2022-08-08

Predictive Methods for Hawkes Processes in High Frequency Finance

Sjogren, Myles Parker

UNIVERSITY OF CALGARY

Predictive Methods for Hawkes Processes in High Frequency Finance

by

Myles Parker Sjogren

A THESIS
SUBMITTED TO THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE
DEGREE OF MASTER OF SCIENCE

GRADUATE PROGRAM IN MATHEMATICS AND STATISTICS

CALGARY, ALBERTA
AUG, 2022

Abstract

High frequency financial data is burdened by a near obligatory level of randomness that obfuscates the task of predictive modelling. Given that events in a limit order book can be self-exciting in nature and influenced by many external sources, Hawkes processes are a common choice for modelling limit order book dynamics. Many stochastic models have been proposed to describe empirical order-book dynamics and to build a framework for prediction tasks. In this thesis an evaluation of one such family of models, the General Compound Hawkes Process models, is presented with the goal of forming a basis for practical applications of the model. Also examined is the broader application of Hawkes processes in developing trading signals that consolidate information contained within the order flow of a limit order book. Through this work we show the feasibility of Hawkes process based models by showing that they can be used to create both prognostic short term indicators and systematic long term predictions for the mid-price.

Preface

This thesis is an original work by the author. No part of this thesis has been previously published.

Acknowledgements

My sincerest thanks all to those that have supported me over the course of my graduate studies. In particular, I would like to thank my advisor Dr. Anatoliy Swishchuk for putting me on the path of this research, providing me with his wisdom on the subject and guiding me over these past couple of years. Further thanks go to my committee members Dr. Deniz Sezer and Dr. Jinniao Qiu for their considerations and for their assistance in refining my work. I would also like to thank all my colleagues and fellow graduate students at the University of Calgary who have contributed work on this topic in the past for laying a foundation upon which I can firmly stand. Next I must thank Fin-ML, Mitacs and the government of Alberta not only for providing me scholarships, but providing me with new opportunities to evolve myself. As such I must also thank Carlo Lamargese and Futures First for providing me with the incredible opportunity to develop and practise the application of this research as well as their constant encouragement. A great thanks also to Tim DeLise for his council and expertise in guiding me towards my goals and to new ideas. Lastly, thank you to my family for their everlasting support of my academic pursuits, and to my father, whose dream is now fulfilled.

Table of Contents

Abstract i

Preface ii

Acknowledgements iii

Table of Contents v

List of Figures vii

List of Tables ix

1 Introduction 1

2 Limit Order Books 5
2.1 Background . . . 5
2.2 Data Descriptions and Visualizations . . . 9

3 Hawkes Process Based Models 14
3.1 Hawkes Processes . . . 14
3.2 General Compound Hawkes Processes . . . 21
3.2.1 Diffusive Limit Theorems for the GCHP . . . 23
3.2.2 GCHP Model Justification . . . 25
3.2.3 GCHP Model Fitting . . . 27

4 Predictive methods for General Compound Hawkes Processes 33
4.1 Experimental Procedures . . . 33
4.2 Prediction using Diffusive Limits . . . 35
4.3 Prediction using Repeated Simulation . . . 42
4.4 Comparison of Prediction Methods . . . 46
4.5 Cross Validation results for GCHP Prediction Methods . . . 46

5 Hawkes Intensity Based Signals 50
5.1 Background . . . 50
5.2 Classifying Order-book Activity . . . 51
5.3 Hawkes Intensity signal . . . 54
5.4 Results . . . 61

6 Conclusions and Future Work 64

A Buy and Sell Event Classification Algorithms 66

B GCHP Model Fitting Results for Stock Data 69

Bibliography **71**

List of Figures

2.1 10 price levels and 10 volume levels of the LOB for sugar futures taken at a select moment in time. . . 6
2.2 Discrete grid of 5 price levels of sugar order-book data over a 15 minute time window highlighting rapid mid-price oscillations and the simultaneous shifting of price levels. 7
2.3 Example of sugar order-book highlighted with corresponding time and sales data. Buys are highlight in green, sells in red and intermittent trades in yellow. 8
2.4 Example of Stock limit order book data including increasing levels of price and volume organized by ticker and timestamp. 9
2.5 Histograms for the average number of mid-price changes of AMZN stock in disjoint time intervals of ten seconds, one minute and one day respectively from left to right. 10
2.6 Bar chart for the percentage of mid-price changes of a given size for AMZN, EBAY, FB and MSFT stock over a one month period in November 2014. 11
2.7 Example of Sugar data including increasing levels of price and volume organized by timestamp. . . . 12
2.8 Histograms for the average number of mid-price changes of sugar futures in disjoint time intervals of ten seconds, one minute and one day respectively from left to right. 12
2.9 Bar chart for the percentage of mid-price changes of a given size for sugar futures over a six month period starting in January 2021. 13

3.1 Example of uni-variate Hawkes process fit to series of mid-price changes. 16
3.2 Example of simulated uni-variate Hawkes process with exponential kernel using algorithm 1 with parameters $\alpha = 0.6$, $\beta = 0.8$ and $\lambda = 1.2$. Rejected points are highlight in red and accepted points are highlighted in green. 20
3.3 Illustration of quantile based approach for one hour of AMZN stock data. 23
3.4 Clustering effect for the number of mid-price changes on four consecutive days of sugar futures data. Note that large price swings are associated with more densely compacted clusters of price changes. . . 26
3.5 Empirical vs. theoretical cumulative distribution functions for the inter-arrivals times of mid-price changes over four days of sugar data. 26
3.6 Empirical and theoretical auto-correlation for increasing time intervals for a subset from one day of sugar data. 28
3.7 Best fit : GCHP2SDO 32
3.8 SB-Best fit : GCHPnSDO 32
3.9 AMZN Best fit : GCHP2SDO 32
3.10 AMZN Best fit : GCHPnSDO 32

4.1 Illustration of walk forward testing scheme with a three hour to two hour train to test split and an one hour step size. 34
4.2 Confusion matrices for three class problem on sugar data using diffusive limits. 37
4.3 Confusion matrices for two class problem on sugar data using diffusive limits. 38
4.4 Histograms of error between predicted and true mid-price for diffusive limit methods for sugar futures data. 38
4.5 Classification reports and confusion matrices for three class problem on AMZN stock data using diffusive limits. 39
4.6 Classification reports and confusion matrices for two class problem on AMZN stock data using diffusive limits. 40

4.7 Histograms of error between predicted and true mid-price for diffusive limit methods for AMZN stock data. . . . 40
4.8 Confusion matrices for thee class problem on FB stock data using diffusive limits. . . . 41
4.9 Confusion matrices for two class problem on FB stock data using diffusive limits. . . . 42
4.10 Histograms of error between predicted and true mid-price for diffusive limit methods for FB stock data. 42
4.11 Example of simulated Mid-price paths for SB data. . . . 43
4.12 Example of simulated Mid-price paths for AMZN data. . . . 43
4.13 Classification reports, confusion matrices and histogram of error between predicted and true mid-price for repeated simulation methods for SB futures data. . . . 44
4.14 Classification reports, confusion matrices and histogram of error between predicted and true mid-price for repeated simulation methods for AMZN stock data. . . . 45
4.15 Classification reports, confusion matrices and histogram of error between predicted and true mid-price for repeated simulation methods for FB stock data. . . . 45
4.16 Confusion matrix for SB cross validation predictions. . . . 47
4.17 Histogram of error for SB cross validation predictions. . . . 47
4.18 Example downwards trading day for GCHP strategy. . . . 48
4.19 Example upwards trading day for GCHP strategy. . . . 49

5.1 Clustering of level one order queue buy and sell events for one hour of SB data. . . . 52
5.2 Empirical and theoretical cdf's for the inter-arrival times of level one order queue buy and sell events for two days of SB data. . . . 53
5.3 Three progressions of net buy and sell order flow based on observed changes in ten fixed price levels of the sugar order-book. . . . 54
5.4 Comparison of Net order flow to Mid-price for four different days of sugar data. . . . 55
5.5 Mid-price, net order flow and uni-variate Hawkes conditional buy and sell intensity functions created using ten fixed price levels for one hour of sugar data. . . . 55
5.6 Exponentially weighted moving averages of buy and sell intensity created using order flow on ten fixed price levels for one hour of sugar data. . . . 56
5.7 Hawkes intensity based directional indicator created using ten fixed price levels for one hour of sugar data. Upwards indicators are highlighted in blue, neutral in yellow and downwards in red. . . . 57
5.8 Example trading day for Hawkes intensity trader based on level one order queues. . . . 58

List of Tables

3.1 Estimated Hawkes parameters for two days of sugar data alongside a goodness of fit test comparing the empirical to observed expected number of events in a unit one second interval. 30
3.2 Computed quantile bins, state-values and stationary probabilities for two test days of sugar data. . . . 31
3.3 Summary statistics for GCHP model fitting process over all sugar data. 32

4.1 Classification reports for three class predictions using diffusive limits on sugar data 37
4.2 SB PDL 3C-CR. 37
4.3 SB EPDL 3C-CR. 37
4.4 SB JDL 3C-CR. 37
4.5 Classification reports for two class predictions using diffusive limits on sugar data 37
4.6 SB PDL 2C-CR. 37
4.7 SB EPDL 2C-CR. 37
4.8 SB JDL 2C-CR. 37
4.9 Classification reports for three class predictions using diffusive limits on AMZN stock data 39
4.10 AMZN PDL 3C-CR. 39
4.11 AMZN EPDL 3C-CR. 39
4.12 AMZN JDL 3C-CR. 39
4.13 Classification reports for two class predictions using diffusive limits on AMZN stock data 40
4.14 AMZN PDL 2C-CR. 40
4.15 AMZN EPDL 2C-CR. 40
4.16 AMZN JDL 2C-CR. 40
4.17 Classification reports for three class predictions using diffusive limits on FB stock data 41
4.18 FB PDL 3C-CR. 41
4.19 FB EPDL 3C-CR. 41
4.20 FB JDL 3C-CR. 41
4.21 Classification reports for two class predictions using diffusive limits on FB stock data 41
4.22 FB PDL 2C-CR. 41
4.23 FB EPDL 2C-CR. 41
4.24 FB JDL 2C-CR. 41
4.25 Summary statistics for mid-price prediction classification problems for all methods. 46
4.26 GCHP in sample classification results. 47
4.27 GCHP out of sample classification results. 47
4.28 GCHP out of sample CV Classification report for subset of sugar data. 48
4.29 Summary trading statistics for GCHP strategy on forty day subset of sugar data. 49

5.1 Cross validation results for Hawkes Process level one order queue trader 61
5.2 Level one algorithm in sample results. 61
5.3 Level one algorithm out of sample results. 61
5.4 Level one algorithm summary statistics. 61
5.5 Cross validation results for Hawkes process fixed ten level order flow trader 62
5.6 Order flow algorithm in sample results. 62
5.7 Order flow algorithm out of sample results. 62
5.8 Order flow algorithm summary statistics. 62

B.1 Summary statistics for GCHP model fitting process over all AMZN data. 70
B.2 Summary statistics for GCHP model fitting process over all EBAY data. 70
B.3 Summary statistics for GCHP model fitting process over all FB data. 70
B.4 Summary statistics for GCHP model fitting process over all MSFT data. 70

List of Algorithms

1 Ogata's modified thinning algorithm for simulation of the uni-variate exponential Hawkes process . . 20

2 Hawkes Intensity Trading Algorithm 59

3 Algorithm for classifying and sizing level one order queue buy events 67

4 Algorithm for classifying buy and sell events based on order-flow 68

Chapter 1

Introduction

In recent times world financial exchanges have become highly integrated with our society as well as increasingly accessible to any and all interested parties. A core component of these exchanges is the electronic limit order book which has become a fundamental way to understand, interact with and engineer financial markets. Limit order books are in fact the main driver and source of data for high-frequency traders [32]. As such, they have become a point of interest for research in machine learning [50] and stochastic modelling [45]. Prominent topics within the latter area include predicting the timing and size of future market events as well as the feasibility of modelling their arrivals using stochastic processes. A pervasive practise in past mathematical literature written on stochastic modelling of limit order books is the use of Poisson processes, for instance in [13], independent Poisson processes were used to model the occurrences of market orders, limit orders and cancellations within an order-book. However, this practise does not align with certain established empirical facts about financial data, such as the clustering of event arrivals as documented in [12, 45]. Further, within the context of an order-book events can be defined as price changes, new orders or various other updates that are either responsible for, or resultant of evolving market conditions. A modification of these approaches is to replace any instance of a Poisson process with a Hawkes process which is characterized by a self-exciting property. The Hawkes process has been shown to embody financial data more aptly, but the potential of Hawkes process based models in handling prediction tasks is less documented.

Working within the framework of a limit order book requires some general adaptations from traditional stochastic modeling of asset prices. The main divergence is that in limit order books, any price is viewed as a discrete (not continuous) process. We cannot assume that the price gradually changes, but instead it jumps from one value to another on a predetermined discrete grid. In fact, at any given time looking at a limit order book shows that there is no single price of an asset, but instead there is a collection of outstanding, unfilled orders that reflect supply and demand. So for the purpose of modelling the value of any asset, the *mid-price* is often used as a proxy [32]. In this way there is a

fundamental difference to understanding a price process via limit order books as opposed to more traditional methods, such as Itô processes [7, 39].

Existing models [39] of price processes as continuous processes begin to disconnect with reality at the micro perspective because there is no real "price" within a limit order book. Instead there is only a history of past transactions and placed orders. Further, these transactions take place at discrete price levels, so their time series evolution exhibits jumps, the effect of which become more profound at smaller time scales. The end result of this is a vast collection of data based on a hierarchy of constantly changing order queues that reflect market conditions on a micro level. Another complication to this picture is the different types of allowable orders on various exchanges that can be placed including market orders, hidden orders, iceberg orders and various others. Certain academics have advocated for the use of continuous time models to explain this complicated market micro-structure, and subsequently introduced a framework based on point processes [17, 22]. In more recent times however, this study of point processes, and in-particular Hawkes processes, has come into more widespread focus [3, 28].

A recent application of Hawkes processes in finance is the development of a new archetype of stochastic model named the General Compound Hawkes Process (GCHP) [40, 42]. This model was first introduced to model risk processes in insurance, but has been adapted for other problems such as modelling the mid-price process of a limit order book. This model consists of two main components; firstly a Hawkes process, which is used to model the number and timing of mid-price changes and secondly a Markov chain, which is used to predict the size and direction of said changes. Simple variations of this model restrict potential mid-price movements to be fixed up and down values while more complex variations define different states for the Markov chain using a quantile based approach on the collection of all observed mid-price movements. Asymptotic analysis of these models has demonstrated a link between order-flow and volatility which can be used to assess the fit of the model [45]. Further the volatility of the mid-price process can be expressed in terms of parameters describing the rate of arrivals of mid-price changes, the direction of these changes and the size of these changes.

In this dissertation we look to study the practical application of the General Compound Hawkes Process to new types of financial data and analyze its predictive capabilities. To this end we apply the model to historical futures data as a means to extend existing theory to a previously untested area of finance. In doing this we can confirm certain universal empirical peculiarities of financial data such as the clustering of events arrivals, and the existence of different volatility regimes over different time frames. Previous literature including [41] has suggested methods defined through diffusive limit theorems for predicting directional change in the mid-price, but stopped short of implementing testing procedures for said methods. As such, thorough testing of this theory is conducted to provide an analysis of the GCHP model as a whole. In doing this, the premise that at any given point in time, we can only use past data to determine the model parameters must be maintained. Only then can a calibrated model be used to evaluate the following section of time and in such a way, the presented procedures fit into the family of *sequential investment strategies*. These are

sequential algorithms that simulate real-life, where we can condition the probability of future events on a history of events leading up to the current time. This methodology is also reminiscent of the application of modern machine learning algorithms, which look to predict the future based on historical data.

After various tests it was generally concluded that the GCHP model was able to both predict the direction and volatility of the mid-price to a reasonable level of accuracy given suitable conditions and optimal parameters. However, maintaining a strong directional prediction of the mid-price across long time spans was found to be challenging for the model. The result of any given test was shown to be heavily dependent on the experimental set-up applied and the parameters used. Nevertheless the model showed strength in predicting the volatility in the mid-price process of several different assets and with sufficient fine tuning, showed potential in predicting directional movements. Despite this one conceptual limitation of the GCHP models is that they solely focus on the mid-price process and because of this, potentially useful information contained within the rest of the order-book can be overlooked. In particular, given that the prompt for a mid-price change is the depletion of either of the level one order queues, it suggests that fluctuations in said order queues can be used as a predictor for price changes. In fact, the net effect of all order queue changes can be shown to be highly correlated with the mid-price itself. This presents another application of Hawkes processes to high frequency finance in modelling the changes in order queues as a means to develop a preemptive signal for mid-price change.

In this dissertation we also look to develop and test real-time adaptable trading strategies based upon the demonstrated value of Hawkes processes in modelling limit order book dynamics. This was done as part of a project to build automated systems to receive, process, store and analyze high-frequency data in a time efficient manner as well as communicate trading signals across different channels. All strategies applied were constructed in a way that was cognizant of time continuity, transaction costs including crossing the spread and computation times. We show that by modelling the changes within an increasing number of levels of bid and ask queues, one can devise a strategy that is predictive enough as well as fast enough to show positive returns over extended periods of time. In this way we show that Hawkes processes can be used to predict more than just the next order-book event and that they are an appropriate tool to help encompass collective market activity in an intra-day time frame.

The research presented in this dissertation looks to investigate the mid-price predication problem using GCHP models, demonstrate the versatility of Hawkes processes in high-frequency finance and show the potential of simple applied trading strategies. The rest of this thesis is organized as follows:

- Chapter 2: Background information on limit order books and data descriptions/visualizations.
- Chapter 3: Background information on Hawkes processes and General Compound Hawkes Processes.
- Chapter 4: An analysis of predictive methods for the General Compound Hawkes Process.

- Chapter 5: An overview of developed order-flow based Hawkes process trading signals.
- Chapter 6: Conclusions and future work.

Chapter 2

Limit Order Books

2.1 Background

Before the advent of the limit order book trade was primarily facilitated by a group of centralized market makers who would publish prices to buy and sell assets in a way that would guarantee they would gain the spread [19]. This was the reward for providing liquidity and undertaking exposure to adverse price movement. In more modern times individual traders, aided by the convenience of order books, have gained the right to place orders using their own specifications creating order driven markets in contrast to the previously existing quote-driven markets. This has lead to the emergence of new phenomenon and an ever expanding bank of high-frequency data. This abundance of data can be used to test new theories as the interactions spawned from order driven markets are rapid in nature and subject to complex dynamics [13].

Limit order books are characterized by an expanding range of price *levels* and *queues*. The collection of all quoted prices for a particular asset form a discrete grid where the smallest allowed space separating possible prices is known as the *tick size*. At any given time one can place an order to buy or sell an asset at a desired price and volume in a variety of ways. The most common way is through limit orders which are called *bids* if they are buy orders below the current market price or *asks* if they are sell orders above the current market price. For example, a level 1 ask (best ask) is the lowest price among all outstanding sell orders. The level 2 ask is then the next highest price with sell orders where the minimum distance between the level 1 and level 2 ask is one tick. These values can be used to define several basic quantities which summarize the current state of a limit order book, a visual illustration of which is shown in figure 2.1.

Definition 2.1. The **ask price** $a(t)$ is lowest price available to buy at time t, the **bid price** $b(t)$ is the highest price available to sell at time t and the **mid price** $m(t) = \frac{1}{2}[a(t) + b(t)]$ is the average of the best bid and ask prices at time t.

Definition 2.2. The **spread** $s(t) = a(t) - b(t)$ is the difference between the best bid and the best ask at time t.

There are many ways of defining features from limit order book data that can be used to track an assets value over time including various manipulations of the best bids, the best asks or simply the last transaction price. Each of these methods have their respective drawbacks in that the best bid and ask only reflect one side of the market and the last transaction price is susceptible to a prohibitive amount of noise due to constant bid-ask bounce [28]. Feature extraction from limit order book data can emphasize different dynamics in differing levels of complexity, examples of this are given in [6, 32]. In this thesis we will instead focus on modelling an assets current and future value through its mid-price because it gives a parsimonious representation of market evolution. In general the mid-price is often used by traders to gauge market direction as well as by market makers to assess stability and manage inventory risk.

An important freedom traders can express with order-books is the freedom to place different kinds of orders at any given time, price and volume. This gives rise to many strategic considerations such as optimal execution time, position sizing and market impact. For instance, if a trader is interested in filling an abnormally large order it is often best to place it bit by bit to ensure a good average price is received, and that they do not drive the price against themselves by consuming all available depth at the current best price.

Definition 2.3. An **order** $x = \{p_x, w_x, t_x\}$ is a commitment to buy/sell up to w_x units at a price no greater/less than p_x. An order is called a **market order** if it is immediately matched or a **limit order** if it is not.

Definition 2.4. A **Limit Order Book (LOB)** $\mathcal{L}(t)$ is the set of all active orders in a market at time t.

