POINT PROCESS MODELLING OF A LIMIT ORDER BOOK

Oden Petersen

Supervisor: Dr. Tom Stindl

School of Mathematics and Statistics
UNSW Sydney

October 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

On a single day in 2023, the US stock market saw an average of around $500 billion dollars worth of shares traded on various exchanges and other venues, exceeding the annual GDP of a typical European country [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, as illustrated in figure 1.1.

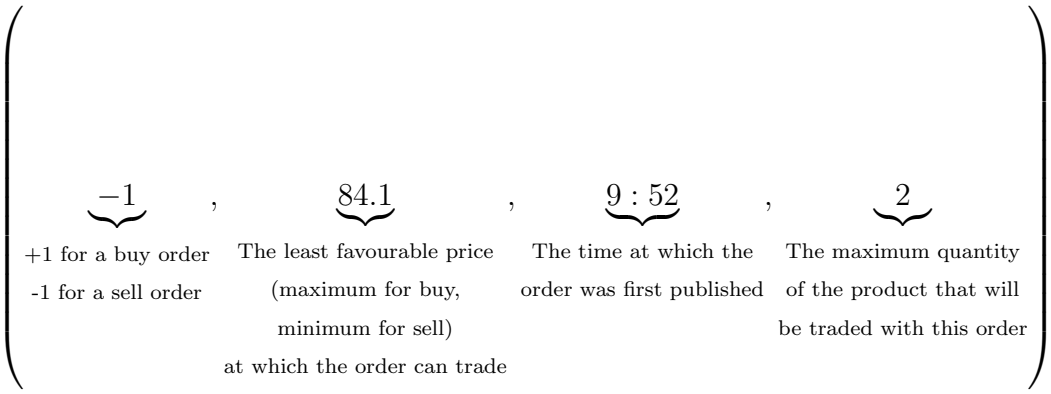


Figure 1.1: Components of an example limit order

1.2 The Matching Algorithm

1.2.1 The Bid and Ask

When an order insertion message is received, the exchange will attempt to form trades with the existing orders in the book such that as much volume as possible is matched. Any volume that cannot be matched will be added to the orders already in the book.

As a result of this matching, the total volume posted to the book may be depleted over time, even in the absence of cancellations. Any time a buy order has a price equal to or greater than that of a sell order, volume will be matched and a trade will occur. Consequently, the most competitive buy price (known as the *bid*) will always be less than the most competitive sell price (known as the *ask*), i.e.

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

An excessively high bid price will be depleted by traders seeking to sell the product at a premium to its true value. Conversely, an excessively low ask price will be pushed 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, and it is common to use these as prediction targets in the construction of trading signals. I might discuss them later

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

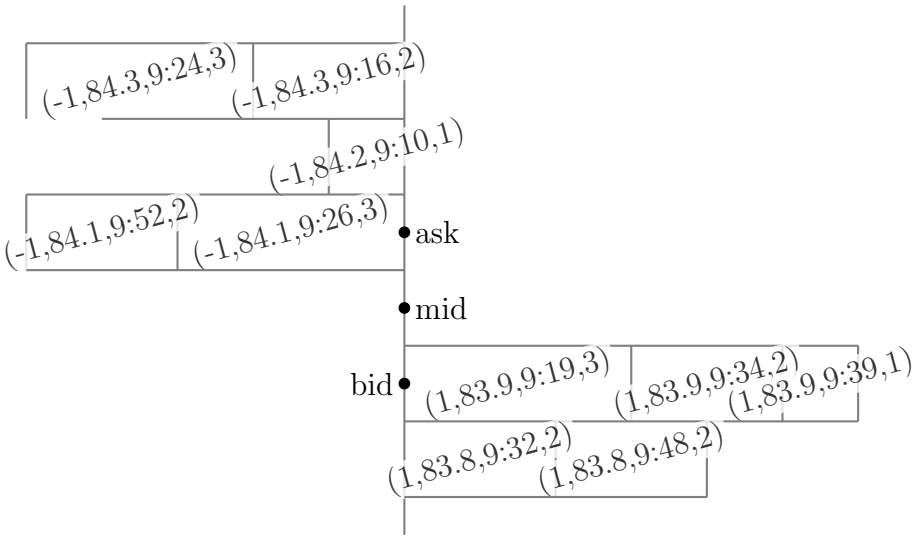


Figure 1.2: An example order book, arranged by order price and time

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.

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

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

Notably, if $\text{mid}_{\mathcal{L}}$ is taken to represent the fair value of the product, 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 components and measures of liquidity which I may discuss later.

