebra generated by X.

expectation conditional on sigma algebra https://math.stackexchange.com/questi
expectation conditional on an event

probability is expectation of indicator function (notation for indicator function is 1_A)

random measure

What is independence independence of RVs if and only if independence of generated sigma algebras

Conditional probability

7.3 Density of a Measure

Absolute continuity: $\mu_1 \ll \mu_2$ (mu_1 is dominated by μ_2) if and only if $\mu_2(A) = 0 \Rightarrow \mu_1(A) = 0$ for every measurable set A. Can also say that one measure is absolutely continuous wrt another on a subset of the measure space Radon-Nikodym Theorem

7.4 Stochastic Process Fundamentals

What is a filtration

What is a stochastic process

What is the natural filtration

What is a realisation of a stochastic process

What is a cadlag The order book is a cadlag over time. Also the counting process is a cadlag.

What is a martingale - do I need this?

References

[1] Palley, T. I., *Financialization: What It Is and Why It Matters*, Working Paper No. 525, The Levy Economics Institute, December 2007. Paper presented at the conference on \Finance-led Capitalism? Macroeconomic Effects of Changes in the Financial Sector," sponsored by the Hans Boeckler Foundation, Berlin, Germany, October 26{27, 2007. Available at: https://www.levyinstitute.org/pubs/wp_525.pdf

[2] Balakrishnan, S., Wainwright, M. J., and Yu, B., *Statistical guarantees for the EM algorithm: From population to sample-based analysis*, *The Annals of Statistics*, 45(1) (2017), 77--120. doi = 10.1214/16-AOS1435 various facts about EM algorithm in general, useful in particular because it shows the general principle that likelihood gradient = EM gradient. balakrishnan, wainwright, yu page 82 as cited at https://stats.stackexchange.com/questions/45652/what-is-the-difference-b

[3] Bouchaud, J.-P., Bonart, J., Donier, J., and Gould, M., *Trades, Quotes and Prices: Financial Markets Under the Microscope*, Cambridge University Press, 2018. ISBN = 9781107156050

[4] Chen, F. and Stindl, T., *Direct Likelihood Evaluation for the Renewal Hawkes Process*, *Journal of Computational and Graphical Statistics*, 27(1) (2018), 119--131. doi = 10.1080/10618600.2017.1341324

[5] Daley, D. J. and Vere-Jones, D., *An Introduction to the Theory of Point Processes: Volume I: Elementary Theory and Methods*, 2nd edition, Springer, 2003. doi = 10.1007/b97277

[6] Embrechts, P., Liniger, T., and Lin, L., *Multivariate Hawkes processes: An application to financial data*, Journal of Applied Probability, 48(A) (2011), 367--378. doi = 10.1239/jap/1318940477

[7] Financial Industry Regulatory Authority, *2024 Industry Snapshot: Market Data*, FINRA, 2024. Available at: https://www.finra.org/media-center/reports-studies/2024-industry-snapsho Accessed: 2 September 2024.

[8] Jamshidian, M. and Jennrich, R. I., *Acceleration of the EM Algorithm by Using Quasi-Newton Methods*, *Journal of the Royal*

Statistical Society. Series B (Methodological), 59(3) (1997), 569--587. Available at: http://www.jstor.org/stable/2346010

[9] Jamshidian, M. and Jennrich, R. I., *Standard errors for EM estimation*, *Biometrika*, 89(1) (2002), 63--75. doi = 10.1111/1467-9868.00230

[10] Jiang, A. Z. and Rodriguez, A., *Improvements on Scalable Stochastic Bayesian Inference Methods for Multivariate Hawkes Processes*, *Statistics and Computing*, 34, article 85 (2024). doi = 10.1007/s11222-024-10392-x

[11] Laub, P. J., Lee, Y., and Taimre, T., *The Elements of Hawkes Processes*, Springer, 2021. doi = 10.1007/978-3-030-84639-8

[12] Lewis, E. and Mohler, G., *A nonparametric EM algorithm for multiscale Hawkes processes*, *Journal of Nonparametric Statistics*, 1(1) (2011), 1--20. one example of EM algorithm applied to point processes. may or may not deserve a direct citation, but can see if anything cited here is useful

[13] Morariu-Patrichi, M. and Pakkanen, M. S., *State-Dependent Hawkes Processes and Their Application to Limit Order Book Modelling*, Quantitative Finance, 22(3) (2022), 563--583 doi = 10.1080/14697688.2021.1983199

[14] Nickel, M. and Le, M., *Learning Multivariate Hawkes Processes at Scale*, arXiv preprint arXiv:2002.12501 [cs.LG], 2020. doi = 10.48550/arXiv.2002.12501

[15] Smith, A. C. and Brown, E. N., *Estimating a state-space model from point process observations*, *Neural Computation*, 15(5) (2003), 965--991. doi = 10.1162/089976603765202622

[16] Veen, A. and Schoenberg, F. P., *Estimation of Space-Time Branching Process Models in Seismology Using an EM-Type Algorithm*, *Journal of the American Statistical Association*, 103(June) (2008), 614--624. doi = 10.1198/016214508000000148

[17] Zhou, F., Li, Z., Fan, X., Wang, Y., Sowmya, A., and Chen, F., *Efficient inference for nonparametric Hawkes processes using auxiliary latent variables*, *The Journal of Machine Learning Research*, 21(1) (2020), 9745--9775.

Unformatted:

[18] On Lewis' simulation method for point processes https://ieeexplore.ieee.org/document/1056305

[19] Ogata paper https://www.ism.ac.jp/editsec/aism/pdf/030$_{20}$243.*pdf*

[20] LONG TIME BEHAVIOUR OF A HAWKES PROCESS-BASED LIMIT ORDER BOOK https://hal.science/hal-01121711v5/document

[21] https://link.springer.com/chapter/10.1007/978-88-470-1766-5$_4$

https://www.idescat.cat/sort/sort461/46.1.1.Worrall-etal.pdf

https://www.tandfonline.com/doi/full/10.1080/10618600.2022.2050247

https://www.santafe.edu/research/results/working-papers/studies-of-the-limit-o

https://arxiv.org/pdf/1903.03223
https://link.springer.com/chapter/10.1007/978-88-470-1766-5$_4$
https://www.jstor.org/stable/2334319
https://www.tandfonline.com/doi/full/10.1080/14697688.2021.1983199
https://www.tandfonline.com/doi/full/10.1080/10618600.2017.1341324
https://www.researchgate.net/publication/4742983$_E stimation_o f_S paceTime_B ranching_P$
https://pubmed.ncbi.nlm.nih.gov/12803953/
http://paleo.sscnet.ucla.edu/Lewis-Molher-EM$_P reprint.pdf$
https://opus.lib.uts.edu.au/bitstream/10453/145676/2/19-930.pdf
https://projecteuclid.org/journals/annals-of-statistics/volume-45/issue-1/Stat
https://www.tandfonline.com/doi/full/10.1080/1351847X.2021.1917441
https://projecteuclid.org/journals/annals-of-probability/volume-24/issue-3/Sta
Data Sketching for Large-Scale Kalman Filtering
https://arxiv.org/pdf/1606.08136