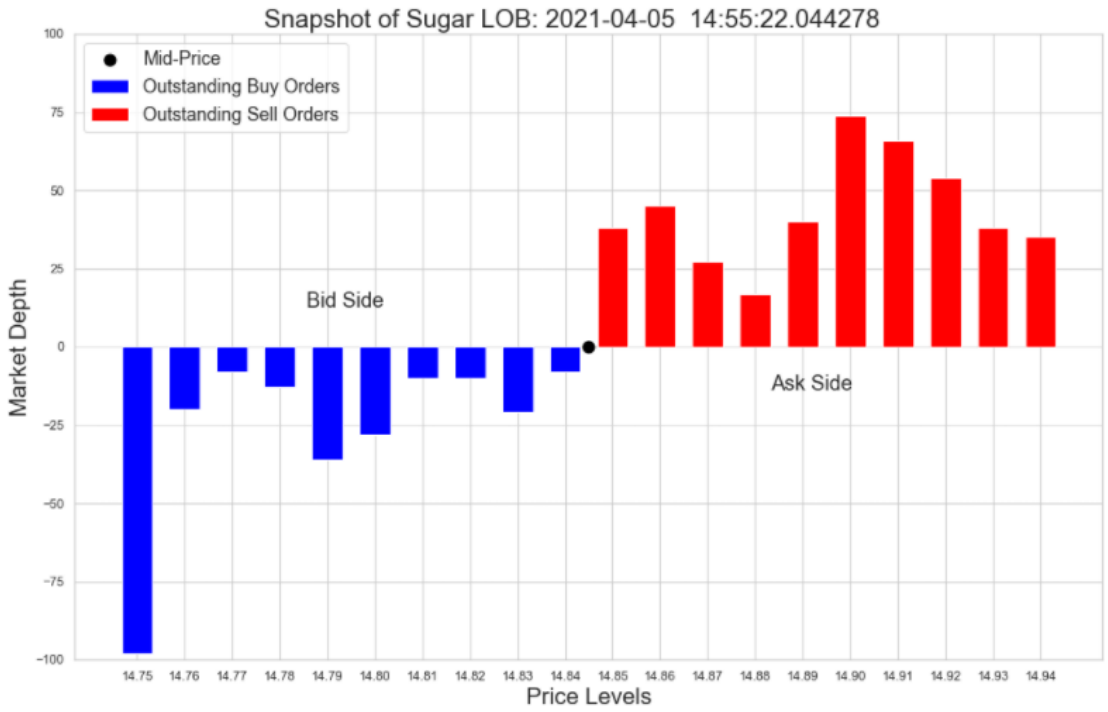


Figure 2.1: 10 price levels and 10 volume levels of the LOB for sugar futures taken at a select moment in time.

Over the course of time new orders settle in the book if they are not matched to an existing opposite order. They will then sit open until either matched, amended or cancelled. Exchanges have set rules to determine how orders get matched typically according to arrival time and size. These rules can be seen to govern how the price of an asset evolves over time as the mid-price will only change if one of the level one queues gets depleted. This serves to support the view that the most relevant information to price evolution is contained towards the centre of the order book and in particular, within the level one order queues. This lays the conceptual foundation for the study of mid-price models as a means to reflect market behaviour.

One prominent challenge to order-book modelling is the size of the state space of an order-book and the high speed at which it is prone to change [19]. The complexity of any discrete grid of allowable prices can lead to run-times for prospective algorithms that can exceed any reasonable time window for the optimal execution of orders in real time. Furthermore, it creates situations where the mid-price has a tendency to oscillate at half tick frequencies over small time intervals. This can bog down the computational effectiveness of mid-price models and produce overly noisy signals that are harder to convert to viable trading strategies. This concept showcases aspects of systemic risk and market instability common to high-frequency data which drives many decision makers to take a probabilistic and stochastic view on potential cash flows [43].

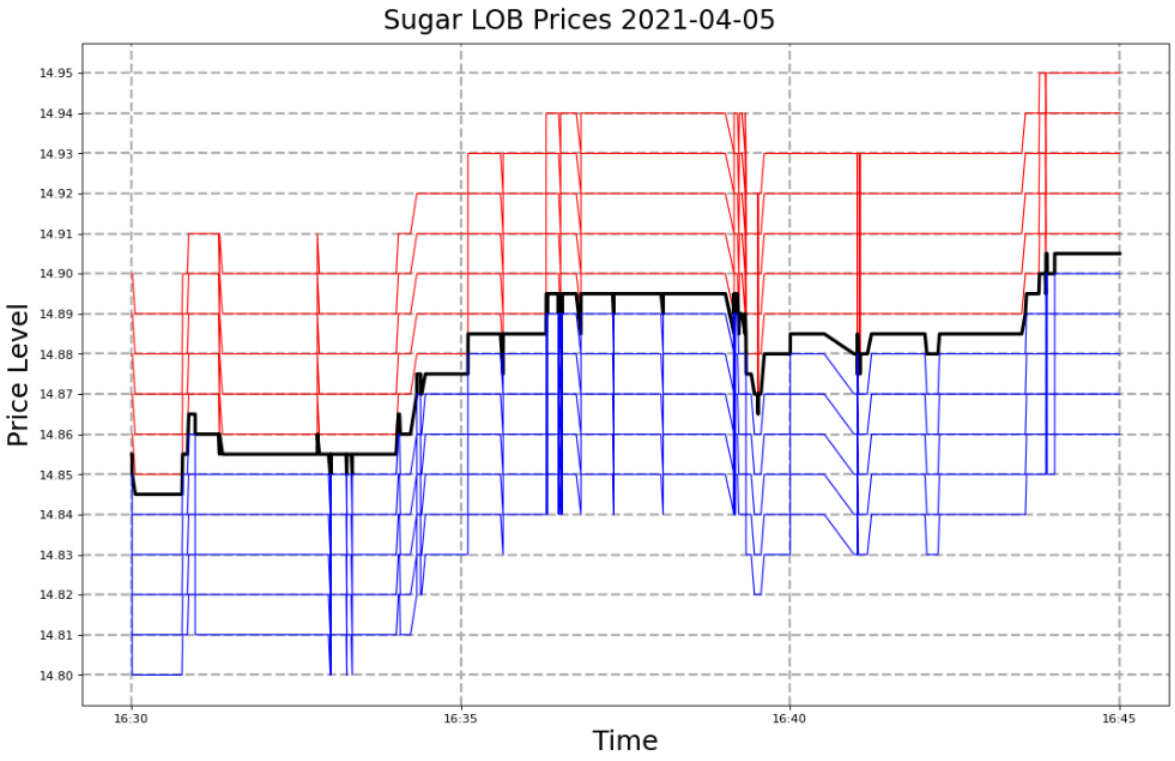


Figure 2.2: Discrete grid of 5 price levels of sugar order-book data over a 15 minute time window highlighting rapid mid-price oscillations and the simultaneous shifting of price levels.

Another challenge to order-book modelling is that the collection of all limit orders, market orders and cancellations comprises the majority of recorded order-book data, but they are by no means a complete representation of a market. One reason for this is that any particular limit order book data-set for a selected asset is only a collection of data provided by a given exchange. Different exchanges can trade the same asset simultaneously and under slightly different conditions, hence any data-set generated from an exchange is a limited picture of the global activity for the underlying asset. Another reason is that the recorded data is indicative of only announced intentions, that is to say traders can hide their intentions through hidden orders or by simply not interacting with the exchange until absolutely necessary, as is the case with all market orders [19]. Furthermore certain traders actions may be directly dependent on $\mathcal{L}(t)$ and $\mathcal{L}(t)$ is dependent on traders actions. This feedback loop can be overtly seen around periods of mid-price oscillation where trades tend to cluster. These effects can be seen in figure 2.3 which shows time and sales data plotted over top of the corresponding order-book data for a short time interval. This overlapping of two separate data-sets creates some ambiguity of the direction of some trades due to slightly misaligned timestamps. In such cases we refer to these as intermittent trades, otherwise we can infer whether a trade was a buy or sell based on whether the fill price was on the ask side or the bid side of the order-book respectively. Figure 2.3 particularly shows that individual trades of higher than average volume can have an immediate, sustained impact on the mid-price, while any impact of trades of smaller than average volume are typically temporary and associated with half-tick oscillations.

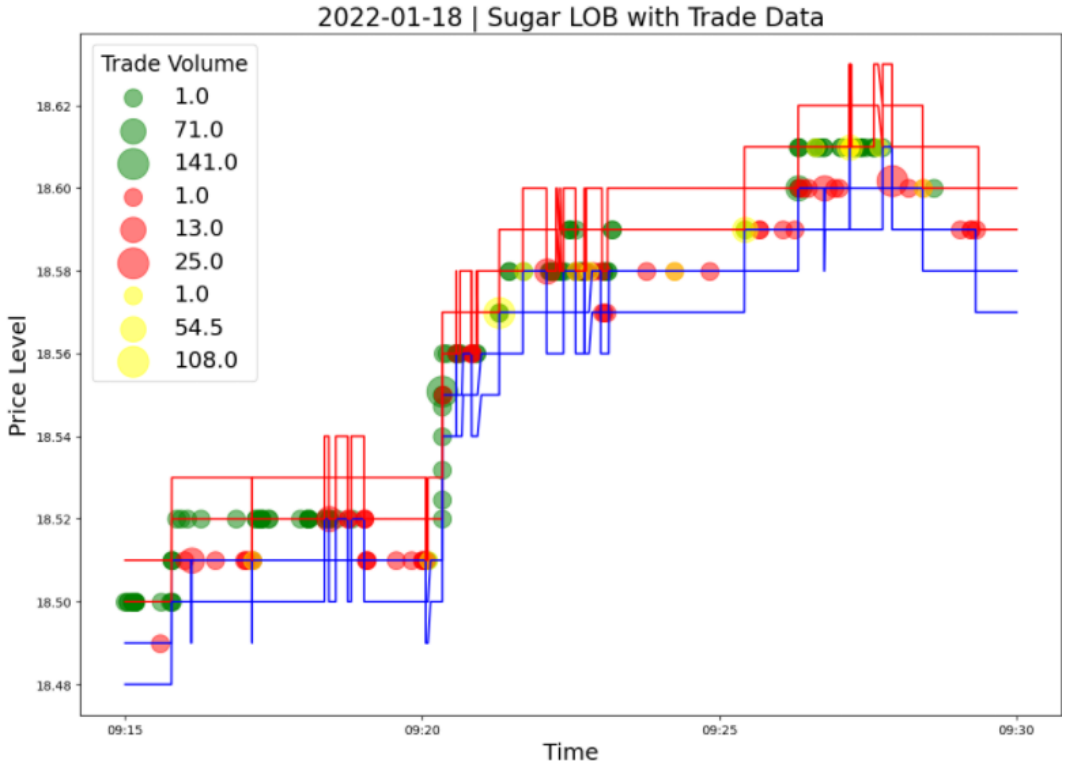


Figure 2.3: Example of sugar order-book highlighted with corresponding time and sales data. Buys are highlight in green, sells in red and intermittent trades in yellow.

2.2 Data Descriptions and Visualizations

Limit order book data-sets from different sources can have different rules for tracking market activity. For instance, some limit order book data-sets contain a different number of levels, some may track only the first five price and volume levels of the order-book while others can track upwards of two hundred. There can also be variation in how different data-sets record order-book updates, some may record a new row every time there is a change in one of the order queues, while others may only record an update after every new market order submission. Further, some data-sets provide categorical details about the type of market events that are recorded while others only show the resulting state of the order-book. In the latter case it is then impossible to determine exactly what type of market-event took place in certain situations. For example, if there is a decrease in the bid queue it is then impossible to tell whether said change was the result of a trade match or a cancellation without the corresponding time and sales data. This gives rise to a conceptual challenge to modelling in that for each distinct case, there is a different implicit connotation towards the state of the market.

Stock Data

The first data-set used in this research is a public historical stock data-set for a variety of stocks including AMZN, FB, MSFT and more as described in and retrieved from [11]. This data contains the top ten levels of price and volume data over a one month period in November 2014. This data is updated after every market order, hence is less abundant than other data-sets while still tracking the intra-day evolution of the order-book.

TimeStamp	Stock	msfm	L1BidPrice	L2BidPrice	L1BidVolume	L2BidVolume	L1AskPrice	L2AskPrice	L1AskVolume	L2AskVolume
2014-11-03 05:14:58.794000+00:00	AMZN	18898794.0	305.01	305.00	100.0	10.0	305.03	305.36	10.0	32.0
2014-11-03 07:08:59.179000+00:00	AMZN	25739179.0	305.10	305.01	100.0	60.0	305.47	310.00	25.0	1.0
2014-11-03 07:09:02.753000+00:00	AMZN	25742753.0	305.01	305.00	60.0	10.0	305.47	310.00	25.0	1.0
2014-11-03 08:33:13.641000+00:00	AMZN	30793641.0	304.03	303.99	40.0	10.0	304.75	305.36	700.0	32.0
2014-11-03 08:33:15.031000+00:00	AMZN	30795031.0	304.03	303.99	40.0	10.0	304.75	305.36	620.0	32.0
...	...	...	...	...	...	...	...	...	...	...
2014-11-26 15:59:59.919000+00:00	VOD	57599919.0	35.64	35.63	2167.0	9400.0	35.65	35.66	7700.0	38200.0
2014-11-26 15:59:59.920000+00:00	VOD	57599920.0	35.64	35.63	2167.0	9400.0	35.65	35.66	7500.0	38200.0
2014-11-26 16:00:42.194000+00:00	VOD	57642194.0	35.56	35.55	100.0	200.0	35.64	35.65	100.0	200.0
2014-11-26 16:00:42.196000+00:00	VOD	57642196.0	35.57	35.56	100.0	100.0	35.65	35.66	100.0	100.0
2014-11-26 16:00:42.196000+00:00	VOD	57642196.0	35.57	35.56	100.0	100.0	35.65	35.66	100.0	100.0

Figure 2.4: Example of Stock limit order book data including increasing levels of price and volume organized by ticker and timestamp.

To demonstrate the high frequency nature of this data we can look at the average number of mid-price changes over set time intervals. Figure 2.5 shows that the mid-price of AMZN stock updates on average at least once every ten seconds, at least ten times per minute and several thousand times per day. The typical size of these changes is shown in figure 2.6 by a bar chart that displays the proportion of all recorded mid-price changes that were of different sizes. Figure 2.6 also shows that different stocks have different overall levels of volatility. In particular, AMZN displays a much broader distribution of absolute mid-price changes as compared to other stocks like MSFT, FB and EBAY where upwards of ninety percent of the recorded mid-price changes were within a single tick in absolute size. This empirical fact can have an impact of the goodness of fit of different proposed mid-price models in that certain models are built around different standards for what is considered a common mid-price change as well as different standards for increased volatility.

Figure 2.5: Histograms for the average number of mid-price changes of AMZN stock in disjoint time intervals of ten seconds, one minute and one day respectively from left to right.

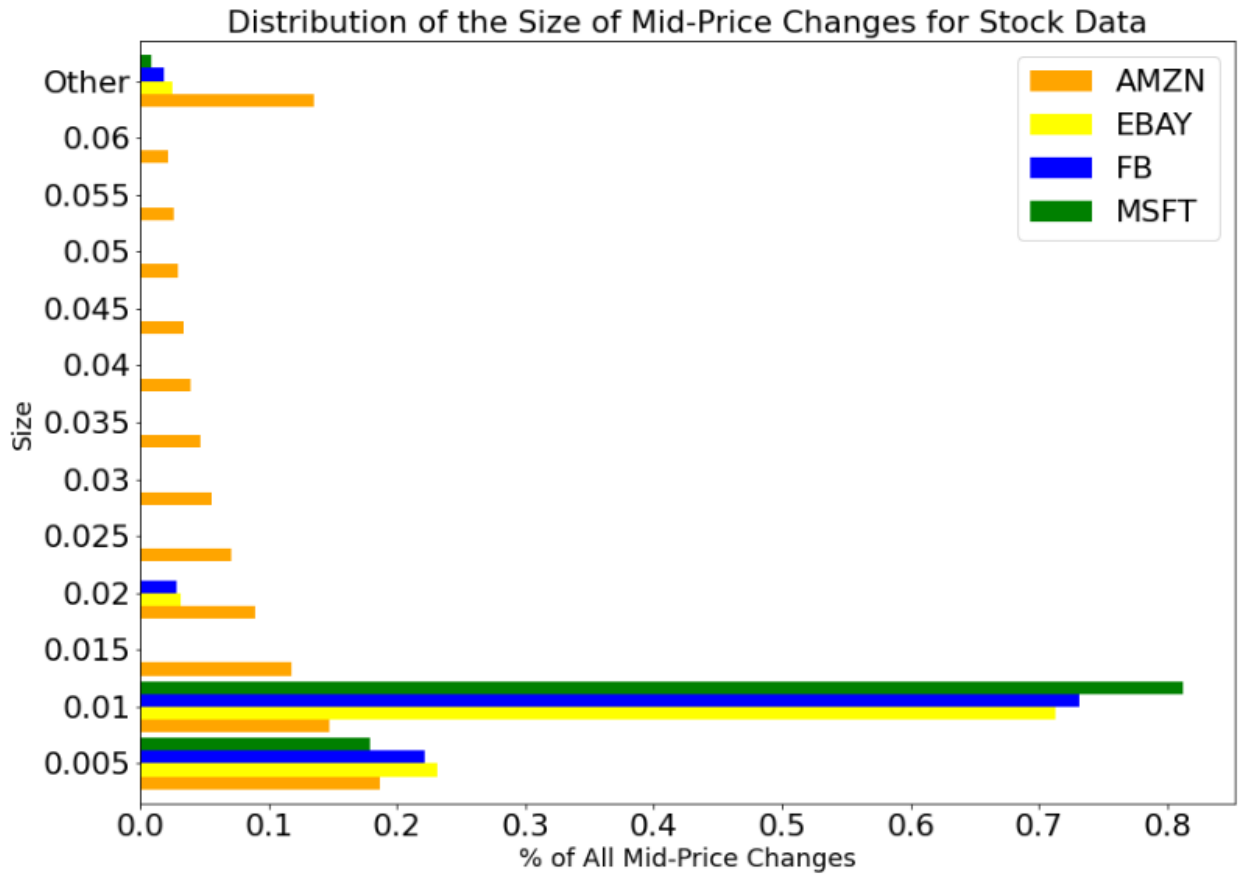


Figure 2.6: Bar chart for the percentage of mid-price changes of a given size for AMZN, EBAY, FB and MSFT stock over a one month period in November 2014.

Futures Data

The largest data-set used in this research is a proprietary sugar futures data-set obtained through Futures First. This data-set contains a row of data corresponding to every single update to the order-book recorded over several different time periods beginning at the start of 2021. That is to say whenever any one of the top ten order queues or price levels changed, a new row of data was recorded detailing the subsequent state of the order-book. Within this data there are many stagnant periods where prices won't update, but the volume available at a given level of depth will. This is a side effect of the ability to cancel currently active limit orders or submit new orders to the limit order book on a free flowing basis. This format gives a finer resolution look at the daily progression of the order-book at the cost of increased computational and storage needs.

Timestamp	Asset	Domain	Type	L1BidPrice	L1BidSize	L1BuyNo	L1AskPrice	L1AskSize	L1SellNo	L2BidPrice	...
2021-04-01 16:53:36.543717+00:00	SBc1	Market Price	Normalized LL2	14.72	109.0	13.0	14.73	100.0	22.0	14.71	...
2021-04-01 16:53:36.656261+00:00	SBc1	Market Price	Normalized LL2	14.72	110.0	13.0	14.73	100.0	22.0	14.71	...
2021-04-01 16:53:36.668180+00:00	SBc1	Market Price	Normalized LL2	14.72	110.0	13.0	14.73	100.0	22.0	14.71	...
2021-04-01 16:53:36.723928+00:00	SBc1	Market Price	Normalized LL2	14.72	110.0	13.0	14.73	100.0	22.0	14.71	...
2021-04-01 16:53:36.783666+00:00	SBc1	Market Price	Normalized LL2	14.72	100.0	13.0	14.73	95.0	22.0	14.71	...
...	...	...	...	...	...	...	...	...	...	...	...
2021-04-28 16:25:58.407554+00:00	SBc1	Market Price	Normalized LL2	17.30	51.0	1.0	17.32	35.0	3.0	17.29	...
2021-04-28 16:25:58.419452+00:00	SBc1	Market Price	Normalized LL2	17.31	5.0	2.0	17.32	7.0	2.0	17.30	...
2021-04-28 16:25:58.435398+00:00	SBc1	Market Price	Normalized LL2	17.31	4.0	1.0	17.32	21.0	2.0	17.30	...
2021-04-28 16:25:58.492194+00:00	SBc1	Market Price	Normalized LL2	17.31	4.0	1.0	17.32	17.0	2.0	17.30	...
2021-04-28 16:25:58.520065+00:00	SBc1	Market Price	Normalized LL2	17.30	54.0	3.0	17.32	19.0	2.0	17.29	...

Figure 2.7: Example of Sugar data including increasing levels of price and volume organized by timestamp.

Figure 2.8 shows that the sugar futures data has a lower overall level of mid-price volatility when compared to the stock data as well as more skewed distribution of the number of mid-price changes. Figure 2.9 then shows that the most common mid-price movement for this sugar data is half a tick with over seventy percent of the recorded movements being of this size, and that over ninety percent of the recorded movements were within one tick in absolute size. This is in contrast to the stock data where a full tick was the most commonly recorded change in the mid-price amongst all the different stocks and more extreme mid-price movements were much more common. These two things together support the premise of using a simpler model for the mid-price of the sugar futures which in turn would serve to lessen computational complexity.

Figure 2.8: Histograms for the average number of mid-price changes of sugar futures in disjoint time intervals of ten seconds, one minute and one day respectively from left to right.