1.2.3 Price-Time Priority

Explain

1.3 Queueing Models of the Order Book

Queue depletions are related to price changes [3]

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`

1.4 Overview of Point Processes

With the goal of modelling the arrival time process, I will now provide an introductory overview of point processes. In the next section, I will highlight some key variants described in existing literature 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.

1.4.1 Definition

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

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

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.

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

1.4.2 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

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

Is this a measure on (T, Σ_T)? Since subsets of disjoint sets are also disjoint, the terms in the union on the right-hand side will be disjoint, and so

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

So this is indeed a measure.

1.4.3 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*,

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

Is this a measure on (T, Σ_T)? 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.

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.

1.4.4 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

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

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

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

as the left limit of the expectation density as the filtration index approaches t.

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

known as the *intensity* or *arrival rate* of the point process. This represents our best estimate of the number of events that will arrive in the next ϵ units of time, for very small ϵ, based on all the information contained in the filtration up to t.

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

1.4.5 Residuals and the Compensator Process

Define residuals, explain why they're $Exp(1)$ distributed Notation for i-th residual is just r_i They're independent according to theorem 3.3 from MORARIU-PATRICHI, M. AND PAKKANEN, M. S.

Define the compensator process $\bar{\Lambda}(t, \omega) = \int_{t_{\min}}^{t} \lambda(s, \omega) ds$ for a point process with intensity process $\lambda(t)$. Note properties of this. I make use of this notation in the estimation chapter.

1.4.6 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 $\mathcal{X}$ is partitioned into a finite number of *event types*, in which case we may refer to $\mathcal{T}$ 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 in vector form, with one entry for each event type.

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.

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

1.5 Point Process Models of Book Updates

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

1.5.1 Poisson Processes

Probably the simplest modeling assumption we can make is to assume that the intensity of the process is a constant $\nu \in \mathbb{R}$.

A Poisson process is a point process for which 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?

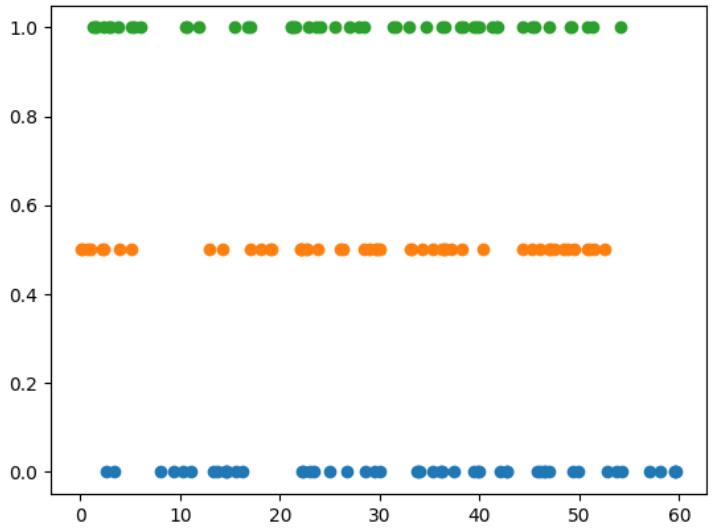
In the multivariate case with n event types, we have a vector $\nu \in \mathbb{R}^n_{>0}$ containing the constant intensities of each individual point process.

This definition might require a reference.

Is there a proof of existence and nonexplosiveness?

Explain why the times between events are exponentially distributed. Can use the linearity of residuals to prove this succinctly, since residuals are exponentially distributed.

There is a link here to the sante fe order book model, described in the Bouchaud source, which is an example of the application of poisson arrivals.



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.

1.5.2 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. Is this weaker version even needed, or is it sufficient to just drop the condition entirely?

As a result, although the intensity $\lambda(t, \omega)$ is constant in ω (i.e. deterministic) how do we know it is?, it can vary as a function of t. This is useful for modeling

deterministic seasonality in the point process, such as having more events near the start/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

1.5.3 Composite Point Processes

Define these in a way that leads into hawkes processes. They should be adapted to the natural filtration on the overall point process (NOT the individual point processes, this would preclude interaction between them)

1.5.4 Autoregressive Point Processes

In the case of the homogenous and inhomogenous poisson processes, the intensity function is deterministic. While such an assumption 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 recency of previous event occurrences. One explicit test for this is to examine the autocorrelation of residuals, 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

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

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

where $\nu \in \mathbb{R}_{>0}$ represents a constant arrival rate, k is a kernel function encoding the self-excitation behaviour of each event, and $t_{\min}$ is the earliest time events can occur, such as the opening time of a trading session. Existence of such processes is established... where?

Long-run mean

$$\overline{\lambda} = \lim_{t \to \infty} \mathbb{E}\left[\lambda(t)\right]$$

$$\bar{\lambda} = \nu + \bar{\lambda} \int_{0}^{\infty} k(t) dt$$

Stability requires $\int_0^\infty k(t)dt < 1$. Not sure if sufficient condition.