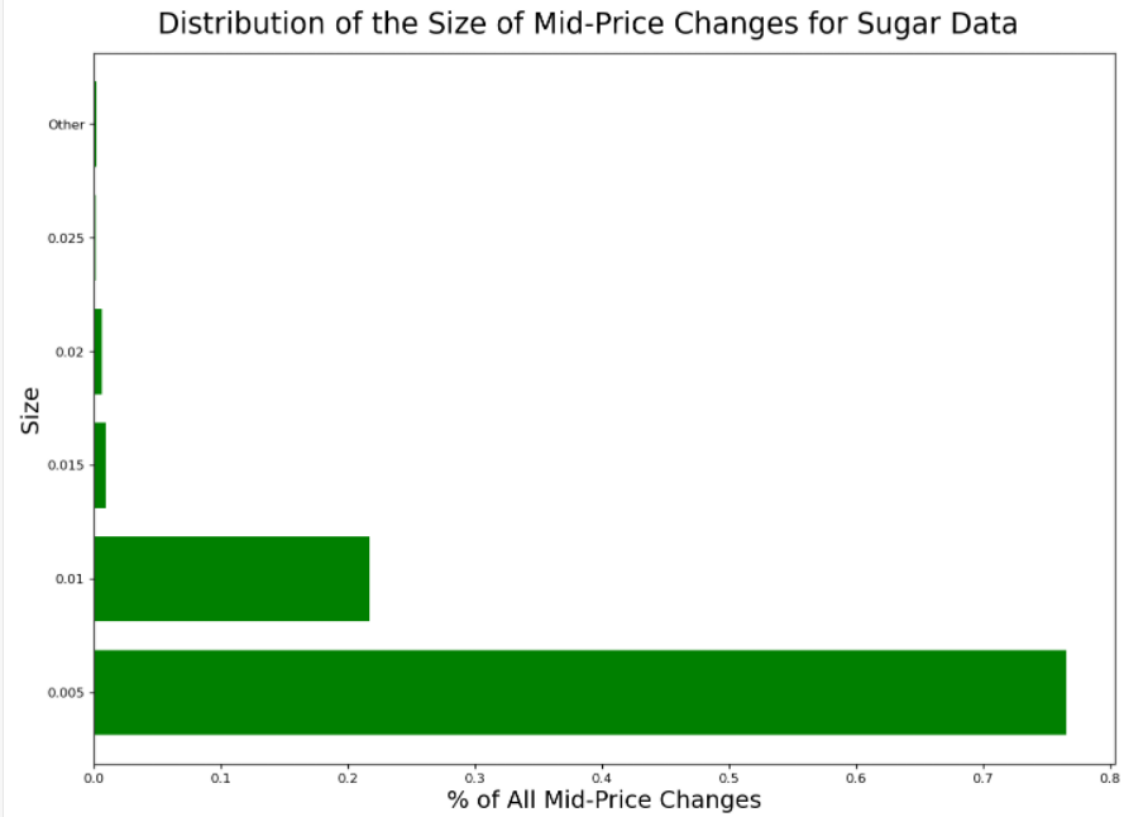


Figure 2.9: Bar chart for the percentage of mid-price changes of a given size for sugar futures over a six month period starting in January 2021.

Chapter 3

Hawkes Process Based Models

3.1 Hawkes Processes

To build up to defining both a Hawkes process and then a General Compound Hawkes Process we start by looking at counting processes which are time-dependent processes characterized by the arrival times of events [16]. In the context of this thesis an event can pertain to a mid-price movement, a change in order flow or an order-book update in general.

Definition 3.1. A stochastic process is a **counting process** if it satisfies the following three properties:

- $N(t) \geq 0, N(0) = 0$
- $N(t+s) \geq N(t) \; \forall \; t, s \geq 0$
- $N(t) \in \mathbb{Z}$

$N(t+s) - N(t)$ is then the number of events occurring during the time interval $[t, t+s]$ and $N(t)$ is a right-continuous step function.

A counting process can be viewed as a cumulative count of the number of event arrivals up to a certain time t. If $s < t$ then $N(t) - N(s)$ is the number of arrivals (counts) on the interval $(s, t]$. A well-known example of a counting process is the Poisson process, derived from the Poisson distribution[26]. Other examples of counting processes would be the arrival of customers at a store or the arrival of orders to a financial market. Counting processes are one class of a more general class of processes called *point processes*, which are a subject of study concerned with the random distribution of points in space and time.

Definition 3.2. Let $T = \{T_1, T_2, T_3, ...\}$ be a sequence of non-negative random variables such that

$$P(0 \leq T_1 \leq T_2 \leq T_3 \leq ...) = 1$$

and the number of points in a bounded region is finite almost surely. Then T is called a **point process** and is characterized by the **conditional intensity function** $\lambda(t)$ in the form

$$\lambda(t) = \lim_{h \to 0} \frac{E[N(t+h) - N(t) \mid \mathcal{F}^N(t)]}{h}$$

where $\lambda(t)$ is a non-negative function and $\mathcal{F}^N(t)$, $t > 0$ is the corresponding natural filtration.

Point processes are a very general formulation that can be entirely defined using their intensity function, where the intensity function can be viewed as the expected infinitesimal rate at which events are expected to occur around time t given the history of events up to that time. In nature we may find specific processes exhibiting behaviour which at first glance may seemingly stem from a random distribution of arrivals, but after closer examination they may fail to be random at all. For example, if we want to understand the probability of an earthquake happening, this process exhibits non-random behavior. Instead such processes can exhibit *self-exciting* behavior causing a temporal clustering of events [3]. In other words one event arrival increases the probability of another happening immediately after. In this way the intensity process is seen to be driven primarily by itself and this suffices to describe a *Hawkes process* [11], named after its creator to model seismic activity[23].

Definition 3.3. 1-D Hawkes Process We say that a counting processes $N(t)$ is a one dimensional Hawkes process if its intensity function has the following form,

$$\lambda(t) = \lambda + \int_0^t \mu(t-s) dN_s$$

where $\lambda > 0$ and $\mu : (0, \infty) \to [0, \infty)$ such that $\int_0^{+\infty} \mu(s) ds < 1$ are called the **background intensity** and the **excitation function** respectively.

Remark 3.4. Given a realized sequence of events $\{t_1, t_2, ..., t_n\}$ with $t_n < t$, these events can form a point process with conditional intensity,

$$\lambda(t) = \lambda + \sum_{t_k < t} \mu(t - t_k)$$

Overwhelmingly existing literature has been concerned with the exponential excitation function, also called the exponential kernel, due to its analytical tractability. This model sets $\mu(t) = \alpha e^{-\beta t}$, and so the *uni-variate linear Hawkes process with exponential decay* is a 3-parameter self-exciting point process. An example of this how this

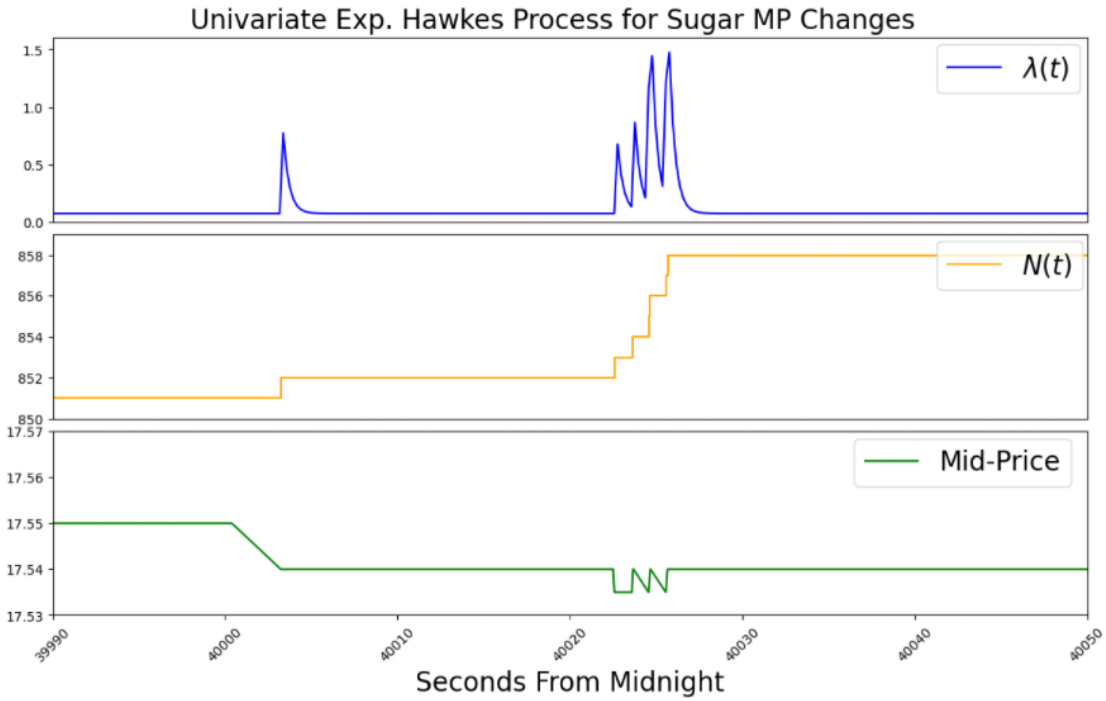


Figure 3.1: Example of uni-variate Hawkes process fit to series of mid-price changes.

process accumulates for a series of events defined as mid-price changes is shown in figure 3.1 which displays the mid-price, conditional intensity function $\lambda(t)$ and Hawkes process $N(t)$ for a minute of sugar data.

To assess the goodness of fit of a Hawkes process to a series of events one can use the fact that conditional intensity function has a long term mean. In particular the Hawkes process has been shown to have desirable asymptotic properties which are demonstrated by the following theorems.

Theorem 3.5. ***LLN for Hawkes Process****[2] Let* $0 < \hat{\mu} := \int_0^\infty \mu(s)ds < 1$*, then it follows that,*

$$\frac{N(t)}{t} \xrightarrow{t\to\infty} \frac{\lambda}{1-\hat{\mu}}$$

In theorem 3.5 the condition that $0 < \hat{\mu} := \int_0^\infty \mu(s)ds < 1$ ensures the process does not explode which itself is a consequence of the stationary of the Hawkes process under certain conditions [28]. In particular for any point process as defined in definition 3.2 it follows that,

$$\lambda(t) = \lim_{h\to 0} \frac{E[N(t+h) - N(t) \mid \mathcal{F}^N(t)]}{h} = \frac{E[dN(t)|\, \mathcal{F}^N(t)]}{dt} \Rightarrow \lambda(t)dt = \mathbb{E}\Big[dN(t)|\, \mathcal{F}^N(t)\Big]$$

then taking expectations of both sides of this equation and applying the tower law gives,

$$\mathbb{E}\left[\lambda(t)\right] dt = \mathbb{E}\left[\mathbb{E}\left[\, dN(t) | \mathcal{F}^N(t)\right]\right] = \mathbb{E}\left[dN(t)\right] = \mathbb{P}\left(t_i \in [t, t+dt)\right) > 0$$

and subsequently $\mathbb{E}[\lambda(t)] > 0$ as $dt > 0$. Now applying the definition of the Hawkes conditional intensity function $\lambda(t)$ and assuming it is stationary,

$$\begin{aligned}\mathbb{E}\left[\lambda(t)\right] &= \mathbb{E}\left[\lambda + \int_0^t \mu(t-s) dN(s)\right] = \lambda + \int_0^t \mu(t-s)\mathbb{E}[dN(s)] \\ &= \lambda + \int_0^t \mu(t-s)\mathbb{E}[\lambda(s)ds] = \lambda + \mathbb{E}[\lambda(t)] \int_0^t \mu(t-s)ds\end{aligned}$$

rearranging and preforming a substitution of variables then gives,

$$\mathbb{E}[\lambda(t)] = \frac{\lambda}{1 - \int_0^t \mu(t-s)ds} = \frac{\lambda}{1 - \int_0^\infty \mu(y)dy} > 0$$

and consequently

$$\int_0^\infty \mu(y)dy > 0$$

given $\lambda > 0$. This quantity $\hat{\mu} = \int_0^\infty \mu(s)ds$ is known as the branching ratio and in the case of the exponential excitation function it can be computed as,

$$\hat{\mu} = \int_0^\infty \alpha e^{-\beta s} ds = \frac{\alpha}{\beta}$$

and this value can be used as a key conceptual aspect of different representations of a Hawkes process in general [28]. Further, the goodness of fit of a Hawkes process can be tested given that,

$$\mathbb{E}\left[N[0,1]\right] = \mathbb{E}[\lambda(t)] = \frac{\lambda}{1-\hat{\mu}} \tag{3.1}$$

where the equality here was documented in [10].

Working within this framework it has been shown that under certain conditions a functional central limit theorem for the linear one dimensional Hawkes process holds true. This is done in [2] as a special case of a more broader theorem for the multivariate Hawkes process where the process converges in the weak sense on the space of cádlág functions on $[0, 1]$ equipped with the Skorokhod topology.

Theorem 3.6. ***FCLT for Hawkes Process****[2] Let* $\hat{\mu} := \int_0^\infty s\mu(s)ds < \infty$*, then it follows that,*

$$\mathbb{P}\left(\frac{N(t) - \frac{\lambda t}{1-\hat{\mu}}}{\sqrt{\frac{\lambda t}{(1-\hat{\mu})^3}}} \leq y\right) \xrightarrow{t\to\infty} \Phi(y)$$

where $\Phi(y)$ *is a standard normal cdf.*

The uni-variate exponential Hawkes process has limits that can be computed with relative ease and that can be used as a basis for finding the limits of modified Hawkes processes. This also gives a convenient way to check the goodness of fit of estimated Hawkes process parameters fitted to any given series of events. Computing such estimated parameters for a Hawkes process can be done using methods such as the expectation maximization algorithm or particle swarm optimization to maximize the likelihood of a given set of data.

Generally speaking in all situations the likelihood can be viewed as the joint density of the observed data to be analyzed and its common use stems from the fact that likelihood methods are known to be asymptotically optimal provided the assumed model is correct [9].

Definition 3.7. Likelihood Let X be a random variable with continuous distribution, x be an outcome of X and θ be the set of parameters of the distribution. Then the likelihood is given by,

$$L(\theta|x) = f_X^\theta(x)$$

In [28] it is noted that the cumulative distribution function of a point process can be characterized by the time of the next arrival conditional to the past. This can be used to define joint density and subsequently the likelihood as follows.

$$f(t_1, t_2, ...t_k) = \prod_{i=l}^{k} f^*(t_i \mid \mathcal{F}^N(t_i))$$

Further given the work done in [37] it was found that for a realization $\omega = \{t_1, ..., t_k\}$ of regular point process the joint density can be written as

$$f_{t_1,...,t_k}(t_1, ..., t_k; T_{k+1} > T) = \prod_{i=1}^{k} \lambda(t_i \mid \omega) \exp\left(-\int_0^T \lambda(t \mid \omega)dt\right)$$

where intuitively a regular process refers to an orderly process in which the probability of having two events happen over an increasingly small interval is tiny. So given such a realization of the Hakwes process the likelihood function becomes deterministic and is defined by the following theorem.

Theorem 3.8. ***Hawkes Process Likelihood****[16, 28] Let $N(\cdot)$ be a regular point processes on $[0,T]$ for some finite positive T and let $t_1, t_2, ..., t_k$ denote a realization of $N(\cdot)$ on $[0,T]$. Then the likelihood L of $N(\cdot)$ given a Hawkes process with intensity function $\lambda(t)$ is expressed in the following form.*

$$L = \left[\prod_{i=1}^{k} \lambda(t_i)\right] \exp\left(-\int_0^T \lambda(u)du\right)$$

Here we should note that theorem 3.8 does not depend on a particular excitation function $\mu(t)$ and this detail motivates it's use for fitting more general Hawkes processes than those with exponential decay. In practice the log-likelihood is often preferred which was introduced in [33] as follows.

$$\ell = \sum_{i=1}^{k} \log\left(\lambda(t_i)\right) - \int_0^T \lambda(u)du \tag{3.2}$$

In the specific case of the exponential kernel the log-likelihood for the Hawkes process has been shown in [34] and is the following.

$$\ell = \sum_{i=1}^{k} \log\left[\lambda + \alpha \sum_{j=1}^{i-1} e^{-\beta(t_i - t_j)}\right] - \lambda t_k + \frac{\alpha}{\beta} \sum_{i=1}^{k} \left[e^{-\beta(t_k - t_i)} - 1\right] \tag{3.3}$$

This general form can be simplified by letting $A(i) = \sum_{j=1}^{i-1} e^{-\beta(t_i - t_j)}$ so that,

$$\ell = \sum_{i=1}^{k} \log\left[\lambda + \alpha A(i)\right] - \lambda t_k + \frac{\alpha}{\beta} \sum_{i=1}^{k} \left[e^{-\beta(t_k - t_i)} - 1\right] \tag{3.4}$$

which has the added effect of reducing the computational complexity given $A(i)$ can be written recursively as follows,

$$A(i) = e^{-\beta t_i + \beta t_{i-1}} \sum_{j=1}^{i-1} e^{-\beta t_{i-1} + \beta t_j} = e^{-\beta(t_i - t_{i-1})} \left(1 + \sum_{j=1}^{i-2} e^{-\beta(t_{i-1} - \beta t_j)}\right) = e^{-\beta(t_i - t_{i-1})} \left(1 + A(i-1)\right)$$

A more detailed summary of this process is given in [28].

For our purposes of market prediction, simulating a Hawkes process is a task worth detailing here. There is more than one way to simulate a Hawkes process, but the most commonly documented approach is Ogata's modified thinning algorithm [34] which is itself an adaptation of [29]. This algorithm is based on the fact that given an event in a series occurs at time t_i, the intensity $\lambda(t)$ is deterministic and bounded from above over the interval $(t_i, t_{i+1}]$. Consequently simulating the next event is then the same as simulating the first event of an in-homogeneous Poisson process with constant rate set to the Hawkes intensity at time t_i. As such the core loop of the algorithm starts by setting $\bar{\lambda} = \sup_{t_i < t \leq t_{i+1}} \lambda(t)$, then generating an inter-arrival time from an exponential distribution with parameter $\bar{\lambda}$, setting the next candidate event time as the previous time plus the generated inter-arrival time and accepting or rejecting this new

point based whether or not an associated uniform random variable falls above or below the current intensity at this new time.

Algorithm 1 Ogata's modified thinning algorithm for simulation of the uni-variate exponential Hawkes process

$s \leftarrow 0 \quad n \leftarrow 0 \quad T = \emptyset$

while $s < T$ **do**

$\bar{\lambda} = \lambda(s)$ ▷ $\bar{\lambda}$ is reset with each new candidate point

$u_1 \sim U(0,1)$

$w = -\ln \frac{u_1}{\bar{\lambda}}$ ▷ inter-arrival times generated at decaying rate using $w \sim Exp(\bar{\lambda})$

$s = s + w$

$u_2 \sim U(0,1)$

if $u_2 * \bar{\lambda} < \lambda(s)$ **then** ▷ candidate points are kept with probability $\frac{\lambda(s)}{\bar{\lambda}}$

$n = n + 1$

$t_n = s$

$T = T \bigcup \{t_n\}$

end if

end while

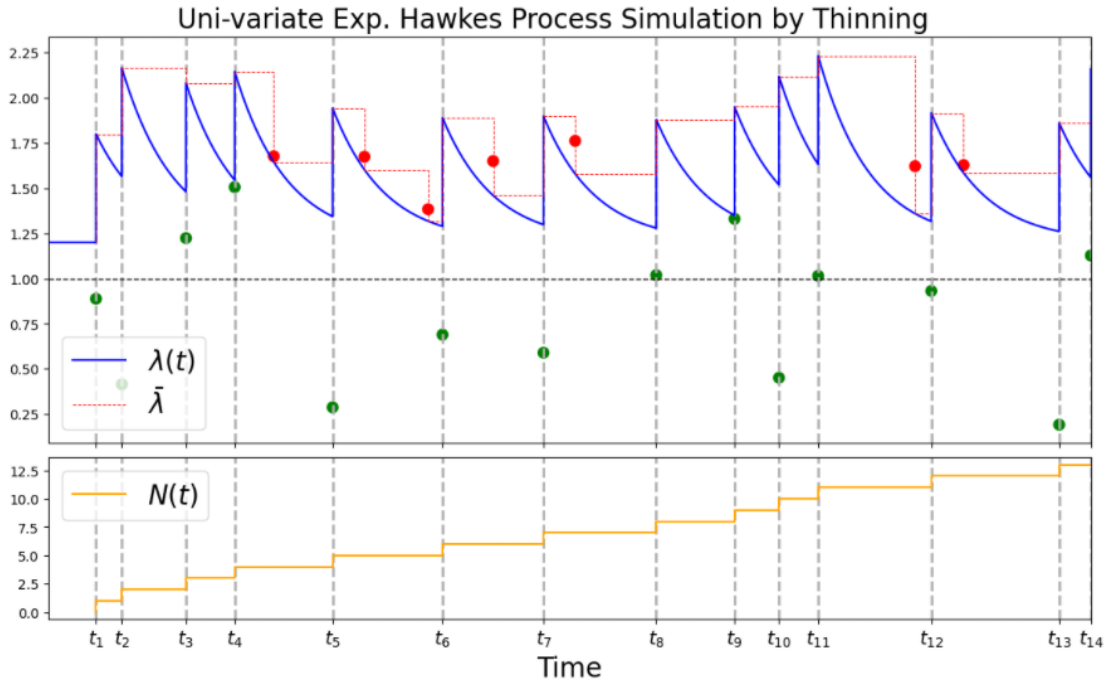


Figure 3.2: Example of simulated uni-variate Hawkes process with exponential kernel using algorithm 1 with parameters $\alpha = 0.6$, $\beta = 0.8$ and $\lambda = 1.2$. Rejected points are highlight in red and accepted points are highlighted in green.

3.2 General Compound Hawkes Processes

For the purpose of our experiments we have implemented four variations of a mid-price model based on the Hawkes process. These models differ only in the number of states used to characterize the size and direction of potential price changes; the basis of the following definitions is explored in [40, 46].

Definition 3.9. General Compound Hawkes Process with Dependent Orders(GCHPDO) Let the mid-price of the underlying be given by,

$$S_t = S_0 + \sum_{k=1}^{N(t)} X_k$$

where $N(t)$ is a Hawkes process and $X_k \in \{+\delta, -\delta\}$ is a Markov chain where δ is fixed to the tick size of the underlying asset. Then we say the process defined by S_t is a General Compound Hawkes Process with Dependent Orders.

Definition 3.10. General Compound Hawkes Process with 2 State Dependent Orders(GCHP2SDO) Let the mid-price of the underlying be given by,

$$S_t = S_0 + \sum_{k=1}^{N(t)} a(X_k)$$

where $N(t)$ is a Hawkes process, X_k is an ergodic continuous time Markov chain independent of $N(t)$ with state space $X = \{1, 2\}$ and $a(.)$ is a continuous and bounded mapping function on X such that,

- $a(1)$ is the mean off all upward price movements
- $a(2)$ is the mean of all downward price movements

Then we say the process defined by S_t is a General Compound Hawkes Process with 2 State Dependent Orders.

Definition 3.11. General Compound Hawkes Process with 4 Dependent Orders(GCHP4DO) Let the mid-price of the underlying be given by,

$$S_t = S_0 + \sum_{k=1}^{N(t)} a(X_k)$$

where $N(t)$ is a Hawkes process, X_k is an ergodic continuous time Markov chain independent of $N(t)$ with state space $X = \{1, 2, 3, 4\}$ and $a(.)$ is a continuous and bounded mapping function on X such that,

- $a(1)$ = mean of all upward price movements $\geq$ one tick in size
- $a(2)$ = mean of all upward price movements $\geq$ one half tick and $<$ one tick in size
- $a(3)$ = mean of all downward price movements $\geq$ one half tick and $<$ one tick in size
- $a(4)$ = mean of all downward price movements $\geq$ one tick in size

then we say the process defined by S_t is a General Compound Hawkes Process with 4 Dependent Orders.

Definition 3.12. General Compound Hawkes Process with n State Dependent Orders(GCHPnSDO) Let the mid-price of the underlying be given by,

$$S_t = S_0 + \sum_{k=1}^{N(t)} a(X_k)$$

where $N(t)$ is a Hawkes process, X_k is an ergodic continuous time Markov chain independent of $N(t)$ with state space $X = \{1, 2, 3, .., n\}$ and $a(.)$ is a continuous and bounded mapping function on X. Then we say the process defined by S_t is a General Compound Hawkes Process with n State Dependent Orders.

These variations on the general compound Hawkes process allow for increasing levels of flexibility in regards to allowable mid-price changes during simulation. This gives an option to adjust for different levels of volatility in the underlying by altering the scope of possible mid-price movements. In particular given an asset known to have low volatility decreasing the number of states and the corresponding size of allowed price changes can often give a better fitted model. Similarly increasing the size of possible movements for more volatile assets will commonly yield a better fitted model overall.

Looking at these definitions together it should also be noted that each of the first three defined models is a specific case of the GCHPnSDO with a select number of states and a specific state mapping function $a(.)$. There any many choices for the function $a(.)$, in the simpler defined models such as the GCHPDO it suffices to use $a(X_k) = X_k$. For higher state models, such as the GCHPnSDO, a quantile based approach is used to break up the collection of all mid-price movements. The approach is outlined as follows:

- Split the data into upwards and downwards price movements
- Compute evenly distributed quantiles on the set of upwards and downwards data disregarding duplicate quantiles
- Compute state values as the mean of all observations which fall between consecutive quantiles
  - Assign state value $a(1)$ to any observation less than the first quantile.
  - Assign state value $a(i)$ to any observation greater than or equal to the i-th quantile and less than the i+1-th quantile.

An illustration of this approach is shown in figure 3.3 in which an hour of amazon stock is processed and the mid-price movements are categorized between a number of mutually exclusive states. Displayed in the figure are the quantiles, counters and associated state values for the thirteen most central states out of a total of fifteen states. This cut-off from fifteen down to thirteen simply was done for the purpose of visual clarity. Also of note in this figure is that although

the majority of movements are collected in more central states, a significant number fall within the outer ranges. This ensures that extreme mid-price movements are separated from average movements which is a feature of the n-state models expressiveness. The number of states chosen for the n-state model is dependent on the data used and the optimal number of states has been shown to vary for different underlying assets. It was noted in [45] that the n-state model has a tendency to outperform others when the number of states is kept relatively small.

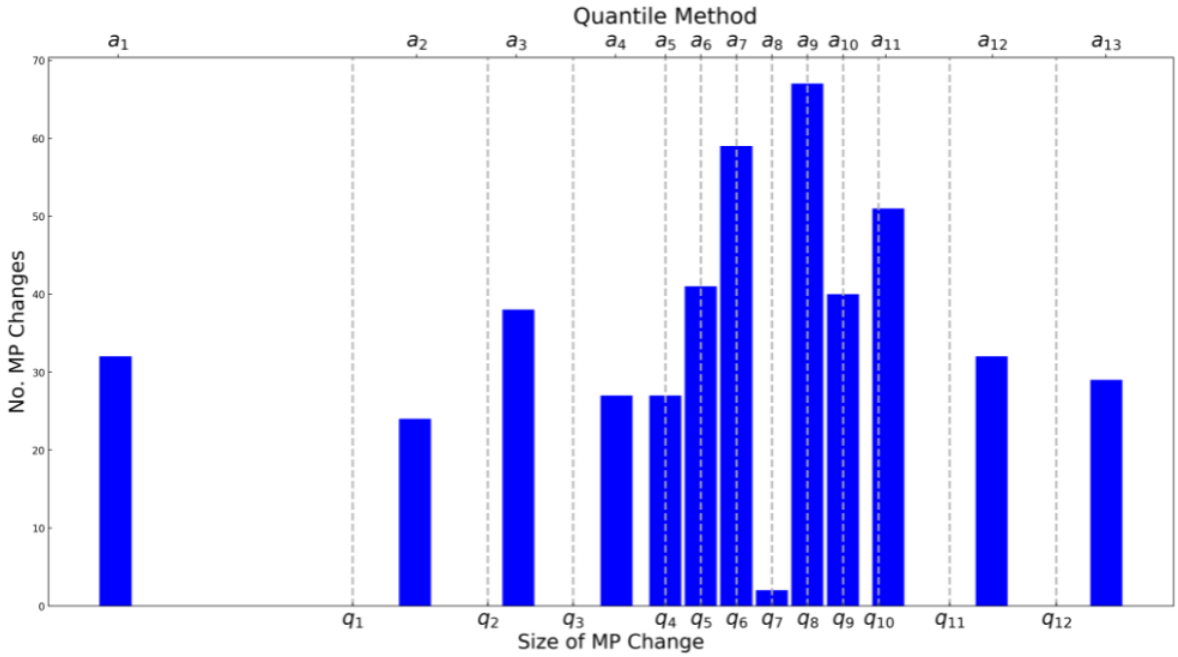


Figure 3.3: Illustration of quantile based approach for one hour of AMZN stock data.

3.2.1 Diffusive Limit Theorems for the GCHP

Asymptotic analysis in the context of high frequency trading can pertain to time scales of just a few minutes or several hours. As such we look to apply the GCHP on a daily time scale to approximate the true mid-price process. In previous works asymptotic analysis was done on the GCHP that uncovered a link between price volatility and order-flow. This link is registered through the diffusive limit of the mid-price process. In particular it can be shown that this limit is expressible in terms of the parameters of the Hawkes processes intensity function and parameters defined through the Markov chain that pertain to mid-price changes. For more details on the following theorems see [45, 41].

Theorem 3.13. *(Jump Diffusion Limit for GCHPnSDO) Let X_k be an ergodic Markov chain with n states $X = \{1, 2, ..., n\}$ with ergodic probabilities $(\pi_1^*, \pi_2^*, ..., \pi_n^*)$ and let S_t be defined as in definition 3.12. Then,*

$$\frac{S(nt) - N(nt)a^*}{\sqrt{n}} \xrightarrow{n\to\infty} \sigma \sqrt{\frac{\lambda}{1-\hat{\mu}}} W(t)$$

where $W(t)$ is a standard Brownian Motion term and the others parameters are defined as follows:

- $0 < \hat{\mu} := \int_0^\infty \mu(s)ds < 1$ *such that* $\int_0^\infty s\mu(s)ds < \infty$
- $(\sigma)^2 := \sum_{i \in X} \pi_i^* v(i)$
- $v(i) = b(i)^2 + \sum_{j \in X} (g(j) - g(i))^2 p(i,j) - 2b(i) \sum_{j \in X} (g(j) - g(i))P(i,j)$
- $a^* := \sum_{i \in X} \pi_i^* a(i)$
- $b(i) := a(i) - a^*$, $b = (b(1), b(2), ..., b(n))'$
- $g := (P + \Pi^* - I)^{-1} b$

Here P is the transition probability matrix for X_k and Π^* is a matrix containing the stationary probabilities of P.

Note that here the transition probability $P(i,j) = P(X_{k+1} = j \mid X_k = i)$ corresponds to the i, j entry of the matrix P and that Π can be approximated using $\Pi \approx \lim_{n \to \infty} P^n$. This theorem suffices for each simpler version of the model but, more convenient expressions can be derived for the parameters in the instance that $n = 2$, see [45] for details. The proof of this theorem stems from the previously stated variations on the LLN and CLT for both the Hawkes process and the GCHP. This theorem gives us a means to test the fit of a GCHP by comparing the volatility of empirical data to an associated theoretical diffusive limit.

Corollary 3.14. *(LLN for GCHPnSDO [40]) The process S_{nt} defined in theorem 3.12 satisfies the following weak convergence in the the Skorohod topology,*

$$\frac{S_{nt}}{n} \xrightarrow{n \to \infty} a^* \frac{\lambda}{1 - \hat{\mu}} t$$

An alternative to theorem 3.12 that uses a pure diffusive limit independent of the number of states exists and can be used to aid in the long term prediction of the price process.

Theorem 3.15. *(Pure Diffusion Limit for GCHPnSDO) Let X_k be an ergodic Markov chain with n states $X = \{1, 2, ..., n\}$ with ergodic probabilities $(\pi_1^*, \pi_2^*, ..., \pi_n^*)$ and let S_t be defined as in definition 4.0.4,* $0 < \hat{\mu} := \int_0^\infty \mu(s)ds < 1$ *and* $\int_0^\infty s\mu(s)ds < \infty$. *Then,*

$$\frac{S(t) - a^* \frac{\lambda}{1-\hat{\mu}} t}{\sqrt{t}} \xrightarrow{t \to \infty} \bar{\sigma} N(0,1)$$

where $N(0,1)$ is a standard normal CDF, $\bar{\sigma}$ is defined as follows,

$$\bar{\sigma} = \sqrt{(\sigma^*)^2 + \left(a^* \sqrt{\frac{\lambda}{(1 - \hat{\mu})^3}} \right)^2}$$

and $\sigma^* = \sigma \sqrt{\frac{\lambda}{(1-\hat{\mu})}}$ where σ and a^* are defined as in the previous theorem.

3.2.2 GCHP Model Justification

The GCHP is a jump model that jumps between the price levels of an order book, this differs from typical jump-diffusion models of asset price in which it is typically assumed that jumps follow a continuous distribution. Further, jumps in a GCHP are not independently and identically distributed but, instead are modeled with a Markov chain which imbibes certain properties like the dependence of consecutive events. What further distinguishes the GCHP from normal jump-diffusion's is the replacement of Poisson processes with Hawkes processes. This seeks to account for some of the major short-comings of jump-diffusion's in that they fail to capture clustering dynamics as well as dependence structures [27]. As such, given we can show the existence of these phenomena in our data it suffices to justify the GCHP model. In particular we use methods first shown in [46] to show that the assumption of uniformity in event arrivals is a bad assumption because in reality, event arrivals within a limit order book are often arbitrarily distributed instead. This plays into the self-exciting property of the Hawkes process which captures the temporal dependence of event arrivals through its stochastic intensity.

Clustering Effect

Due to the self-exciting intensity function of the Hawkes process it has a tendency to generate events in clusters. This is because when one event occurs it creates a burst in intensity which increases the subsequent likelihood of seeing another event immediately afterwords. The end result of this is clustered data and this pattern can be shown to exist in real data by plotting the average number of mid-price changes over one minute bars. This clustering pattern is shown in figure 3.4 for four consecutive days of sugar data.

Arbitrary Distribution of Inter-Arrival Times

In the past a prevailing assumption was that limit order book events arrive according to a Poisson process and as such their inter-arrival times were exponentially distributed. This assumption can be used to define models with high levels of tractability, but in more recent times it has been shown to be a bad assumption given that inter-arrival times more often follow arbitrary distributions. This in turn advocates the use Hawkes processes as their rate of arrivals is heavily influenced by the timing and number of recent events which directly leads to non-standard distributions. To verify this for our data we plot the empirical cumulative distribution function of the inter-arrival times of mid-price changes against theoretical counterparts for various distributions such as the Exponential, Gamma, Weibull, Reciprocal and Wald distributions. All of these distributions are known to have similar shapes to the typical empirical distribution of our data. In figure 3.5 we see that on different days the best fitting distribution can often be Weibull or Gamma, but it was found to never be the case that the best fitting distribution was exponential.

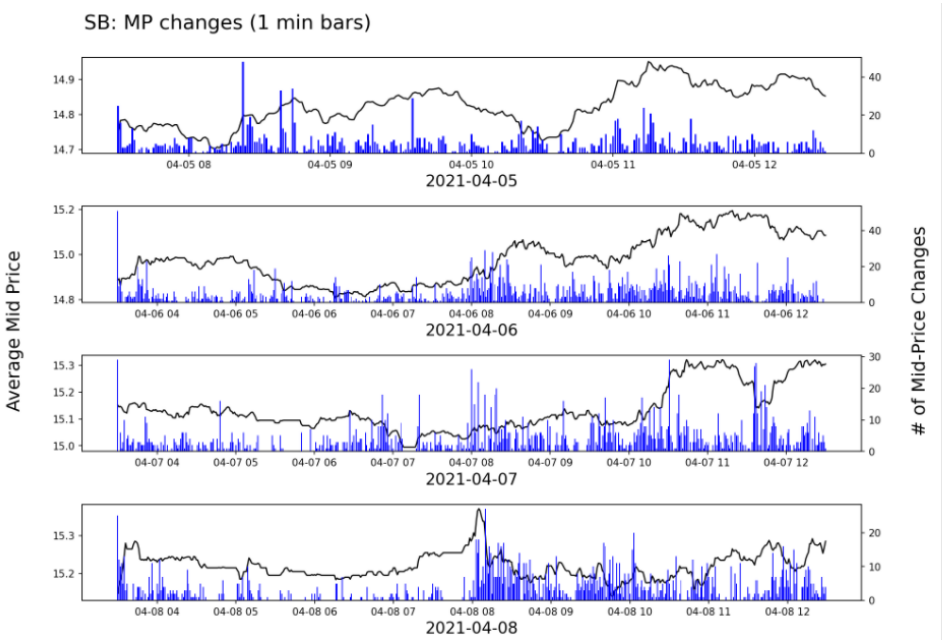


Figure 3.4: Clustering effect for the number of mid-price changes on four consecutive days of sugar futures data. Note that large price swings are associated with more densely compacted clusters of price changes.

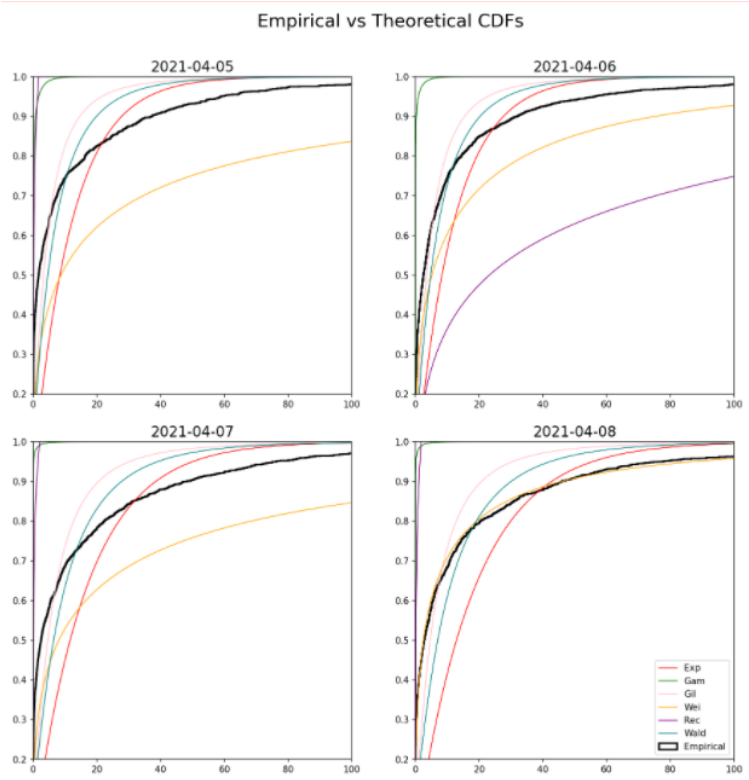


Figure 3.5: Empirical vs. theoretical cumulative distribution functions for the inter-arrivals times of mid-price changes over four days of sugar data.

Dependence of Consecutive Book Events

Another order-book dynamic to consider is how dependent one event is on the previous and to a greater extent, how long does the influence of one book event last? To explain this we can look towards the intensity function of the Hawkes process to see how large every spike in intensity created by an arrival is, and at what rate that created intensity decays. We will later see in table 3.1 that when fitted to a series of mid-price changes the estimated Hawkes parameters for our sugar data give a beta which is much larger than alpha. This infers that decay will often negate the effects of any generated spikes in intensity quite quickly, but this is not always the case. For short time windows we can show the existence of a sophisticated dependence between consecutive events using the auto-correlation of the number of mid-price changes over set time intervals. In particular we can use the following expression suggested in [24] for the empirical auto-correlation,

$$C(\tau,\delta) = \frac{\mathbb{E}\left[(N_{t+\tau} - N_t)\,(N_{t+2\tau+\delta} - N_{t+\tau+\delta})\right] - \mathbb{E}\left[N_{t+\tau} - N_t\right]\,\mathbb{E}\left[N_{t+2\tau+\delta} - N_{t+\tau+\delta}\right]}{\sqrt{var(N_{t+\tau} - N_t)\;var(N_{t+2\tau+\delta} - N_{t+\tau+\delta})}} \tag{3.5}$$

where τ is the length of the time interval considered and δ is the time lag between compared intervals. This function compares the number of mid-price changes over an interval of length τ to increasingly lagged intervals of the same size. This can then be compared to the corresponding theoretical value stated in [14] for a Hawkes process with specific dynamics.

$$C(\tau,\delta) = \frac{e^{-2\beta\tau}(e^{\alpha\tau} - e^{\beta\tau})^2\alpha(\alpha - 2\beta)}{2\left(\alpha(\alpha - 2\beta)(e^{(\alpha-\beta)\tau} - 1) + \beta^2\tau(\alpha - \beta)\right)}e^{(\alpha-\beta)\delta} \tag{3.6}$$

The results of computing equations (3.5) and (3.6) for a subset of sugar data are shown in figure 3.6 where we compute each for increasing values of τ and time lags of up to sixty seconds. From said figure we can see that for all values of τ tested there exists a non-negligible level of positive correlation for the first few time lags. This positive theoretical correlation is seen to subsequently decay, but empirical correlations continue to exist intermittently despite the theoretical value approaching zero. So we can conclude that for this given subset of sugar data the impact of a mid-price change lasts at least few seconds and that one change is not independent from the last. Both of these ideas together reiterate the potential use of Hawkes processes in modelling mid-price changes in an order-book.

3.2.3 GCHP Model Fitting

The best model for any given subset of data may vary based on the volatility intrinsic to said subset. For instance subsets of data taken for time windows closer to market closing tend to be more volatile than those taken from the middle of the day and based on figure 2.6 it is known that different underlying stocks have different distributions for the typical mid-price movement. Consequently given the right circumstances simpler versions of the GCHP model with a limited number of states model may persistently under estimate the true volatility of the data. Choosing the

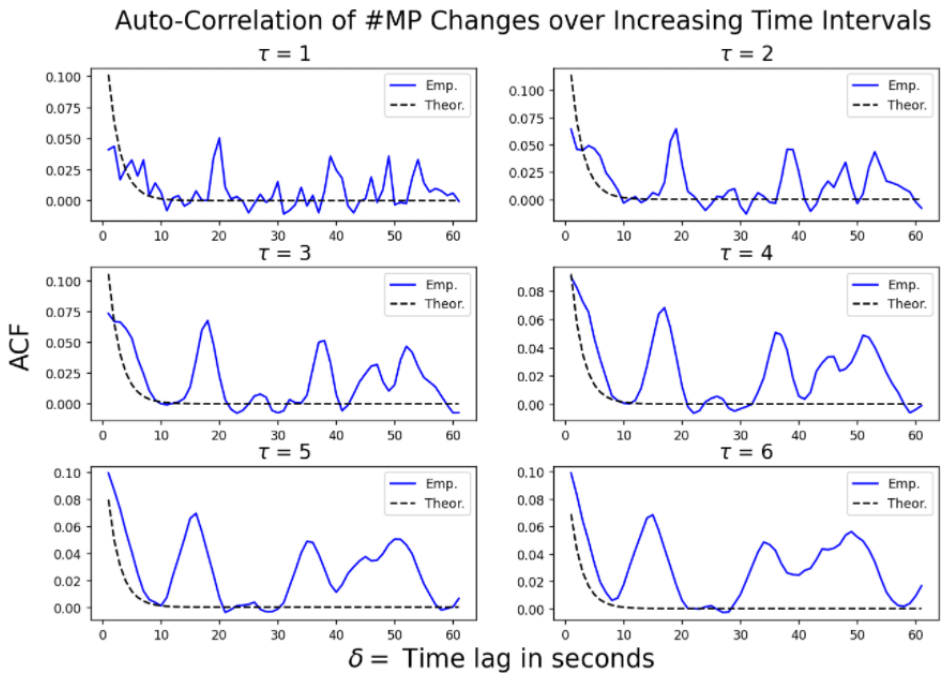


Figure 3.6: Empirical and theoretical auto-correlation for increasing time intervals for a subset from one day of sugar data.

best model for any given subset of data is a process involving fitting each component of the defined model and can be broken down loosely into the following steps:

- Fit a Hawkes Process to the series of arrival times of mid-price changes and record optimal parameters of the arrival process.
- Define transition probabilities between states of the Markov chain as the empirical frequencies of consecutive mid-price movements occurring over the training interval.
- Apply diffusive limit theorems to find theoretical volatility and compare to observed empirical volatility over the training interval.