[1] In the case of multivariate point processes, interactions between 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*.

Multidimensional:

$$\bar{\lambda} = \nu + K\bar{\lambda}$$

$$\bar{\lambda} = (I - K)^{-1}\nu$$

Requires $|\det K| < 1$ (why? and what about if there are directions orthogonal to ν?) Define K

Why use exponential kernel? Possible reasons why others might be more correct eg polynomial kernel captures long memory better Special case of $\beta = 0$ may be worth including in the final kernel. (Integrated point process?)

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.

1.5.5 Quadratic Hawkes Processes

We'll see

1.5.6 State Dependence

Vinovskaya proposes regime-switching independent of the point process Moriaru-Patrichi and Pakkanen generalise this by coupling state transitions with the event sequence. They consider conditioning on only the state of the triggering event (their 2.3), as well as conditioning on both the state of the triggering event and the current state (their 4.2). Moriaru-Patrichi and Pakkanen find that conditioning on the state of the order book improves the fit of the model. (How?)

"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)

" This two-way interaction between N and X makes them fully coupled, just like the order flow and the state of the LOB in an order-driven market. It also distinguishes the model from the existing regime switching Hawkes processes (Wang et al., 2012; Cohen and Elliott, 2013; Vinkovskaya, 2014; Swishchuk and Huffman, 2020), where the state process evolves exogenously, receiving no feedback from the point process"

The first of these:

$$\tilde{\lambda}_{ex}(t) = \phi_e(X(t), x)\left(\nu_e + \sum_{e' \in \mathcal{E}, x' \in \mathcal{X}} \int_{[0,t)} k_{e',e}(t-s, x') d\tilde{N}_{e'x'}(s)\right), t \geq 0, e \in \mathcal{E}, x \in \mathcal{X}.$$

The second:

$$\tilde{\lambda}_{ex}(t) = \nu_{ex} + \sum_{e' \in \mathcal{E}, x' \in \mathcal{X}} \int_{[0,t)} k_{e'x'ex}(t-s) d\tilde{N}_{e'x'}(s), t \geq 0, e \in \mathcal{E}, x \in \mathcal{X}.$$

1. markov switching 2. linear state space model / kalman filter (we'll see if i get around to figuring out how to work this)

1.5.7 *Regression on marks*

Multiply the kernel by $\exp(X \cdot \beta)$ where X is a vector of mark variables and β are coefficients. transfer functions should be separable, see [6.25 Definition] in https://www.research-collection.ethz.ch/bitstream/handle/20.500.11850/151886/eth-1112-02.pdf

1.5.8 *Adding structure to the event space*

For a finite number of event types (e.g. 10 book levels) this just corresponds to constraining the coefficient matrices.

For a continuous event space it is a bit more complicated

Another way to look at this is just having marks but the distribution of the marks is not constant - this works too. This approach is nice because it admits Kalman filtering. (Is it possible to mix Kalman filtering in event time, volume time, and wall time? Maybe better to just use wall time though.)

1.5.9 *Hidden Events*

Seems relevant https://www.cambridge.org/core/journals/advances-in-applied-probability/arti
elementary-derivation-of-moments-of-hawkes-processes/79A7355542F08087C8AE828C664A304
Hawkes Process Inference with Missing Data https://aaai.org/papers/12116-hawkes-process-inference-with-missing-data/ makes use of monte carlo EM algorithm https://www.jstor
Hawkes processes with hidden marks https://www.tandfonline.com/doi/full/10.1080/1351847X

Meta-orders. Kernel is product of an ordinary kernel and an EMA on the meta-order hidden events. See model in [3] 10.4.3.

Could also be used to correct for the quasi-EM approximation.

1.6 the most general model

Here for ideas/reference but I dont think this section will be in the final report

State-dependent multi-kernel multivariate hawkes process with increasing kernel components and possible inhibition.

Nonparametric estimate of nu, sim study for u-shaped true nu. Probably splines, maybe kde if its fast enough. Compare to quadratic regression

Simple: the marks evolve according to state-space model involving various model variables, but have no causal impact on the point process (though obviously there is state-dependence, and the state may include the marks). More complicated: hidden marks influence the point process; evaluate this with monte carlo EM.

The state might be generalised to a kalman-filter-ish model [15]

Intensity might be affected by noise introduced at each event time. Probably can use monte carlo EM for this.