The details of the first two steps and the origination of this comparative analysis of volatility's can be found in [24, 45]. For the last step recall theorem 3.13 in which the jump diffusive limit for the mid-price process as modelled by the GCHPnSDO was stated.

$$\frac{S_{nt} - N(nt)a^*}{\sqrt{n}} \xrightarrow{\text{n}\to\infty} \sigma \sqrt{\frac{\lambda}{1-\hat{\mu}}} W(t) \tag{3.7}$$

A well fitted GCHP model should be able to approximate the empirical volatility of the mid-price over set time intervals. To provide a means to assess this, one can look at disjoint intervals of increasing size $[in, (i+1)n]$ and by

multiplying each side of the above expression by $\sqrt{n}$ we can define,

$$S_i^* = S_{(i+1)n} - S_{in} - (N((i+1)n) - N(in))a^* \tag{3.8}$$

where $t = 1$. It then follows that for large values of n,

$$std(S_i^*) \approx \sqrt{n}\sigma\sqrt{\frac{\lambda}{(1-\hat{\mu})}} \tag{3.9}$$

where n is the number of steps forward in the discretized process. Given this expression it is possible to plot the empirical standard deviations for increasing n and compare to the corresponding theoretical limit given by the right hand side of equation (3.9) which can be calculated from the fitted parameters of the Hawkes process and Markov chain. To quantify the goodness of fit for each variation of the GCHP model a least squares regression can be preformed to construct a hypothetical best fitting curve which can then be compared to the empirical standard deviations. In particular squaring on both sides the previous expression for the empirical standard deviation gives,

$$std(S_i^*)^2 \approx n\sigma^2\left(\frac{\lambda}{(1-\hat{\mu})}\right) \tag{3.10}$$

where the right hand side can be viewed as a linear regression function with respect to n which has coefficient $c = \sigma^2\left(\frac{\lambda}{(1-\hat{\mu})}\right)$. Thus the following equation (3.11) gives an error rate for the fitting process which can be used to directly compare the fit of different model variations.

$$\left|\frac{\sqrt{c} - \sigma\sqrt{\frac{\lambda}{1-\hat{\mu}}}}{\sqrt{c}}\right| \tag{3.11}$$

Implementing this fitting process over active trading hours for different days of data gives varied results and often the best model can be reflective of different volatility regimes. In particular it was generally found that two state models have a tendency to underestimate the empirical volatility while the more flexible models have a tendency to overestimate it. Also of interest is the spread of the empirical volatility scatter for increasing time windows; the scope of this spread was found to be indicative of the average absolute size of mid-price change observed over a given day.

To give a more detailed account of the model fitting process for the GCHP we look at two example days of sugar futures data. The fitting process starts by fitting a Hawkes process to the series of events as determined by all the observed times of mid-price changes in the recorded data. This is done through maximum likelihood estimation as detailed in section 3.1 and the goodness of fit is tested by comparing the empirical expected number of events in a unit time interval to its theoretical value as was noted in equation 3.1.

Day	α	β	λ	$\mathbb{E}[N([0,1])]$	MLE
01/20/21	2.306	8.823	0.051	0.068293	0.068414
02/02/21	1.990	5.539	0.053	0.082585	0.083073

Table 3.1: Estimated Hawkes parameters for two days of sugar data alongside a goodness of fit test comparing the empirical to observed expected number of events in a unit one second interval.

Comparing the last two columns of data in table 3.1 it's clear that the data is well fit by the Hawkes process with exponential kernel. Now, having determined the estimated Hawkes parameters we then find the parameters of the Markov chain which govern the size of potential mid-price changes. This is done using the empirical frequency of observed mid-price movements over the training data. For the GCHPDO and the GCHP2SDO we can define the transition probabilities by looking at the sign of each set of two consecutive mid-price movements and tallying the results. For our two test days of data the two state transition matrices are as follows,

$$P_{2_1} = \begin{bmatrix} 0.3980 & 0.6020 \\ 0.6650 & 0.3350 \end{bmatrix} \quad P_{2_2} = \begin{bmatrix} 0.3202 & 0.6798 \\ 0.6759 & 0.3241 \end{bmatrix}$$

and we see that is it much more common for consecutive movements to be of opposite direction than the same direction. This in indicative of a significant amount of mean-reversion in the mid-price process and is consistent with the observed behaviour of the mid-price process in section 2. For the GCHPnSDO to define the transition matrix we must first define the quantiles that determine the classification boundaries for the observed events. In practise to avoiding doing an unnecessarily large number of computations we set a default value for the number of quantiles used to divide the data. In the case of duplicate quantiles the resulting number of states n used in the model will often be less than the number of specified quantiles. For the two days of test data shown here the number of non-duplicate quantiles was six, hence the n-state transition matrices had dimensions six by six. However, for more volatile data the optimal value for n could double in size and this optimal value can be found by running iterative tests on an increasing number of quantiles until there is no noticeable improvement in model fit, an example of this process is given [45].

$$P_{n_1} = \begin{bmatrix} 0 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0 & 0.0000 & 0.0769 & 0.2308 & 0.5385 & 0.1538 \\ 0 & 0.0111 & 0.1407 & 0.1037 & 0.4444 & 0.3000 \\ 0 & 0.0075 & 0.1030 & 0.2640 & 0.4963 & 0.1292 \\ 0 & 0.0063 & 0.1502 & 0.4460 & 0.3192 & 0.0782 \\ 0 & 0.0076 & 0.3042 & 0.2890 & 0.1635 & 0.2357 \end{bmatrix} \quad P_{n_2} = \begin{bmatrix} 0 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0 & 0.0000 & 0.1333 & 0.1333 & 0.667 & 0.0667 \\ 0 & 0.0109 & 0.1202 & 0.1749 & 0.4536 & 0.2404 \\ 0 & 0.0107 & 0.0473 & 0.2710 & 0.5787 & 0.0923 \\ 0 & 0.0049 & 0.0930 & 0.5777 & 0.2521 & 0.0722 \\ 0 & 0.0000 & 0.1918 & 0.5023 & 0.1324 & 0.1735 \end{bmatrix}$$

2021-01-20 : Quantiles = $\{q_1, q_2, q_3, q_4, q_5, q_6\} = \{-0.025, -0.01, -0.005, 0.005, 0.01, 0.03\}$							
State	1	2	3	4	5	6	
Bin	$(-\infty, -0.025)$	$[-0.025, -0.01)$	$[-0.01, -0.005)$	$[0.005, 0.005)$	$[0.005, 0.01)$	$[0.01, \infty)$	
$a(i)$	0	-0.0185	-0.01	-0.005	0.005	0.01051	$a^* = 0.00021$
$\pi(i)$	0	0.0076	0.1572	0.3100	0.3715	0.1537	$\sigma = 3.2826e-05$
2021-02-02 : Quantiles = $\{q_1, q_2, q_3, q_4, q_5, q_6\} = \{-0.03, -0.01, -0.005, 0.005, 0.01, 0.025\}$							
State	1	2	3	4	5	6	
Bin	$(-\infty, -0.03)$	$[-0.03, -0.01)$	$[-0.01, -0.005)$	$[0.005, 0.005)$	$[0.005, 0.01)$	$[0.01, \infty)$	
$a(i)$	0	-0.0187	-0.0010	-0.0050	0.0050	0.0106	$a^* = 4.6486e-05$
$\pi(i)$	0	0.0072	0.0876	0.4066	0.3929	0.1058	$\sigma = 2.5320e-05$

Table 3.2: Computed quantile bins, state-values and stationary probabilities for two test days of sugar data.

In these n dimensional matrices is should be noted that the first row and column reflect the transition probabilities associated with the most downward state which is defined to collect all movements strictly less than the smallest computed quantile. So when that smallest quantile corresponds exactly to the smallest observed movement, no data will fall into that particular bin. A breakdown of the values for the bins, state values and stationary probabilities is given in table 3.2 for the n-state model. In that table it can be noted that more often than not the state-values line up directly with computed quantiles and that the density of the stationary probabilities is highest around the bins associated with the smallest possible mid-price movements. The calculated values for a^* are given in the last column alongside the volatility parameter σ; a positive value of a^* indicates that the majority of mid-price movements observed were upward movements and a negative value indicates that the majority of the movements were downwards. This parameter serves as the primary directional indicator for the predictive methods of the GCHP detailed later.

We can then apply the previously detailed process in equations (3.8) - (3.11) to produce fitting curves for each proposed model and determine visually as well as numerically the best fitting model. The two examples days in which the best fitting model is different are displayed in figure 3.7 and figure 3.8. Within these figures the blue scatter points show the empirical standard deviation of the modified mid-price given by equation (3.8), the black curve shows the corresponding theoretical standard deviation given by the right hand side of equation (3.9) and the red curve is the best fitting curve to the empirical scatter. These figures present an example of how a day with lower overall volatility was better fit by a two state model and a day with higher overall volatility was better fit by a n-state model.

Applying this same process across our entire data-set shows that the best fitting model is often dependent on the day. Despite this the best model for the sugar data was consistently the n-state model as it was the best fitted model close to half the time, however, it has a higher overall mean error rate across all days than the other models and a higher worst case fit. Summary statistics for the fitting process over the entire collection of sugar data is shown in table 3.3.

For the collection of stock data there are collectively fewer days to work with than the futures data so the following results are less likely to generalize, but the fitting process still produces some interesting outcomes. For instance the

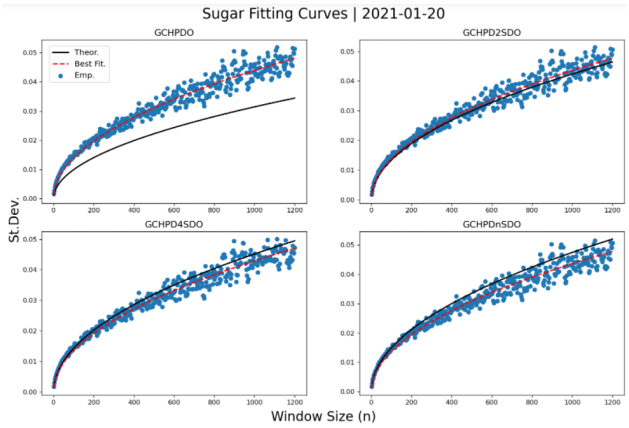


Figure 3.7: Best fit : GCHP2SDO

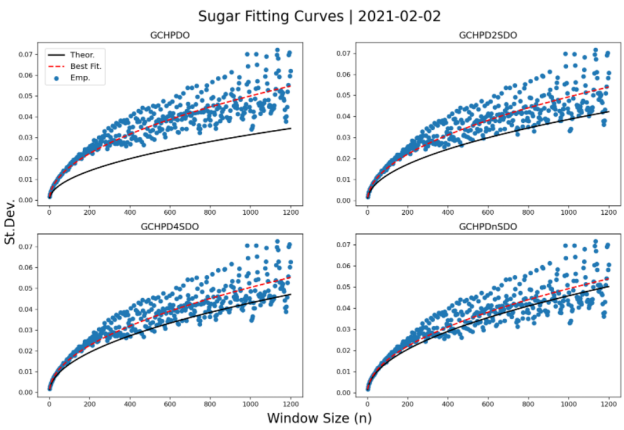


Figure 3.8: SB-Best fit : GCHPnSDO

	Best Fit's	% of Time Best Fit	Mean Error Rate	Best Error Rate	Worst Error Rate
GCHPDO	4	3.2	8.26	2.43	10.54
GCHP2sDO	37	29.6	5.41	0.01	14.13
GCHP4DO	25	20	3.24	0.33	10.68
GCHPnSDO	59	47.2	7.96	0.34	23.06

Table 3.3: Summary statistics for GCHP model fitting process over all sugar data.

model that fits best most often varies for distinct stocks and sometimes one model may never be chosen for a particular stock. For instance the AMZN stock data is not fit well by fixed tick models, but is well fit by adaptive tick models. This correlates with figure 2.5 which shows that AMZN has a more varied distribution of mid-price movements than the other stocks. Further analyzing the results of the fitting process it can be noted that days for which the best fitting model has a large error rate usually correspond to a day in which there was an abnormally large price swing. Summary statistics for the fitting process on a few different stocks are shown in tables B.1-B.4 in appendix B.

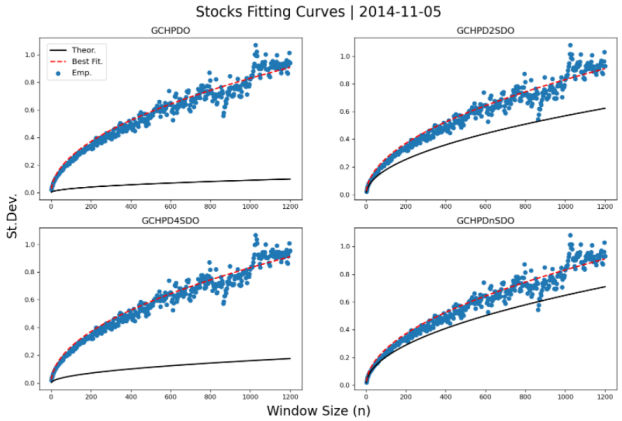


Figure 3.9: AMZN Best fit : GCHP2SDO

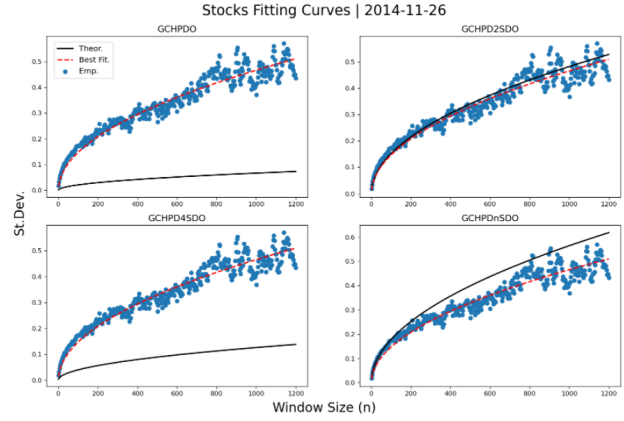


Figure 3.10: AMZN Best fit : GCHPnSDO

Chapter 4

Predictive methods for General Compound Hawkes Processes

4.1 Experimental Procedures

Now that we have detailed the procedures for fitting we now turn to the task of generating predictions using the GCHP. In particular we seek to apply the GCHP models to the practical problem of predicting strictly intra-day mid-price movements in a way that is prudent to the task of trading. As such we develop an experimental procedure with three three main objectives in mind:

- accurately predict directional mid-price movements above a certain size.
- gauge potential volatility in the mid-price process.
- minimize absolute error between true and predicted mid-prices.

These objectives though not numerically exact, if endeavoured to achieve will increase the probability of success for any potential trading strategy. For the first objective we create a machine learning like classification problem by setting a threshold α to organize upward and downward mid-price movements. That is any mid-price movement larger than α will be classified as an upward movement, any mid-price movement below $-\alpha$ will be classified as a downwards movement and any movement in between will be classified as stationary. This gives another parameter to tweak as the optimal value of this threshold can vary based on various factors such as the length of and volatility inherent to the chosen training interval. This threshold is also important in creating balanced classes which helps avoid over-predicting minority classes which in this context can lead to inaction if the minority class is the stationary class, or to trigger happy behaviour if the minority class is any other. Two approaches for setting this parameter are as follows:

- fix α manually across all tests.
- set α proportional to the volatility of the mid-price process over the training interval.

The first of these approaches provides standardized test results, but can also overlook the impact of local behaviour. As such the second approach better accounts for market events and looks to adapt tests to intra-day conditions. For the second objective a 2-class classification problem is constructed similarly to the 3-class problem with the difference being that all mid-price movements are classified according to a threshold β in absolute value. This provides a means to assess the models potential to predict extreme events irrespective of their direction. Lastly for the third objective we simply compare the final predicted mid-price to the corresponding realized value using historical data and analyze the distribution of differences between the corresponding true and predicted mid-prices. Together these goals give two separate target labels for each test conducted based on the difference between the future and current mid-price which are defined as follows.

$$\text{Price Pred.} = \begin{cases} 1 \ , \ S_t - S_0 \geq \alpha \\ 0 \ , \ \alpha > S_t - S_0 \geq -\alpha \\ -1 \ , \ -\alpha > S_t - S_0 \end{cases} \qquad \text{Vol. pred.} = \begin{cases} 1 \ , \ |S_t - S_0| \geq \beta \\ 0 \ , \ |S_t - S_0| < \beta \end{cases}$$

Another consideration is the amount of time allotted to calibrate model parameters alongside the length of the corresponding testing interval. In general the longer the testing interval the less reliable any prediction is going to be due to the intrinsic randomness of financial data. Also of note is the fact that the optimal model parameters can vary greatly between different days and that overnight activity, which is not represented in the data used, can change market conditions drastically. For these reasons all experiments are conducted using an out of sample testing scheme illustrated in figure 4.1 in which we treat every day separately. That is to say that at the start of every day we re-train the model with no reference to the previous days parameters.

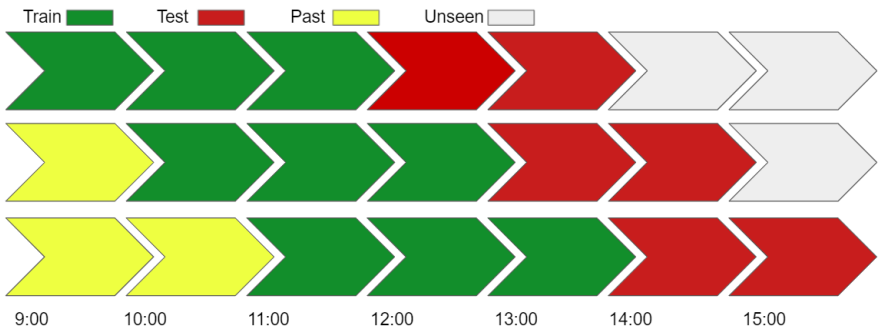


Figure 4.1: Illustration of walk forward testing scheme with a three hour to two hour train to test split and an one hour step size.

This set-up generates multiple tests per day for a given data-set which helps alleviate the problems typically associated with limited sample size, however, decreasing the length of the training sets in this scheme can result in choosing more subsets of data that that are not well-fit by the GCHP models due to small stents of atypical behaviour. Further, given the intent to trade we must also be aware of periods of extreme volatility, whether high or low, as this scheme is rigid in the times at which it will generate a new prediction. Conscious of these structural issues, tests are filtered to help maximize prediction accuracy and subsequently trading results. This is done by firstly by ensuring the error rate of the model fitting process over each respective training set is below 20%, and secondly by checking the level of volatility in the mid-price process over each training interval relative to recent times. In particular tests are filtered based on whether or not the standard deviation of the mid-price falls within extreme upper and lower quantiles of the standard deviation of a number of previous training sets. These criteria were implemented in attempts to improve performance statistics by ensuring that ill fitting models are not used and extreme volatility regimes do not unduly influence test results simply as a result of bad timing. In this way the focus of any tests preformed is centered towards predicting mid-price movements of more than a couple of ticks which generally give more actionable grounds for trading. All the described parameters were set after extensive testing in order to create balanced classes and give a sufficient time window for the model fitting process to function appropriately.

4.2 Prediction using Diffusive Limits

In section 3.2 diffusive limits for the GCHP were stated in theorems 3.13 and 3.15 respectively. The first approach to prediction done in this work is to apply these diffusive limits of the mid-price process to make a prediction for the future mid-price at a set time past the end of a designated training window. For each iterative training window this entails fitting all variations of the GCHP model to find the best fitting model, calibrating the corresponding parameters and making a numeric prediction for the future mid-price. This approach is further broken down into three different methods that use slightly different formulations for the predicted mid-price. The first of these formulations is derived from the following corollary to theorem 3.15.

Corollary 4.1. *[41] Given the pure diffusive limit for the mid-price process stated it follows that $S(t)$ can be approximated by,*

$$S(t) \approx S(0) + a^* \frac{\lambda}{1-\mu} t + \bar{\sigma} W(t) \tag{4.1}$$

where $W(t)$ is a standard Brownian motion.

Now, given $W(t) \sim \sqrt{t}\, N(0,1)$ it follows that,

$$S(t) \approx S(0) + a^* \frac{\lambda}{1-\hat{\mu}} t + \bar{\sigma} W(t) \Longrightarrow S(t) \approx S(0) + a^* \frac{\lambda}{1-\hat{\mu}} t + \bar{\sigma} \sqrt{t} N(0,1) \tag{3.11}$$

and from here forward (3.11) will be referred to as the pure diffusive limit of the mid-price process. Based off this corollary a simpler formula can be given by noting that the expectation of a standard Brownian motion is zero. Thus the following approximation can be used as a proxy for the pure-diffusive limit for the mid-price on a large time interval. This lessens computation times at the cost of the generality given by the drift term of the original formula.

$$\mathbb{E}[S(t)] \approx S(0) + a^* \frac{\lambda}{1-\mu} t \tag{3.12}$$

This equation (3.12) will be referred to as the expectation of the pure diffusive limit.

Lastly, based on theorem 3.12 [24, 45], a formula for the future mid-price can be derived from the jump diffusion limit which directly incorporates a simulated Hawkes process into the predicted value.

$$\begin{aligned}
&\frac{S(nt) - N(nt)a^*}{\sqrt{n}} \to \sigma \sqrt{\frac{\lambda}{1-\hat{\mu}}} W(t) \\
&\implies S(nt) - N(nt)a^* \to \sigma \sqrt{nt} \sqrt{\frac{\lambda}{1-\hat{\mu}}} N(0,1) \\
&\implies S(nt) \to N(nt)a^* + \sigma \sqrt{nt} \sqrt{\frac{\lambda}{1-\hat{\mu}}} N(0,1)
\end{aligned}$$

It then follows that,

$$S(nt) \approx S(0) + N(nt)a^* + \sigma \sqrt{nt} \sqrt{\frac{\lambda}{1-\hat{\mu}}} N(0,1) \tag{3.13}$$

is another viable formula for the predicted mid-price and this equation (3.13) shall be referred to as the jump diffusive limit. In this approach many predictions are generated based off repeated simulations of a standard normal random variable and the mean of these simulations is taken as the final prediction.

Results

The following is a preliminary test done on the sugar futures data-set using fixed hyper-parameters to establish the generality of our experimental set-up over a large time interval, and to establish a means of comparison for each proposed prediction method. These hyper-parameters are as follows,

- three hour training window, two hour prediction horizon and one hour step size.
- α threshold set to $2/3$ of the standard deviation of the mid-price process over the training interval.
- β threshold set to the standard deviation of the mid-price process over the training interval .

Further tests will be done later to optimize these hyper-parameters to more localized values better reflecting intra-day market conditions. To summarize the results for the three class and two class problems classification reports are shown in tables 4.1 and 4.6 and confusion matrices are shown in figures 4.2 and 4.3 respectively.

Table 4.1: Classification reports for three class predictions using diffusive limits on sugar data

Class	prec	recall	f1	sup
-1	0.53	0.42	0.47	112
0	0.19	0.27	0.22	33
1	0.41	0.45	0.43	77
acc	0.41			
m-avg	0.38	0.38	0.37	222
w-avg	0.44	0.41	0.42	222

Table 4.2: SB PDL 3C-CR.

class	prec	recall	f1	sup
-1	0.53	0.41	0.46	112
0	0.16	0.24	0.19	33
1	0.42	0.47	0.44	77
acc	0.41			
m-avg	0.37	0.37	0.37	222
w-avg	0.44	0.41	0.42	222

Table 4.3: SB EPDL 3C-CR.

class	prec	recall	f1	sup
-1	0.53	0.41	0.46	112
0	0.16	0.24	0.19	33
1	0.42	0.47	0.44	77
acc	0.41			
m-avg	0.37	0.37	0.37	222
w-avg	0.44	0.41	0.42	222

Table 4.4: SB JDL 3C-CR.

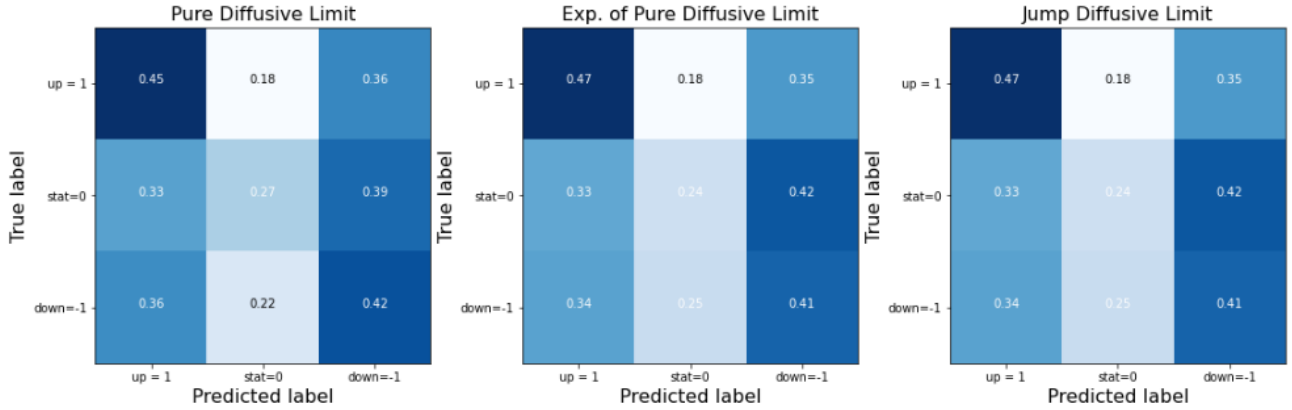


Figure 4.2: Confusion matrices for three class problem on sugar data using diffusive limits.

Table 4.5: Classification reports for two class predictions using diffusive limits on sugar data

Class	prec	recall	f1	sup
0	0.29	0.43	0.35	60
1	0.74	0.60	0.67	162
acc	0.56			
m-avg	0.52	0.52	0.51	222
w-avg	0.62	0.56	0.58	222

Table 4.6: SB PDL 2C-CR.

class	prec	recall	f1	sup
0	0.30	0.45	0.36	60
1	0.75	0.61	0.67	162
acc	0.57			
m-avg	0.53	0.53	0.52	222
w-avg	0.63	0.57	0.59	222

Table 4.7: SB EPDL 2C-CR.

class	prec	recall	f1	sup
0	0.30	0.45	0.36	60
1	0.75	0.60	0.67	162
acc	0.56			
m-avg	0.52	0.53	0.51	222
w-avg	0.63	0.56	0.58	222

Table 4.8: SB JDL 2C-CR.

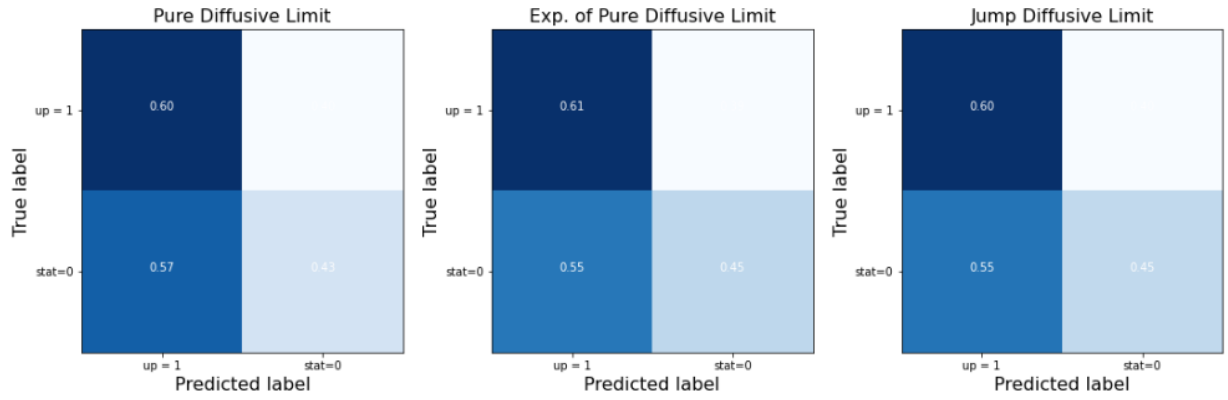


Figure 4.3: Confusion matrices for two class problem on sugar data using diffusive limits.

In this test each respective approach generated similar predictions with small variations in size usually not more than a couple of ticks. Overall predictive accuracy was several percent above random change with a strength in predicting non-stationary movements as evidenced by the precision and recall statistics shown in table 4.1 being above 0.4 for both the upwards and downwards classes. Also of note is that when the model prediction was wrong there was a higher tendency to predict an opposite directional movement than a stationary movement as shown in the confusion matrices for each approach in figure 4.2. This is caused by the removal of tests that took place over an abnormally low volatility regime in which the likely predicted movement was likely to be stationary resulting in a class imbalance in favour of the upwards and downwards classes comparative to the stationary class. Looking the statistics for the two class label the overall predictive accuracy is strong, but there was also a heavy class imbalance that is influencing the results.

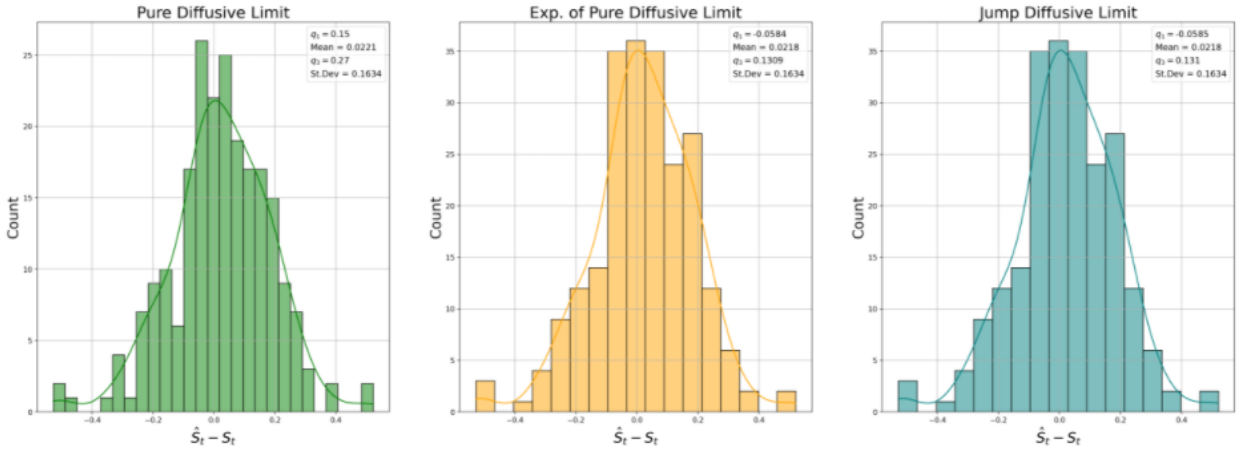


Figure 4.4: Histograms of error between predicted and true mid-price for diffusive limit methods for sugar futures data.

From figure 4.4 it can be seen that the error between the true values and the predicted values for the mid-price at the end of each testing interval do not follow a normal distribution. A positive note is that the mean error was around zero for all tests, but the presence of heavy tails in these histograms is indicative of a non-insignificant number of bad

predictions. So on average the final predicted value was close to the true value, but there are many cases where it can be off by many ticks as the standard deviation of these predictions is noticeably high.

Now considering the stock data-set, given that each individual stock within it shows different levels of liquidity as well as different standards for average mid-price changes, different hyper-parameters should be specifically calibrated for each different stock. However, to again establish a comparative analysis of the prediction methods the following standardized hyper-parameters are used.

- one hour training window, half hour prediction horizon and half hour step size.
- α threshold set to $2/3$ of the standard deviation of the mid-price process over the training interval.
- β threshold set to $4/5$ standard deviation of the mid-price process over the training interval.

Lowering the length of the training and test windows helps with issues related to the sample size of this stock data-set, we also adjust the volatility threshold β to reflect this shortening.

Table 4.9: Classification reports for three class predictions using diffusive limits on AMZN stock data

Class	prec	recall	f1	sup
-1	0.44	0.50	0.47	22
0	0.36	0.55	0.44	22
1	0.58	0.37	0.45	38
acc	0.45			
m-avg	0.46	0.47	0.45	82
w-avg	0.49	0.45	0.45	82

Table 4.10: AMZN PDL 3C-CR.

class	prec	recall	f1	sup
-1	0.46	0.50	0.48	22
0	0.38	0.59	0.46	22
1	0.58	0.37	0.45	38
acc	0.46			
m-avg	0.47	0.49	0.46	82
w-avg	0.50	0.46	0.46	82

Table 4.11: AMZN EPDL 3C-CR.

class	prec	recall	f1	sup
-1	0.46	0.50	0.48	22
0	0.38	0.59	0.46	22
1	0.58	0.37	0.45	38
acc	0.46			
m-avg	0.47	0.49	0.46	82
w-avg	0.50	0.46	0.46	82

Table 4.12: AMZN JDL 3C-CR.

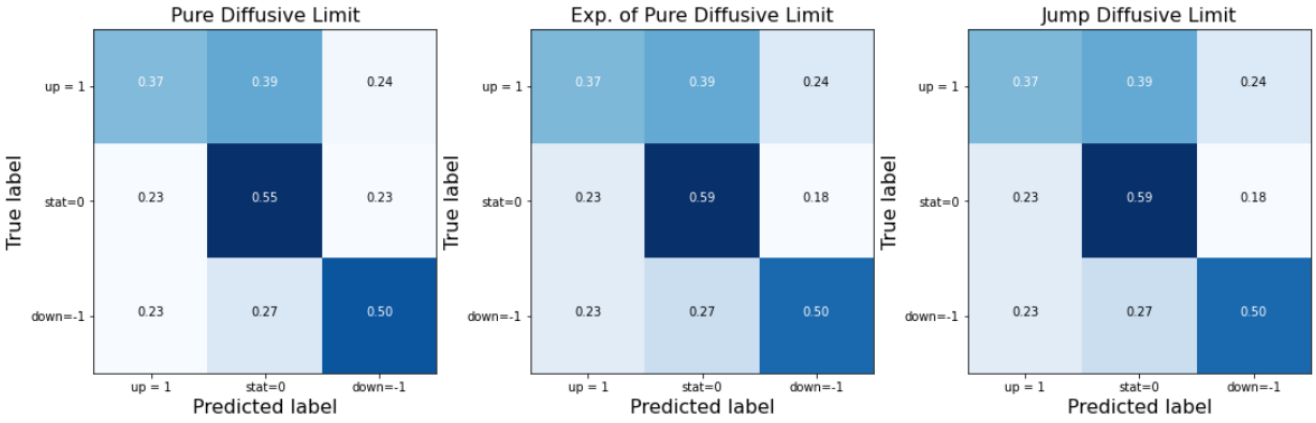


Figure 4.5: Classification reports and confusion matrices for three class problem on AMZN stock data using diffusive limits.

In this test each approach generated almost exactly the same directional predictions leading to identical summary statistics. The overall predictive accuracy is noted at 46% which is considerably high and the precision and recall

Table 4.13: Classification reports for two class predictions using diffusive limits on AMZN stock data

Class	prec	recall	f1	sup
0	0.39	0.59	0.47	27
1	0.73	0.55	0.62	55
acc	0.56			
m-avg	0.56	0.57	0.55	82
w-avg	0.62	0.56	0.57	82

Table 4.14: AMZN PDL 2C-CR.

class	prec	recall	f1	sup
0	0.41	0.59	0.48	27
1	0.74	0.58	0.65	55
acc	0.59			
m-avg	0.58	0.59	0.57	82
w-avg	0.63	0.59	0.60	82

Table 4.15: AMZN EPDL 2C-CR.

class	prec	recall	f1	sup
0	0.41	0.59	0.48	27
1	0.74	0.58	0.65	55
acc	0.59			
m-avg	0.58	0.59	0.57	82
w-avg	0.63	0.59	0.60	82

Table 4.16: AMZN JDL 2C-CR.

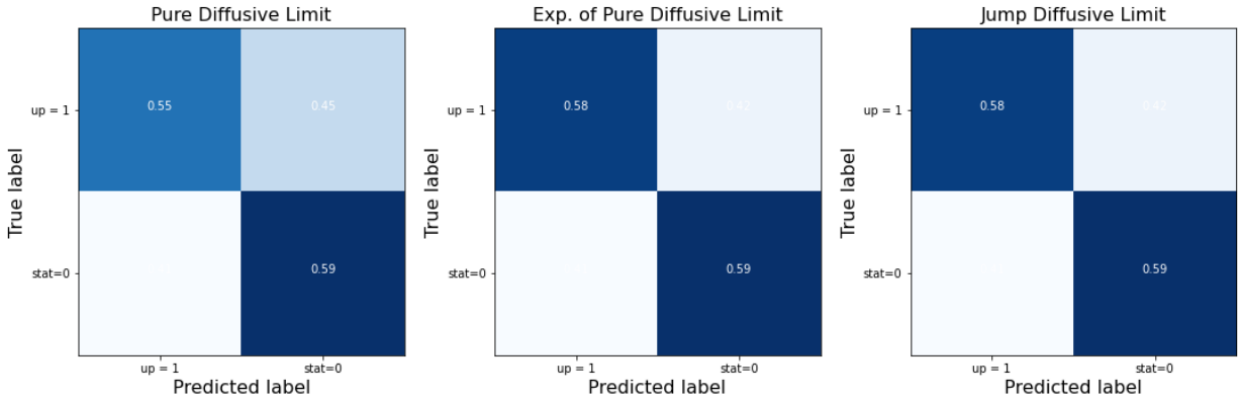


Figure 4.6: Classification reports and confusion matrices for two class problem on AMZN stock data using diffusive limits.

statistics are consistently strong across all classes. Also of note is that the confusion matrices for the directional prediction are close to diagonally dominant with a particular peak in predicting upwards movements. Looking the statistics for the two class label shows some variation in the predictions generated by each different approach while still maintaining the overall predictive accuracy multiple points above random chance.

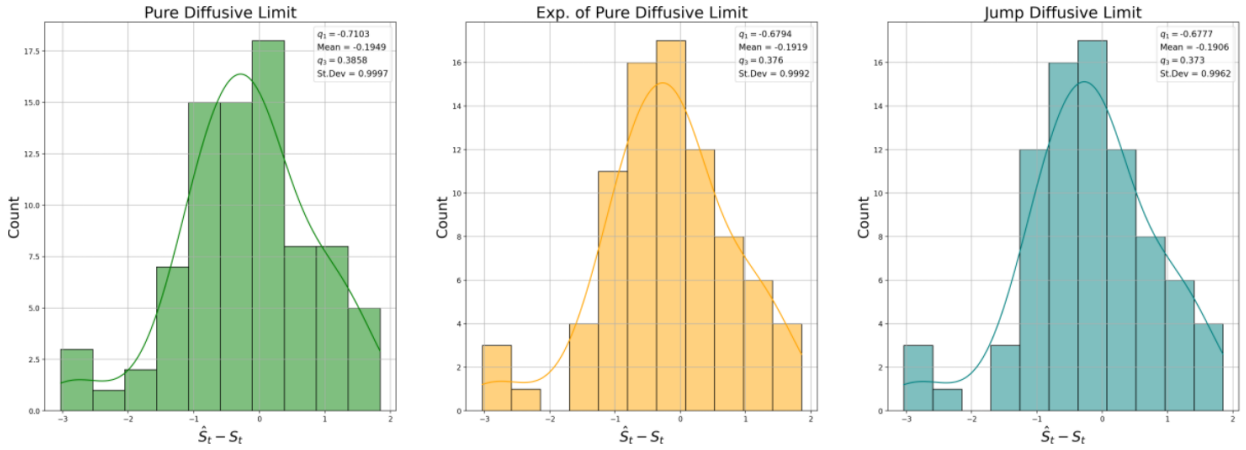


Figure 4.7: Histograms of error between predicted and true mid-price for diffusive limit methods for AMZN stock data.

From figure 4.7 it can be seen that the distribution of absolute errors are negatively skewed across each approach indicating a propensity to under predict the true value throughout this test which is further indicated by the mean error for all tests being negative. This leads to more variance in the predictions than those of the sugar futures predictions which is reflective of the overall volatility of each respective underlying.

It should be noted that conducting the same tests on different underlying assets using the same parameters will produce varying results. Thus it is agreeable that model parameters and experimental hyper-parameters should be optimized for each specific underlying asset tested. To show this the following test maintains the previously used parameters and applies them to FB stock data.

Table 4.17: Classification reports for three class predictions using diffusive limits on FB stock data

Class	prec	recall	f1	sup
-1	0.27	0.24	0.25	29
0	0.37	0.44	0.40	32
1	0.46	0.42	0.44	38
acc	0.37			
m-avg	0.36	0.37	0.36	99
w-avg	0.37	0.37	0.37	99

Table 4.18: FB PDL 3C-CR.

class	prec	recall	f1	sup
-1	0.27	0.24	0.25	29
0	0.38	0.47	0.42	32
1	0.45	0.39	0.42	38
acc	0.37			
m-avg	0.37	0.37	0.36	99
w-avg	0.37	0.37	0.37	99

Table 4.19: FB EPDL 3C-CR.

class	prec	recall	f1	sup
-1	0.27	0.24	0.25	29.00
0	0.38	0.47	0.42	32.00
1	0.45	0.39	0.42	38.00
acc	0.37			
m-avg	0.37	0.37	0.36	99
w-avg	0.37	0.37	0.37	99

Table 4.20: FB JDL 3C-CR.

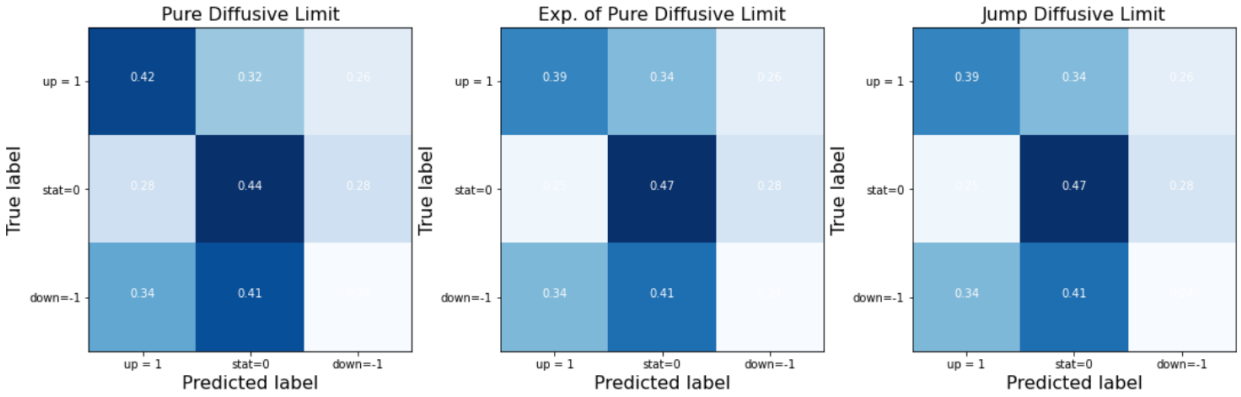


Figure 4.8: Confusion matrices for thee class problem on FB stock data using diffusive limits.

Table 4.21: Classification reports for two class predictions using diffusive limits on FB stock data

Class	prec	recall	f1	sup
0	0.50	0.53	0.52	45
1	0.59	0.56	0.57	54
acc	0.55		0	
m-avg	0.54	0.54	0.54	99
w-avg	0.55	0.55	0.55	99

Table 4.22: FB PDL 2C-CR.

class	prec	recall	f1	sup
0	0.50	0.53	0.52	45
1	0.59	0.56	0.57	54
acc	0.55			
m-avg	0.54	0.54	0.54	99
w-avg	0.55	0.55	0.55	99

Table 4.23: FB EPDL 2C-CR.

class	prec	recall	f1	sup
0	0.50	0.53	0.52	45
1	0.59	0.56	0.57	54
acc	0.55			
m-avg	0.54	0.54	0.54	99
w-avg	0.55	0.55	0.55	99

Table 4.24: FB JDL 2C-CR.