Quadratic hawkes process

$$IRF_{e,e',k}(s,t) = \alpha_{e,e',k} e^{X(s)\cdot c_{e,e',k}} \exp\left(-\beta_{e,e',k}(t-s)\right)$$

$$\lambda_{e,e',k}(t,\omega) = \int_0^t IRF_{e,e',k}(s,t) d\Lambda_{e',\omega}(s)$$

$$\lambda_e(t,\omega) = \sum_{e'} \sum_{k} \lambda_{e,e',k}(t,\omega) + \text{spline intensity}(t)$$

Chapter 2

Estimation

To enable analysis of empirical order book data, I will now explain the principles involved in parametric estimation of point process models.

We 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.

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 $\mathcal{D}_t$. How is this well-defined? The point process realising $\mathcal{D}_t$ exactly has probability zero, and probability-zeor events cant be conditioned on. I suspect the solution here is to construct it using the janossy density. In particular, instead of conditioning on $\mathcal{D}_t$, it is proper to condition on $\mathcal{D}_t$ being in a certain event set and having a certain number of events, then use this to define a measure which $\lambda_{\mathcal{D}}$ is constructed as a density of

For the purposes of defining the likelihood function, the janossy density also needs to be derived or cited

$$p(\mathcal{D}) = \exp\left(\int_{t_{\min}}^{t_{\max}} \log(\lambda_{\mathcal{D}}(t))d\Lambda_{\mathcal{D}} - \int_{t_{\min}}^{t_{\max}} \lambda_{\mathcal{D}}(t)dt\right).$$

2.1 Branching Probabilities

I originally wrote this for hawkes processes, but can be made more general now since I'm gonna describe what it means to condition on $\mathcal{D}, \mathcal{F}$ above, and these are just probabilities wrt this conditional measure. Endo/exo can still be given as an example. If I don't make it more general I need to make it clear that it's only relevant to hawkes processes. It is common to regard events in the point process as having either been generated by a "background" poisson process with arrival rate ν, or else "triggered" by some previous event. The first kind are known as *exogenous events*, while the second kind are *endogenous events*. From the form of the intensity, it follows why? that the probability of an event at time t being from the background process, conditioning on the natural filtration $\mathcal{F}_t$,Is this defined properly? will be

$$B_{\text{exo},t} \coloneqq \frac{\nu}{\lambda(t,\omega)},$$

while the probability of it being endogenous will be

$$B_{\text{endo},t} \coloneqq \int_{t_{\min}}^{t} B_{s,t} d\Lambda_{\omega}(s)$$

where

$$B_{s,t} \coloneqq \frac{k(t-s)}{\lambda(t,\omega)}$$

is the probability of an event at t having been triggered by an event at s.

Notice the dependence on ω, implying that these probabilities (often known as *branching probabilities*) are specific to a particular realisation of the process.

Observe that these are truly probabilities for composite point processes, but if some components of intensity are negative this is no longer the correct interpretation; nonetheless we will still use the B notation

2.2 Maximum Likelihood Estimation

Given a parametric family of multivariate point processes,

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

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

Letting $p_\theta(\mathcal{D})$ be the probability density of an observed dataset $\mathcal{D}$ under $\mathcal{T}_\theta$, we can write the *likelihood function* for the dataset

$$L(\theta; \mathcal{D}) \coloneqq p_\theta(\mathcal{D}) = \exp\left(\int_{t_{\min}}^{t_{\max}} \log(\lambda^\theta_{\mathcal{D}}(t))d\Lambda_{\mathcal{D}} - \int_{t_{\min}}^{t_{\max}} \lambda^\theta_{\mathcal{D}}(t)dt\right),$$

where $\lambda^\theta_{\mathcal{D}}$ represents the empirical intensity function of $\mathcal{T}_\theta$.

2.2.1 *Maximum Likelihood Estimator*

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

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

this is known as the *maximum likelihood estimate*, and is a popular method of selecting a single "best" model from a parametric family. Discuss why and what properties it has. Ogata proves consistency and asymptotic normality for stationary processes estimate = on a particular dataset. estimator = considered as a random variable Ogata paper https://www.ism.ac.jp/editsec/aism/pdf/030$_{20}$243.*pdf*
Is it biased? preferably downward, this is nice for hidden events to stop the event count from increasing too much
Because finding global maxima is in general a hard problem, it is common to first weaken our search to finding stationary points of L using its first derivative with respect to θ. We can then use second-order information to eliminate local minima and saddle points, and select the best local maxima among those found in hopes that it is either the global maxima, or at least has a similar enough likelihood to be a good model choice.

2.2.2 *Log-Likelihood Maximisation*

Observe that we can write L in the form

$$L(\theta; \mathcal{D}) = \exp(\ell(\theta; \mathcal{D}))$$

where $\ell(\theta; \mathcal{D})$ is the *log-likelihood function*,

$$\begin{aligned}
\ell(\theta; \mathcal{D}) &\coloneqq \log(p_\theta(\mathcal{D})) \\
&= \int_{t_{\min}}^{t_{\max}} \log(\lambda^\theta_{\mathcal{D}}(t)) d\Lambda_{\mathcal{D}} - \int_{t_{\min}}^{t_{\max}} \lambda^\theta_{\mathcal{D}}(t) dt \\
&= \int_{t_{\min}}^{t_{\max}} \log(\lambda^\theta_{\mathcal{D}}(t)) d\Lambda_{\mathcal{D}} - \bar{\Lambda}\left(t_{\max}\right).
\end{aligned}$$

Then, by the chain rule, it follows that

$$\frac{\partial}{\partial\theta} L(\theta; \mathcal{D}) = \left(\frac{\partial}{\partial\theta} \ell(\theta; \mathcal{D})\right) \exp\left(\ell(\theta; \mathcal{D})\right).$$

Because $\exp\left(\ell(\theta; \mathcal{D})\right) > 0$ for all θ, the left-hand side is zero exactly when

$$\frac{\partial}{\partial\theta} \ell\left(\theta; \mathcal{D}\right) = 0,$$

meaning that the stationary points of L and ℓ coincide. This is also clear if we consider that log is a monotonic function, meaning that if L has a unique maxima in a region, it will also be the unique maxima of ℓ in that region.

Furthermore, we have

$$\begin{aligned}
\frac{\partial}{\partial\theta}\ell(\theta;\mathcal{D}) &= \int_{t_{\min}}^{t_{\max}} \frac{\partial}{\partial\theta}\log(\lambda_{\mathcal{D}}^{\theta}(t))d\Lambda_{\mathcal{D}} - \int_{t_{\min}}^{t_{\max}} \frac{\partial}{\partial\theta}\lambda_{\mathcal{D}}^{\theta}(t)dt \\
&= \int_{t_{\min}}^{t_{\max}} \frac{\frac{\partial}{\partial\theta}\lambda_{\mathcal{D}}^{\theta}(t)}{\lambda_{\mathcal{D}}^{\theta}(t)}d\Lambda_{\mathcal{D}} - \int_{t_{\min}}^{t_{\max}} \frac{\partial}{\partial\theta}\lambda_{\mathcal{D}}^{\theta}(t)dt, \\
\frac{\partial^2}{\partial\theta^2}\ell(\theta;\mathcal{D}) &= \int_{t_{\min}}^{t_{\max}} \frac{\frac{\partial^2}{\partial\theta^2}\lambda_{\mathcal{D}}^{\theta}(t) - \left(\frac{\partial}{\partial\theta}\lambda_{\mathcal{D}}^{\theta}(t)\right)\left(\frac{\partial}{\partial\theta}\lambda_{\mathcal{D}}^{\theta}(t)\right)^T}{\lambda_{\mathcal{D}}^{\theta}(t)}d\Lambda_{\mathcal{D}} - \int_{t_{\min}}^{t_{\max}} \frac{\partial^2}{\partial\theta^2}\lambda_{\mathcal{D}}^{\theta}(t)dt.
\end{aligned}$$

why can we bring the derivative inside the integral?

2.2.3 Poisson Processes

Suppose we have a family of poisson process models parameterised by their constant intensity $\nu \in \mathbb{R}_{>0}$.
The log-likelihood of each model is then given by

$$\begin{aligned}
\ell(\nu;\mathcal{D}) &= \int_{t_{\min}}^{t_{\max}} \log(\nu)d\Lambda_{\mathcal{D}} - \int_{t_{\min}}^{t_{\max}} \nu dt \\
&= \log(\nu)\int_{t_{\min}}^{t_{\max}} d\Lambda_{\mathcal{D}} - \nu\int_{t_{\min}}^{t_{\max}} dt \\
&= \log(\nu)N_{\mathcal{D}}\left(t_{\max}\right) - \nu\left(t_{\max} - t_{\min}\right).
\end{aligned} \tag{2.2.1}$$

At stationary points, we will then have

$$\frac{\partial}{\partial\nu}\ell(\nu;\mathcal{D}) = \frac{1}{\nu}N_{\mathcal{D}}\left(t_{\max}\right) - \left(t_{\max} - t_{\min}\right) = 0,$$

implying that

$$\hat{\nu}_{\text{MLE}} = \frac{N_{\mathcal{D}}(t_{\max})}{t_{\max} - t_{\min}}.^{1} \tag{2.2.2}$$

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

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

In the multivariate case, where we have a vector $\nu \in \mathbb{R}_{>0}^n$ containing the arrival rates for each event type E, we similarly obtain

$$\left(\hat{\nu}_{\text{MLE}}\right)_E = \frac{N_{\mathcal{D}}^E\left(t_{\max}\right)}{t_{\max} - t_{\min}}.$$

[1] Subject to the Hessian of ℓ being negative definite, which is true here.