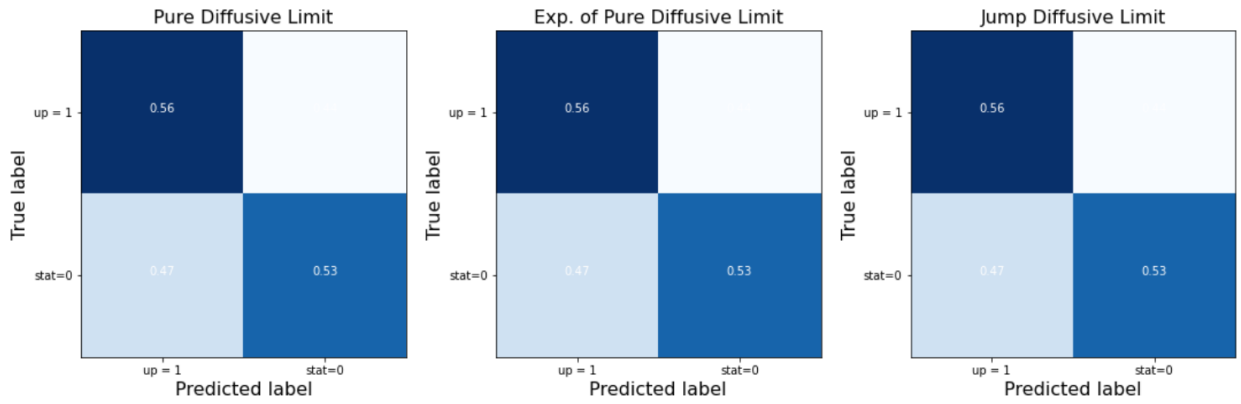


Figure 4.9: Confusion matrices for two class problem on FB stock data using diffusive limits.

Again each approach generated similar directional predictions, but the overall predictive accuracy is noted at a high of 37% across all approaches which is much lower than that reported in the AMZN stock tests. Looking at the confusion matrices in figure 4.8 shows that in this test the model struggled to predict downwards movements, but the statistics for the two class label shown in figure 4.9 indicate the model was still able to consistency predict the size of movements at a rate above random chance. From figure 4.10 it can be seen that the distribution of absolute errors has no significant skew for each approach indicating a higher level of randomness to the predictions in this test as compared to the previous one. Taken together these results connote that the model was better at predicting the size of mid-price movements in the FB data as compared to their direction when given the standard parameters.

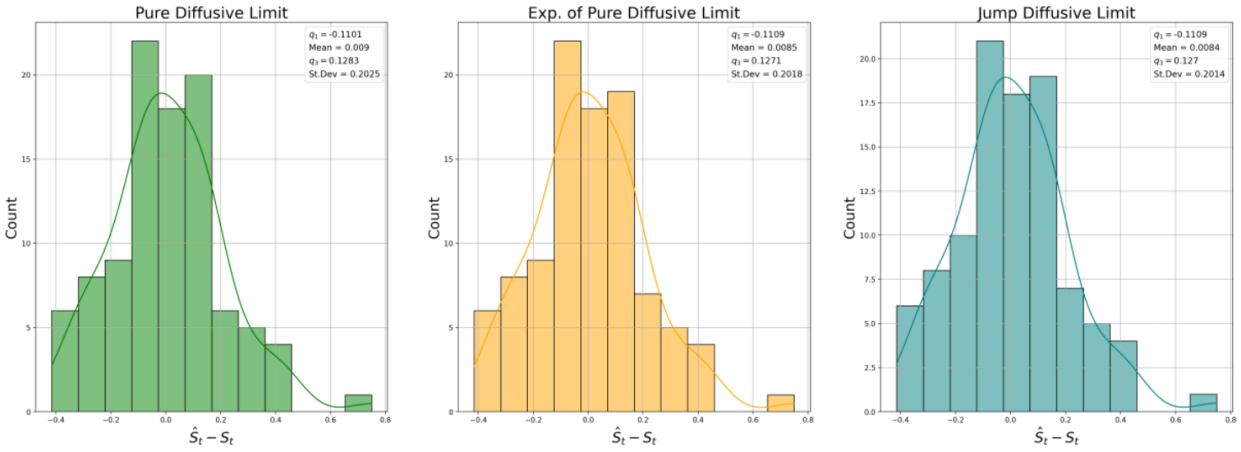


Figure 4.10: Histograms of error between predicted and true mid-price for diffusive limit methods for FB stock data.

4.3 Prediction using Repeated Simulation

An alternative approach to attempting prediction tasks is to simulate many paths of the underlying mid-price and use an aggregation of these paths to form a final prediction in a Monte-Carlo-esque manner. To accomplish this using

a GCHP model there are two steps, the first of which is to simulate a Hawkes process over the testing interval to yield a series of event times. Then for each one of these event times a mid-price movement can be simulated based on the state space of the best fitting GCHP model over the current training interval. That is to say that using the current state of the model corresponding to the previous observed mid-price change, the next mid-price movement is simulated using the distribution of transition probabilities taken from the corresponding row of the empirical transition matrix. Together these steps produce a random number of timed mid-price changes each of which is given a direction and size. The cumulative effect of these simulated events then forms a predicted mid-price path. To form the final predicted mid-price many paths are generated and the mean of the end points of all these simulated paths is taken as the final numeric prediction. Figures 4.11 and 4.12 shows examples of a collection of simulated mid-price paths for the sugar futures data and AMZN stock data respectively. It can be observed that no one path exactly follows the realized mid-price, but through repeated simulation the direction of the mid-price can be captured by the general trajectory of all the simulated paths.

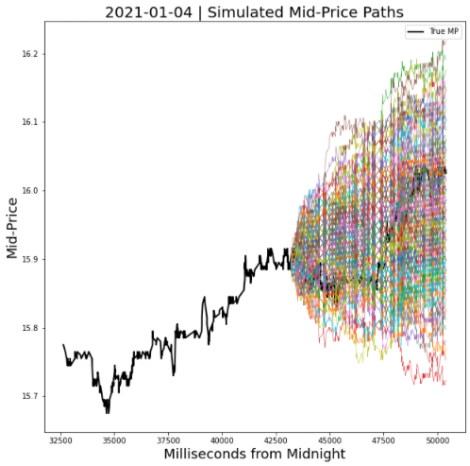


Figure 4.11: Example of simulated Mid-price paths for SB data.

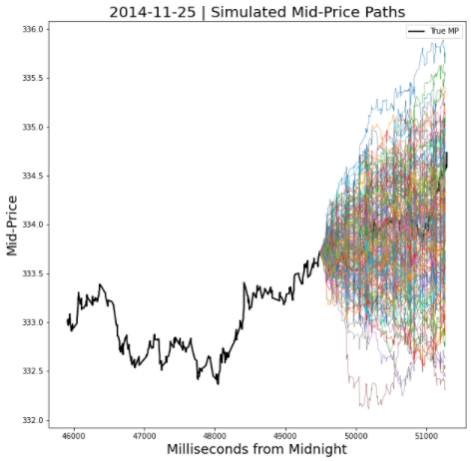


Figure 4.12: Example of simulated Mid-price paths for AMZN data.

Using this method introduces a new parameter in the number of simulated paths generated, this affects both the computational cost of running tests and has the potential to introduce mean reversion into the results. To elaborate, this method is burdened by an inherent component of randomness that can be mitigated by increasing the number of simulated paths used in creating the final prediction. As such, running the same test multiple times will produce slightly different results every time, but it was found that performance statistics would only tend to vary on average by a couple of percentage points. There are both pros and cons to increasing and decreasing the number of simulated paths generated. Overall after various tests it was found that this method can often struggle to capture large price swings and trend reversals. As such the train to test split can drastically affect the results, but to gather a means of

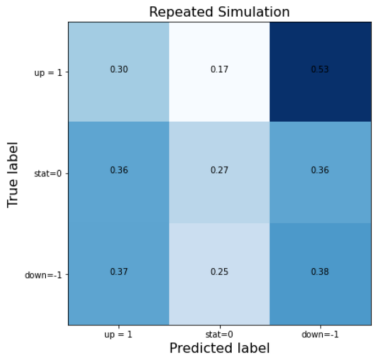
comparison to the diffusive limit approaches all hyper-parameters are kept the same and results are reported.

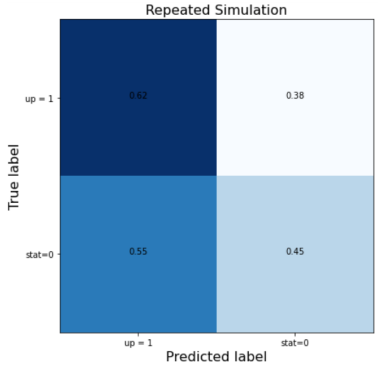
Results

The following tests were all conducted using the same parameters as were previously used for the diffusive limit tests. In the tests for the sugars futures data and the AMZN stock data the directional accuracy was found to be lower than the diffusive limits methods, but the mean error was also found to be lower. While in the test for the FB stock data the three class accuracy was found to be higher and the mean error was significantly lower. In general it was found that the repeated simulation approach produced better or worse results depending on the underlying asset at the cost of higher computation times than the diffusive limit methods. All results for this method are shown in figures 4.13, 4.14 and 4.15 for the SB, AMZN and FB data respectively.

Class	precision	recall	f1	sup
-1	0.45	0.38	0.41	112
0	0.18	0.27	0.22	33
1	0.30	0.30	0.30	77
acc	0.34			
m-avg	0.31	0.32	0.31	222
w-avg	0.36	0.34	0.35	222

Class	precision	recall	f1	sup
0	0.30	0.45	0.36	60
1	0.75	0.62	0.68	162
acc	0.57			
m-avg	0.53	0.53	0.52	222
w-avg	0.63	0.57	0.59	222





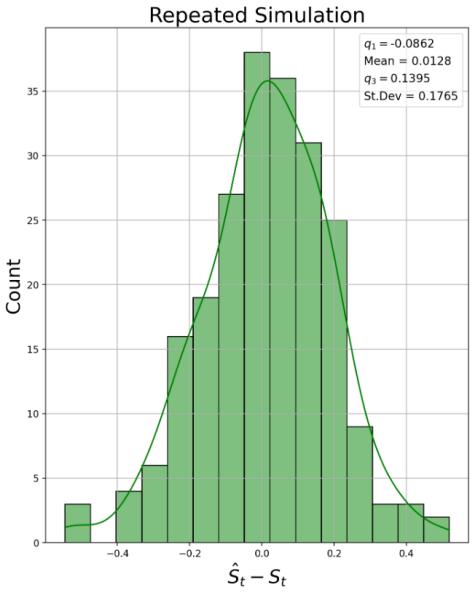
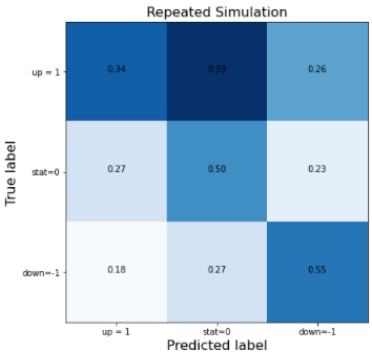
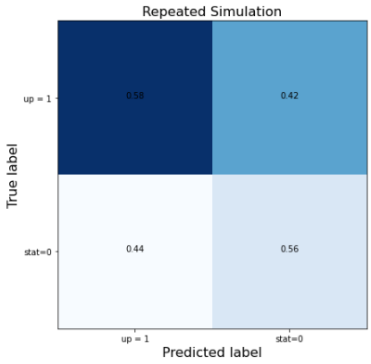


Figure 4.13: Classification reports, confusion matrices and histogram of error between predicted and true mid-price for repeated simulation methods for SB futures data.

Class	precision	recall	f1	sup
-1	0.44	0.55	0.49	22
0	0.34	0.50	0.41	22
1	0.57	0.34	0.43	38
acc	0.44			
m-avg	0.45	0.46	0.44	82
w-avg	0.47	0.44	0.44	82



Class	precision	recall	f1	sup
0	0.39	0.56	0.46	27
1	0.73	0.58	0.65	55
acc	0.57			
m-avg	0.56	0.57	0.55	82
w-avg	0.62	0.57	0.59	82



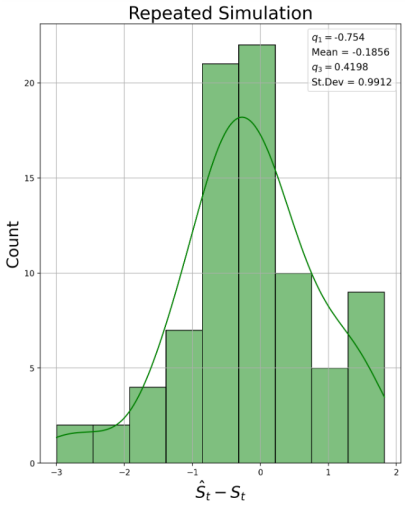
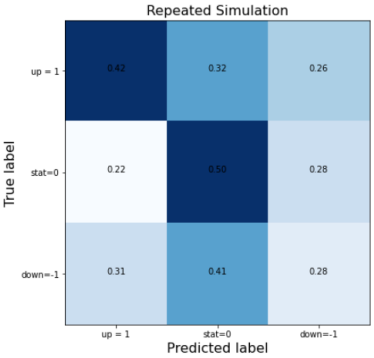
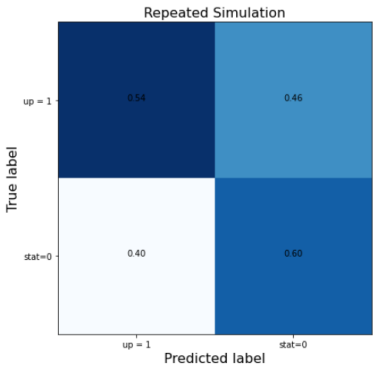


Figure 4.14: Classification reports, confusion matrices and histogram of error between predicted and true mid-price for repeated simulation methods for AMZN stock data.

Class	precision	recall	f1	sup
-1	0.30	0.28	0.29	29
0	0.40	0.50	0.44	32
1	0.50	0.42	0.46	38
acc	0.40			
m-avg	0.40	0.40	0.40	99
w-avg	0.41	0.40	0.40	99



Class	precision	recall	f1	sup
0	0.52	0.60	0.56	45
1	0.62	0.54	0.57	54
acc	0.57			
m-avg	0.57	0.57	0.57	99
w-avg	0.57	0.57	0.57	99



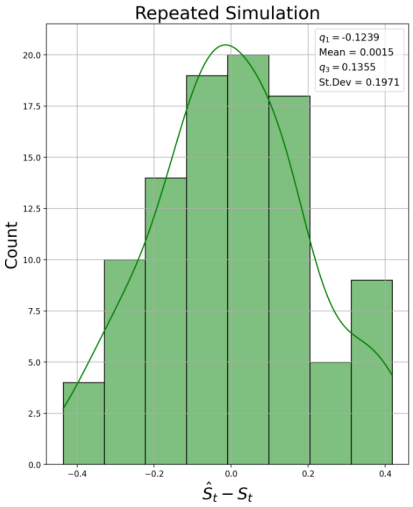


Figure 4.15: Classification reports, confusion matrices and histogram of error between predicted and true mid-price for repeated simulation methods for FB stock data.

4.4 Comparison of Prediction Methods

When comparing the end results for all tests it can generally stated that each prediction method generates highly similar results with a couple of exceptions. In particular it was not uncommon for the different methods to produce exactly the same level of accuracy irregardless of the underlying asset. In all tests conducted the overall accuracy on the three class and two class predictions as well as the marco averaged f1-scores were shown to be several percentage points above random chance levels. Further the mean error of the numeric predictions was shown to be no more than a couple of ticks in all instances. A summary of test statistics is shown in table 4.25 below. Given these results it

Asset	Method	3c - Acc	3c - M-Avg F1	2c - Acc	2c - M-Avg F1	Error Mean	Error Std-Dev
SB	PDL	0.41	0.37	0.56	0.51	0.0221	0.1634
	EPDL	0.41	0.37	0.57	0.52	0.0218	0.1634
	JDL	0.41	0.37	0.56	0.51	0.0218	0.1634
	RS	0.34	0.31	0.57	0.52	0.0128	0.1765
AMZN	PDL	0.45	0.45	0.56	0.55	-0.1949	0.9997
	EPDL	0.46	0.46	0.59	0.57	-0.1919	0.9992
	JDL	0.46	0.46	0.59	0.57	-0.1906	0.9962
	RS	0.44	0.44	0.57	0.55	-0.1856	0.9912
FB	PDL	0.37	0.36	0.55	0.54	0.0090	0.2025
	EPDL	0.37	0.36	0.55	0.54	0.0085	0.2018
	JDL	0.37	0.36	0.55	0.54	0.0084	0.2014
	RS	0.40	0.40	0.57	0.57	0.0015	0.1971

Table 4.25: Summary statistics for mid-price prediction classification problems for all methods.

is our conclusion that each prediction method is relatively interchangeable and that the practical application of these methods is not inhibited or enabled by any particular method tested. The small differences in classification metrics across methods can be seen as attributable to the fixed margins for classes combined with the fixed testing times. One practical consideration when choosing which method to use is the computational cost associated with each method and given the previously stated conclusion, using the expectation of the pure diffusive limit can be seen as desirable in situations where lower computation times are valued.

4.5 Cross Validation results for GCHP Prediction Methods

One restriction inherent to the previous test results is that they were computed using standardized hyper-parameters across a large time interval. To get a better sense of the generalizability of the GCHP model predictions and improve performance statistics we now look to improve the previous results by optimizing the experimental hyper-parameters to localized values using cross validation. These hyper-parameters include the length of the training and test intervals, the step size of the walk forward scheme and the multiplier applied to the standard deviation where computing class thresholds. Over the course of testing it was found that prioritizing parameters that give high f1 scores typically leads

to better out of sample generalization and that using the expectation of the pure diffusive limit to lessen computation times was ideal. The following results are for a forty day subset of the sugar data over which an eight to eight day train to test split was applied. Tables 4.26 and 4.27 show classification statistics for each cross validation fold and figures 4.16 and 4.17 give the three class confusion matrix and histogram of error respectively for all out of sample predictions.

Fold	Acc	MA F1	1 F1	0 F1	-1 F1	Supp
0	0.46	0.46	0.54	0.30	0.52	52.00
1	0.57	0.56	0.55	0.60	0.53	23.00
2	0.55	0.44	0.21	0.70	0.40	47.00
3	0.33	0.32	0.34	0.39	0.23	91.00

Table 4.26: GCHP in sample classification results.

Fold	Acc	MA F1	1 F1	0 F1	-1 F1	Supp
0	0.50	0.51	0.55	0.47	0.50	42.00
1	0.38	0.38	0.35	0.33	0.47	29.00
2	0.41	0.25	0.17	0.56	0.00	49.00
3	0.42	0.43	0.55	0.30	0.45	80.00

Table 4.27: GCHP out of sample classification results.

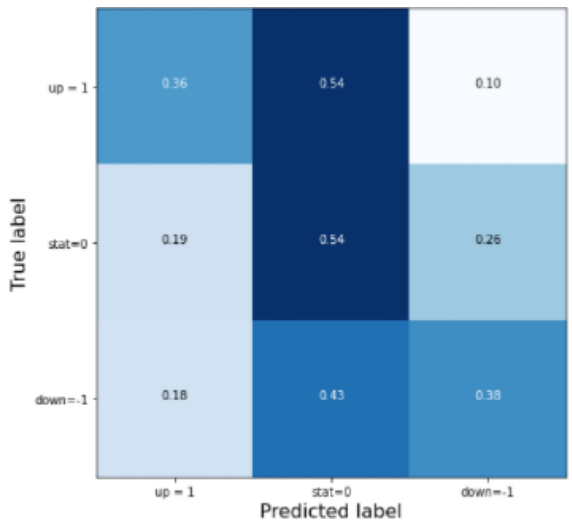


Figure 4.16: Confusion matrix for SB cross validation predictions.

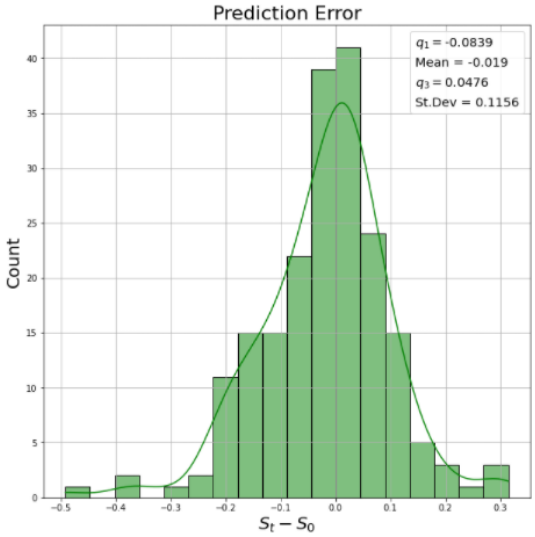


Figure 4.17: Histogram of error for SB cross validation predictions.

When compared to the results of the standardized test the corresponding in sample results showed improvements across all metrics with accuracy scores for certain folds peaking above fifty percent. Despite this it can also be noted that there can be significant variation between the performance for different folds as well as selected sample size. Classification statistics across all filtered out of sample predictions are shown in table 4.28 and together these results indicate that there was a tendency to over predict the middle class, but promising overall prediction accuracy.

Class	precision	recall	f1	sup
-1	0.48	0.38	0.43	60
0	0.36	0.54	0.44	68
1	0.52	0.36	0.46	72
acc	0.43			
m-avg	0.45	0.43	0.43	200
w-avg	0.45	0.43	0.43	200

Table 4.28: GCHP out of sample CV Classification report for subset of sugar data.