2.2.4 *Composite Processes*

Suppose that the intensity of each point process $\mathcal{T}_{\theta_1,\theta_2,\dots}$ in a parametric family has the form

$$\lambda^{\theta_1,\theta_2,\cdots}(t,\omega) \coloneqq \sum_i \lambda_i^{\theta_i}(t,\omega),$$

where each component $\lambda_i(t,\omega)$ is parameterised by some vector θ_i and adapted with respect to the natural filtration $\mathcal{F}$ of $\mathcal{T}_{\theta_1,\theta_2,\dots}$. Then the loglikelihood will be

$$\ell(\theta_1,\theta_2,\dots;\mathcal{D}) = \int_{t_{\min}}^{t_{\max}} \log\left(\sum_i \left(\lambda_i^{\theta_i}\right)_{\mathcal{D}}(t)\right) d\Lambda_{\mathcal{D}}(t) - \sum_i \bar{\Lambda}_i\left(t_{\max}\right),$$

where

$$(\lambda_i^{\theta_i})_{\mathcal{D}} \coloneqq \lim_{s\to t^+} \mathbb{E}_{\mathbb{P}|\mathcal{F}_t,\mathcal{D}_t}\left[\lambda_i^{\theta_i}(s)\right]$$

Is this the right definition to use? it should coincide with component intensities in the case of a composite point process, as noted below. It seems to depend on continuity of the first moment, and so maybe more restrictions need to be placed on the component processes

and

$$\bar{\Lambda}_i(t) \coloneqq \int_{t_{\min}}^{t} \left(\lambda_i^{\theta_i}\right)_{\mathcal{D}}(s)ds.$$

Define

$$B_{i,t}^{\theta_1,\theta_2,\cdots} \coloneqq \frac{\left(\lambda_i^{\theta_i}\right)_{\mathcal{D}}(t)}{\lambda_{\mathcal{D}}^{\theta_1,\theta_2,\cdots}(t)}$$

to be the ratios of each component to the overall intensity.

By the chain rule, we then have Why can I exchange derivative and integral here

$$\begin{aligned}
\frac{\partial}{\partial\theta_i}\ell(\theta_1,\theta_2,\dots;\mathcal{D}) &= \int_{t_{\min}}^{t_{\max}} \frac{\partial}{\partial\theta_i}\log\left(\sum_i \left(\lambda_i^{\theta_i}\right)_{\mathcal{D}}(t)\right) d\Lambda_{\mathcal{D}} - \frac{\partial}{\partial\theta_i}\bar{\Lambda}_i(t_{\max}) \\
&= \int_{t_{\min}}^{t_{\max}} \frac{\frac{\partial}{\partial\theta_i}\left(\lambda_i^{\theta_i}\right)_{\mathcal{D}}(t)}{\lambda_{\mathcal{D}}^{\theta_1,\theta_2,\cdots}(t)} d\Lambda_{\mathcal{D}} - \int_{t_{\min}}^{t_{\max}} \frac{\partial}{\partial\theta_i}\left(\lambda_i^{\theta_i}\right)_{\mathcal{D}}(t)dt \\
&= \int_{t_{\min}}^{t_{\max}} \frac{\frac{\partial}{\partial\theta_i}\left(\lambda_i^{\theta_i}\right)_{\mathcal{D}}(t)}{\left(\lambda_i^{\theta_i}\right)_{\mathcal{D}}(t)} B_{i,t} d\Lambda_{\mathcal{D}} - \int_{t_{\min}}^{t_{\max}} \frac{\partial}{\partial\theta_i}\left(\lambda_i^{\theta_i}\right)_{\mathcal{D}}(t)dt \\
&= \int_{t_{\min}}^{t_{\max}} \left(\frac{\partial}{\partial\theta_i}\log\left(\left|\left(\lambda_i^{\theta_i}\right)_{\mathcal{D}}(t)\right|\right)\right)\left(|B_{i,t}|\, d\Lambda_{\mathcal{D}}\right) - \int_{t_{\min}}^{t_{\max}} \frac{\partial}{\partial\theta_i}\left(\lambda_i^{\theta_i}\right)_{\mathcal{D}}(t)dt,
\end{aligned}$$

Can the last equation be written in terms of log?

Fixing B and iteratively maximising (or newtons method, or gradient steps) can only stop at a stationary point; these are the fixed points of such a procedure. If the Bs are positive then convergence is guaranteed, covered in next section. Also notice that expected gradient is equal to ℓ gradient, meaning that small improvements to $E[\ell]$ should also improve ℓ. So it makes sense to iteratively maximise, and hope for convergence.

2.3 Expectation Maximisation

If $\left(\lambda_i^{\theta_i}\right)_{\mathcal{D}}$ are positive,[2] 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

2.3.1 EM for State-Dependent Hawkes Processes

- Equations
- 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)

2.4 Inference for the Mark Process

Kalman filter? Ways to couple this with the point process? (Monte carlo EM?)

2.5 Hidden Events

$$\begin{aligned}
&\mathbb{P}^\theta(t \in \mathcal{T} | \mathcal{D}, \mathcal{H}_t) \\
=&\mathbb{P}^\theta(t \in \mathcal{T} | \mathcal{D}_t, \mathcal{H}_t) \frac{\mathbb{P}^\theta(\mathcal{D}_{t\to T} | t \in \mathcal{T}, \mathcal{D}_t, \mathcal{H}_t)}{\mathbb{P}^\theta(\mathcal{D}_{t\to T} | t \in \mathcal{T}, \mathcal{D}_t, \mathcal{H}_t)\lambda_{\mathcal{D},\mathcal{H}} dt + \mathbb{P}^\theta(\mathcal{D}_{t\to T} | t \notin \mathcal{T}, \mathcal{D}_t, \mathcal{H}_t)(1 - \lambda_{\mathcal{D},\mathcal{H}} dt)} \\
=&\mathbb{P}^\theta(t \in \mathcal{T} | \mathcal{D}_t, \mathcal{H}_t) \frac{\mathbb{P}^\theta(\mathcal{D}_{t\to T} | t \in \mathcal{T}, \mathcal{D}_t, \mathcal{H}_t)}{\mathbb{P}^\theta(\mathcal{D}_{t\to T} | t \notin \mathcal{T}, \mathcal{D}_t, \mathcal{H}_t)}
\end{aligned}$$

$$\lambda^\theta_{\mathbb{P}|\mathcal{D},\mathcal{H}_t} = \lambda^\theta_{\mathcal{D},\mathcal{H}} \frac{\mathbb{P}^\theta(\mathcal{D}_{t\to T} | t \in \mathcal{T}, \mathcal{D}_t, \mathcal{H}_t)}{\mathbb{P}^\theta(\mathcal{D}_{t\to T} | t \notin \mathcal{T}, \mathcal{D}_t, \mathcal{H}_t)}$$

[2] In the case where $\lambda^{\theta_1,\theta_2,\cdots}$ is the intensity of a composite point process whose components are adapted to the natural filtration of $\mathcal{T}_{\theta_1,\theta_2,\ldots}$, and have intensities $\lambda_i^{\theta_i}$, the functions $\left(\lambda_i^{\theta_i}\right)_{\mathcal{D}}$ will just be the empirical intensity of the componentsIs this true? is there a simple proof?, and therefore will each be positive definite functions. The values of $B_{i,t}$ will then represent the probability that an event at t is from the ith component point process.Is this true? is there a simple proof?

$$
\begin{aligned}
&\lambda^{\theta}_{\mathbb{P}|\mathcal{D}} \\
=&\lambda^{\theta}_{\mathcal{D}} \frac{\mathbb{P}^{\theta}(\mathcal{D}_{t\to T}|t \in \mathcal{T}, \mathcal{D}_t)}{\mathbb{P}^{\theta}(\mathcal{D}_{t\to T}|t \notin \mathcal{T}, \mathcal{D}_t)} \\
=&\lambda^{\theta}_{\mathcal{D}} \exp\left(\int_t^T \log\left(1 + \frac{\lambda_{\text{from t and hidden descendants}}}{\lambda_{\text{exogenous to t plus observed and descendents}}}\right) d\Lambda - \int_t^T \lambda_{\text{from t and hidden descendants}}
\end{aligned}
$$

Can bound this above by bounding denominator below and numerator above using monotonicity of exponential kernel :)

For metaorders with fixed duration and no descendants this can be evaluated quickly. Unfortunately this makes things nondifferentiable:(But can approximated with a 'bump' function with bounded support More generally, can use bump times exp(polynomial), possibly with a dynamic 'duration' component that varies as a function of history, and possibly with descendants

2.6 Monte Carlo EM

I think this makes it possible to do hidden marks

$$
\begin{aligned}
&\mathbb{E}\left[\int \log\lambda d\Lambda - \int \lambda dt|\mathcal{D}\right] \\
&= \int \left[\int \log\lambda d\Lambda - \int \lambda dt\right] p(\lambda|\mathcal{D}) d\mu(\lambda) \\
&= \int \left[\int \log\lambda d\Lambda - \int \lambda dt\right] \frac{p(\mathcal{D}_{after}|\lambda, \mathcal{D}_{before}) p(\lambda|\mathcal{D}_{before})}{p(\mathcal{D}_{after}|\mathcal{D}_{before})} d\mu(\lambda) \\
&= \frac{\mathbb{E}\left[\left(\int \log\lambda d\Lambda - \int \lambda dt\right) p(\mathcal{D}|\lambda)\right]}{p(\mathcal{D})} \\
&= \frac{\mathbb{E}\left[\left(\int \log\lambda d\Lambda - \int \lambda dt\right) p(\mathcal{D}|\lambda)\right]}{\mathbb{E}\left[p(\mathcal{D}|\lambda)\right]}
\end{aligned}
$$

This is basically particle filtering, I think?

$$
\begin{aligned}
H(s) &= UH(s) + \epsilon_u(s) \\
X(s) &= OH(s) + \epsilon_i(s)
\end{aligned}
$$

$$
\begin{aligned}
\sum_s \mathbb{E}\Bigg[\int &\left(\log(\alpha) + X(s)c + H(s)c_H - \beta(t-s)\right) B_{s,t} d\Lambda(t) \\
&- \int \alpha \exp\left(X(s)c + H(s)c_H - \beta(t-s)\right) dt \Big| \mathcal{D}
\end{aligned}
$$

$$
\mathbb{E}\left(\sum_s \int H(s) B_{s,t} d\Lambda(t) - \int \alpha H(s) \exp\left(X(s)c + H(s)c_H - \beta(t-s)\right) dt \Big| \mathcal{D}\right)
$$

2.7 Hidden Marks

2.7.1 Discrete

Markov

2.8 Quadratic Hawkes Processes

$$\int k \log(\alpha \sum \exp(-\beta(t-s)))d\Lambda - \alpha^k \int \left(\sum \exp(-\beta(t-s))\right)^k dt$$

$$= \int k \log(\alpha) d\Lambda + \int k \log(\sum \exp(-\beta(t-s)))d\Lambda - \alpha^k \int \left(\sum \exp(-\beta(t-s))\right)^k dt$$

$$\frac{\partial}{\partial \alpha}$$

$$= \int \frac{k}{\alpha} d\Lambda - k\alpha^{k-1} \int \left(\sum \exp(-\beta(t-s))\right)^k dt$$

$$\frac{\partial}{\partial \beta}$$

$$= \int k \frac{\sum \exp(-\beta(t-s))(s-t)}{\sum \exp(-\beta(t-s))} d\Lambda - \alpha^k \int k \left(\sum \exp(-\beta(t-s))\right)^{k-1} \left(\sum \exp(-\beta(t-s))(s-\right.$$

$$\frac{\partial}{\partial k}$$

$$= \int \log(\alpha \sum \exp(-\beta(t-s)))d\Lambda - \int \log\left(\alpha \sum \exp(-\beta(t-s))\right) \left(\alpha \sum \exp(-\beta(t-s))\right)^k dt$$

$$\int_a^b \log(ce^{sx}) ce^x dx$$

$$= \log(c) \int_a^b e^x dx + cs \int_a^b x e^x dx$$

$$= \log(c)(e^b - e^a) + cs \left([xe^x]_a^b - \int_a^b e^x dx \right)$$

$$= \log(c)(e^b - e^a) + cs \left(be^b - ae^a - e^b + e^a \right)$$

2.9 Uncertainty Quantification

Asymptotic normality

Parametric Bootstrap

2.10 Computational Concerns

Sensor fusion for parallelisation

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

Exploiting sparsity [14]

2.11 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 3

Generative Sampling

3.1 Simulation Methods

Immigration-birth interpretation [13] (aka Watson-Galton models)

Pseudocode

Ogata thinning

Pseudocode

Do these have different time complexity? Memory complexity?

3.2 Simulation Study of Estimation Methods

Convergence analysis for simple models (eg univariate)

3.3 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?

[19] [20] sources that use hawkes process models to simulate an order book for analysis purposes

Chapter 4

Application to KOSPI/SPY Data

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`$_p apers/w29369/w29369.pdf.Optionspri$

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

4.0.1 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$_fil$
$Process.pdf$

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

4.1 Summary of datasets

- data structure and granularity - which assets are involved

4.2 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

4.3 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

4.4 Self-Excitation and Mutual Excitation

i.e. poisson vs hawkes

4.5 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

4.6 Regression on marks

not sure what this will look like yet. how to justify? which variables are reasonable to use?

Chapter 5

Conclusion

This is the conclusion

Chapter 6

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.

6.1 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.

6.2 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

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

6.2.1 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

6.3 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

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

[20] 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