To assess the practical viability of this set-up we tested a simple trading strategy that uses these out of sample predictions to trade one lot at a time based on the most viable directional predictions. This strategy enters a trade if an upwards or downwards prediction is made and exits said trade if either an opposing prediction is made, a bad model fit is detected, an extreme volatility regime is detected or an early exit criteria is hit. Having an early exit criteria is advisable in this situation given that predictions are made at fixed times due to the computation time it takes to generate said prediction. This leaves the strategy susceptible to sudden moves or trend reversals and as such, if either a profit target or a stop loss is hit the trade will be exited accordingly and a new trade will not be considered until the time of the next prediction. Ideal example trading days are shown in figures 4.18 and 4.19 in which a single trade was made over the course of a series repeating downwards and upwards predictions respectively generated once every fifteen minutes.

Figure 4.18: Example downwards trading day for GCHP strategy.

Figure 4.19: Example upwards trading day for GCHP strategy.

This strategy seeks only to make trades at times of relative stability and to capitalize on consistent trends. It is naturally limited in the number of times it can trade in a day and actively avoids times of high volatility in which the GCHP model fit is found to be worse than average. The cumulative results for this strategy are shown below in table 4.29.

PnL	Mean Trade Result	Median Trade Result	Trades	Win Rate
1.03	0.02102	0.01	49.0	51.02

Table 4.29: Summary trading statistics for GCHP strategy on forty day subset of sugar data.

Overall the results are positive, but despite this it should be noted that this test was limited in sample size and that it makes some assumptions that are not realistic in that it trades at the mid-price without having to use order queues. Nevertheless, it suffices to indicate that the application of GCHP model predictions is compatible with trading activities given proper set-up and optimized parameters.

Chapter 5

Hawkes Intensity Based Signals

5.1 Background

The GCHP models give a parsimonious representation of the stochastic dynamics of a limit order book through a direct analysis of the mid-price. This approach is justified given that the most influential activity in an order-book book occurs around its centre, but this ignores the tangible relationship between order-flow and mid-price action. As a result of the proliferation of limit order book data many have documented this relationship and shown that the impact of order-flow is one of the main drivers of price variations [3]. Further Hawkes processes, being known to exhibit properties of interest in modelling high frequency financial data, have been used to model order-flow in various ways [30]. One typical approach is to specify distinct sets of order types, most commonly buy and sell market orders, and fit either a uni-variate Hawkes process to their independent arrival processes, or a multivariate Hawkes process to their joint arrival process. In [12] the intensity of these processes was then used to define a trading signal based on a threshold of the ratio of buy to sell intensity at any given time.

Different views exist in regards to how to best manage the information contained within the expanding levels of an order-book and what specific variation on the Hawkes process best embodies the stylized facts of the corresponding high-frequency data. For instance, in [1] a multivariate marked Hawkes process was used to build a model of the first level of an order-book and it was found that the arrival process of orders is dependent on the volume as well as the past history of orders. In [30] a state dependent Hawkes process which is a multivariate Hawkes process coupled with a state process was used to model the relationship between order-flow and the shape of a LOB. So, given this established importance of order-flow in existing literature this section details how to build a trade-able signal from historical order-flow using simple uni-variate exponential Hawkes processes. It is not our goal to build a specific execution strategy, if this were the case there are many dynamics to consider including the types of orders to use, state variables such as the

spread, target setting given observed volumes and volatility, permanent vs. temporary price impact and many more. These dynamics have prompted the definition of optimization problems such as finding optimal acquisition/liquidation strategies for large orders or finding the optimal time to switch a limit order to a market order. For an overview of the framework and examples of such problems one can see [11, 25]. We instead look to establish proof of concept through the simplifying assumptions of being able to trade instantaneously at any time throughout active trading hours, having no price impact and of always using market orders to back-test computed signals. To this end we implement procedures to build a Hawkes intensity based signal calibrated to the timing of recorded changes in order queues. In this way we show that the application Hawkes processes is not limited to modelling the mid-price and that they can be used to effectively model any level of bid or ask price and volume. The proposed model and strategy serves as a baseline for future enhancements and was developed in awareness of the potential challenges associated with real time trading.

5.2 Classifying Order-book Activity

The expeditious while broad nature of order-book data suggests two contrasting approaches to surmise market activity within an intra-day time-frame. These two approaches give differing levels of detail that highlight the trade-offs between the amount of information available, the amount of information that can be processed reasonably in finite time and how much of that information is relevant to predictable outcomes. The first approach applied here is to directly monitor for changes in the level one order queues which float around the mid-price. This approach is supported by the fact that the trigger for a mid-price change is a level one queue depletion and that the level one queues are typically the most active of all queues in the order-book. Monitoring for changes in these queues, we can define buy and sell events based on the directional pressure they exert on the mid-price. In particular, buy events can be seen as those which exert upward pressure on the mid-price and sell events can be seen as those that exert downward pressure on the mid-price. As such, the following is a simple way of classifying buy and sell events based on observed order queue changes:

- Buy Event - increase in bid queue or decrease in ask queue.
- Sell Event - decrease in bid queue or increase in ask queue.

An algorithmic breakdown of this process can be found in appendix A. This classification is reasonably apt, but does not account for the size of these events nor does it account for the way in which the data progresses at times of price movement. In particular at times in which price movement is observed, it is common that the price level at which the level one order queues sit shifts up or down in response. This is further complicated by the fact that the direction of price movement changes the implied effect. Given this, we assume that if the bid or ask price changes the size of the

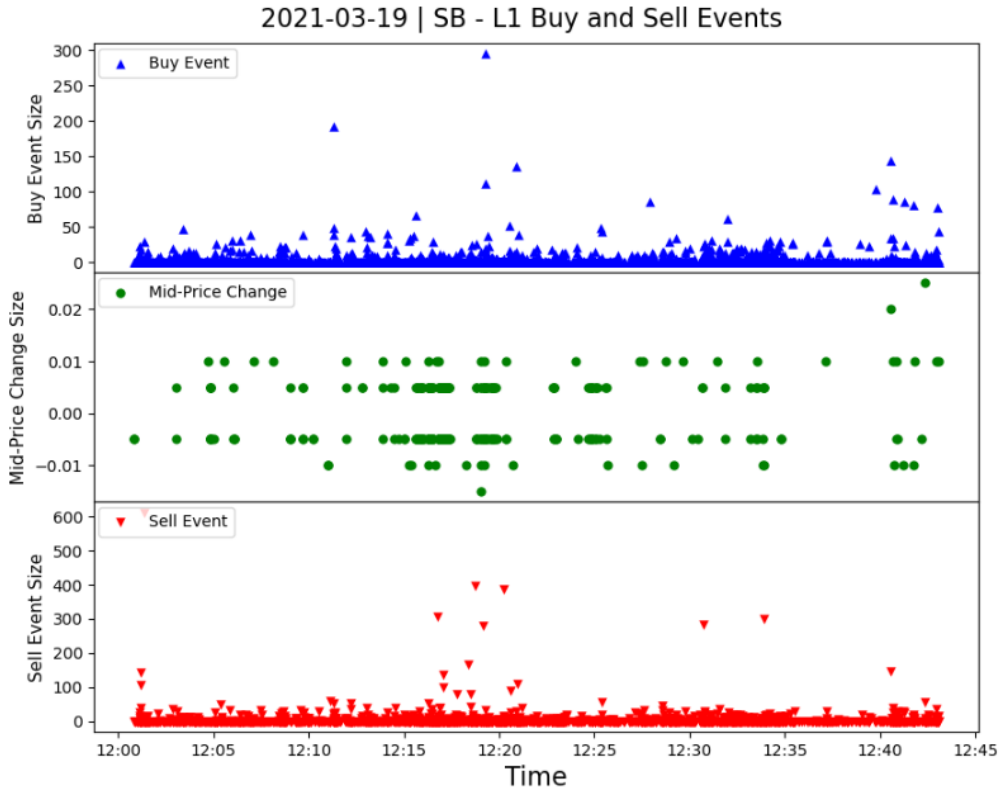


Figure 5.1: Clustering of level one order queue buy and sell events for one hour of SB data.

corresponding event is defined in terms of the size of the queue at the new price level or the last observed queue size depending on the direction of the price movement.

The distinction between buy and sell events serves to highlight times of significant mid-price movement as compared to standard fluctuations. In particular sequences of buy events of large size can be shown to cluster alongside upward mid-price changes and the opposite is true for sell events. These effects can be seen in figure 5.1 where a scatter-plot of buy and sell events is plotted alongside all the mid-price changes that occurred over the corresponding time interval. This clustering effect again justifies the use of Hawkes processes in modelling the inter-arrival times of these buy and sell events. Examining the empirical cumulative distribution functions of the inter-arrival times of these buy or sell events against various theoretical counterparts shows that their inter-arrival times follow arbitrary distributions, which is further justification for the use Hawkes processes. This is shown in figure 5.2 in which it can also be noted that the collection of buy and sell events for the same day can follow different distributions, and that the exponential distribution fails to capture the empirical distribution in each example case. This first approach to classifying buy and sell events using the order-book is limited by the fact that within our data we cannot differentiate between orders fills and cancellations. It is further limited by the fact that the price levels at which the level one queues sit can move frequently causing this signal to lose track of any particular order. So although tracking level one queue

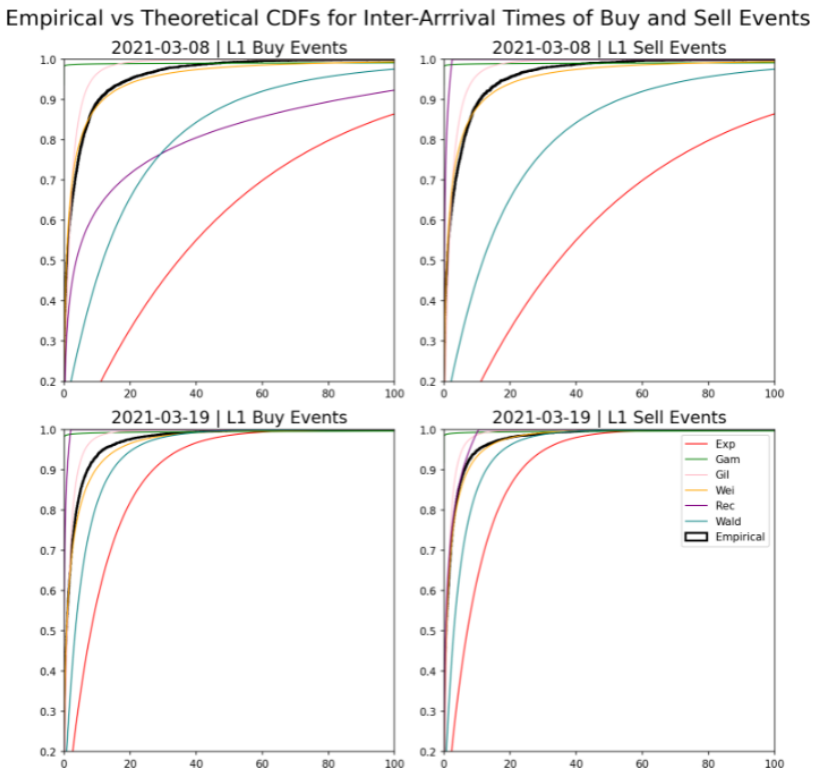


Figure 5.2: Empirical and theoretical cdf's for the inter-arrival times of level one order queue buy and sell events for two days of SB data.

activity can give a short preemptive signal of mid-price movement, it does not capture the full scope of the order flow within the limit order book. Given these ideas the second approach applied here to classifying buy and sell events is to monitor queues at all recently observed price levels within the data and focus on the cumulative change across all these levels. That is to say, given the order queues at twenty distinct price levels are observed, we look for instances of these fixed price levels in the next row of data and record any changes in a time congruent manner. The net buy flow is then defined as the summation of all observed changes in the currently tracked price levels that exert upward pressure on the mid-price and the net sell flow is defined similarly. Buy events are then defined as instances where the net buy flow exceeds some positive threshold and sell events are defined likewise for the net sell flow. This approach does not capture the full extent of the order flow for the underlying at a given time because the number of levels recorded in our data is finite. Also, when the mid-price moves the price levels move as well, this can cause the data to lose track of order queues on the outside levels of the order-book. In this case we are careful to not count these effects in the data as actual queue movement in defining buy and sell events. A breakdown of this approach can be found in appendix A.

The cumulative difference in net buy and sell flow is shown in figure 5.4 to be highly correlated with the mid-price over short time lags and to almost always follow the same trajectory. Thus the mid-price can be seen as relatively reactive to the net-order flow, but the time in which this net order flow is predictive of the mid-price was found not to

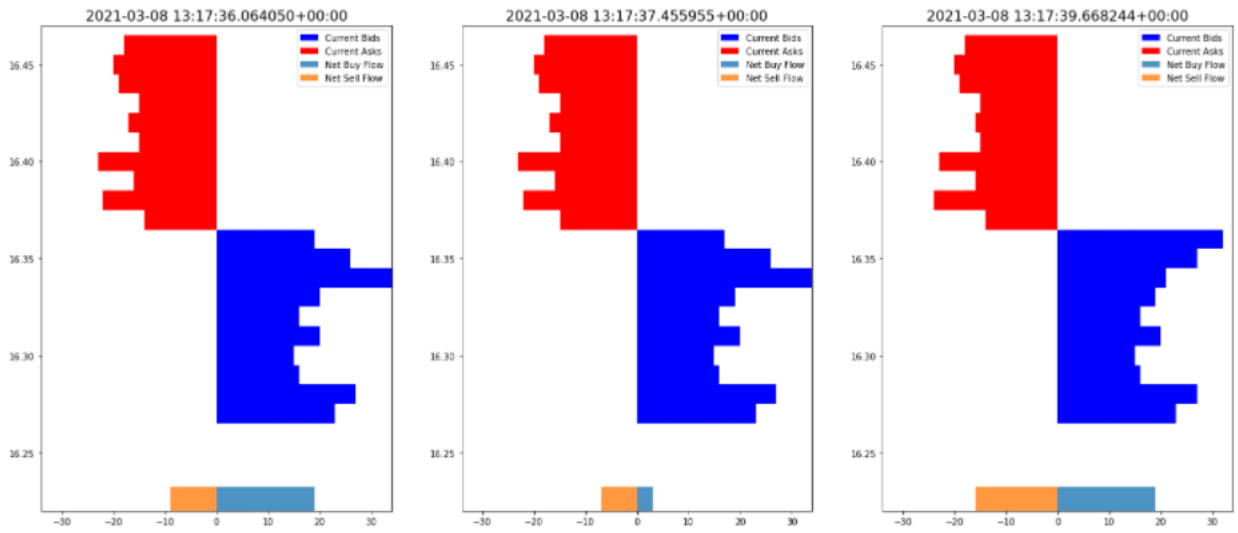


Figure 5.3: Three progressions of net buy and sell order flow based on observed changes in ten fixed price levels of the sugar order-book.

lend itself to implementing a viable trading signal by itself. This is due to the prospective number of trades this would incur and the associated slippage costs. Alternatively, we use the net order flow to define distinct buy and sell events and fit a Hawkes process to each of these respective series of events. We then use these Hawkes processes to develop a strategy based on the relationship between buy flow and sell flow intensity which limits the amount of prospective trades by focusing on times of locally extreme convergence or divergence in these two so called flow intensities.

5.3 Hawkes Intensity signal

Given a sequence of buy or sell events observed at times $\{t_1, t_2, ...t_k\}$ using the previously stated definition of the uni-variate Hawkes conditional intensity function given in remark 3.4, the following function is used to calculate the buy and sell intensity at a specified time t,

$$\lambda(t) = \lambda + \sum_{t_k<t} \alpha e^{-\beta(t_k - t)} \tag{5.1}$$

To maintain an emphasis towards an intra-day framework the parameters λ, α and β are fitted using the likelihood methods described in section 3.1 using a pre-determined amount of past data, typically not more than a couple of hours, and refitted on a set schedule. This allows the intensity function to be better fitted to distinct time regimes and recognize the different overall levels of volatility seen across different days. To ease computation times it was generally found that this function can be computed using a subset of the most recent events prior to the time t without comprising its expressiveness.

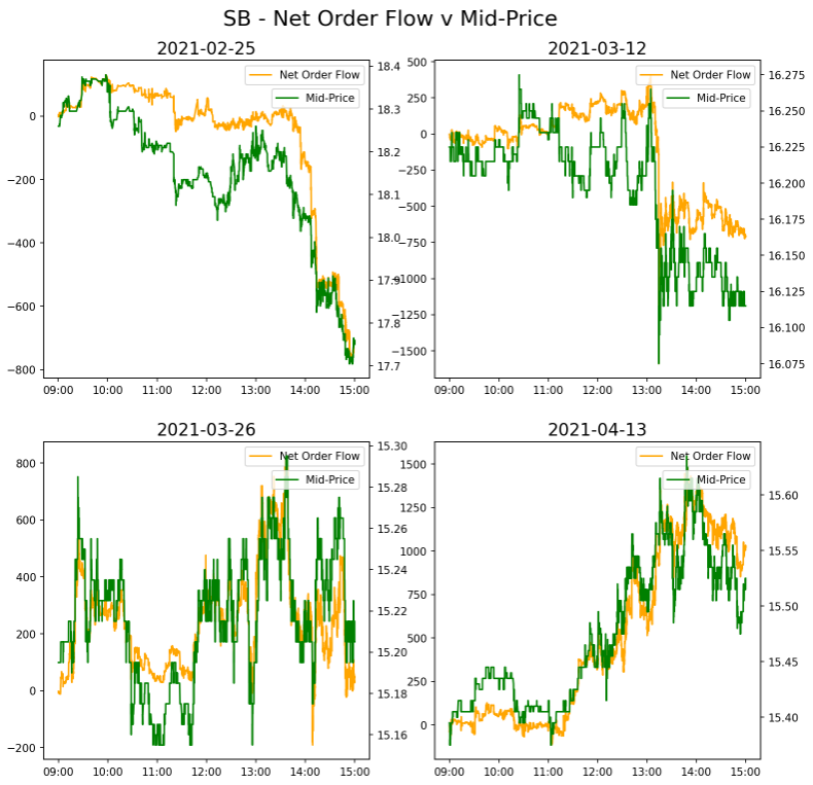


Figure 5.4: Comparison of Net order flow to Mid-price for four different days of sugar data.

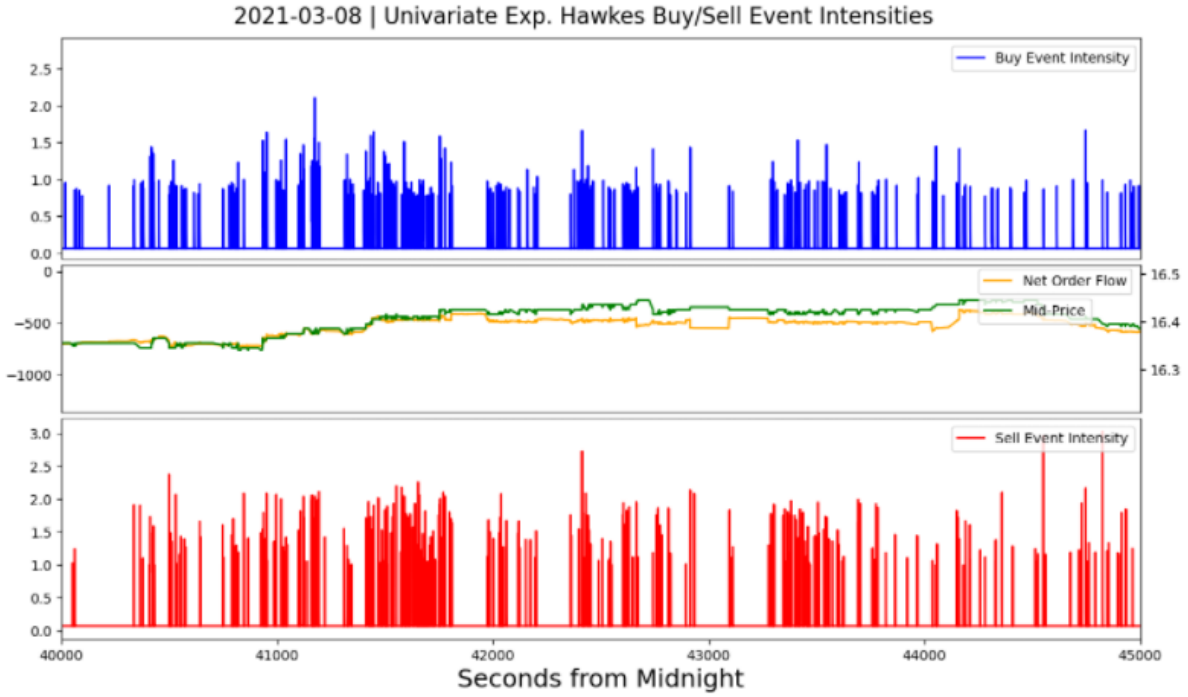


Figure 5.5: Mid-price, net order flow and uni-variate Hawkes conditional buy and sell intensity functions created using ten fixed price levels for one hour of sugar data.

It was found that the maximum likelihood estimates for the Hawkes parameters were usually such that any intrinsically generated spikes in intensity were short lived; an example of this is shown in figure 5.5. This was due to a typically high decay parameter β which negates most spikes in intensity before they can generate multiple new events. To smooth out the buy and sell intensity functions an exponentially weighted moving average computed recursively as follows was applied with a smoothing factor of $0 < \theta < 1$.

$$y_t = \lambda_0$$

$$y_{t+1} = (1 - \theta)y_t + \theta\lambda_t$$

This approach weights the most recent events the highest and can be used to provide a more visually smooth indicator market buy and sell activity. An example of these moving averages for one hour of sugar data is displayed in figure 5.6 in which it can be seen that the mid-price rises as the intensity difference increases, then levels out as the intensity difference approaches zero.

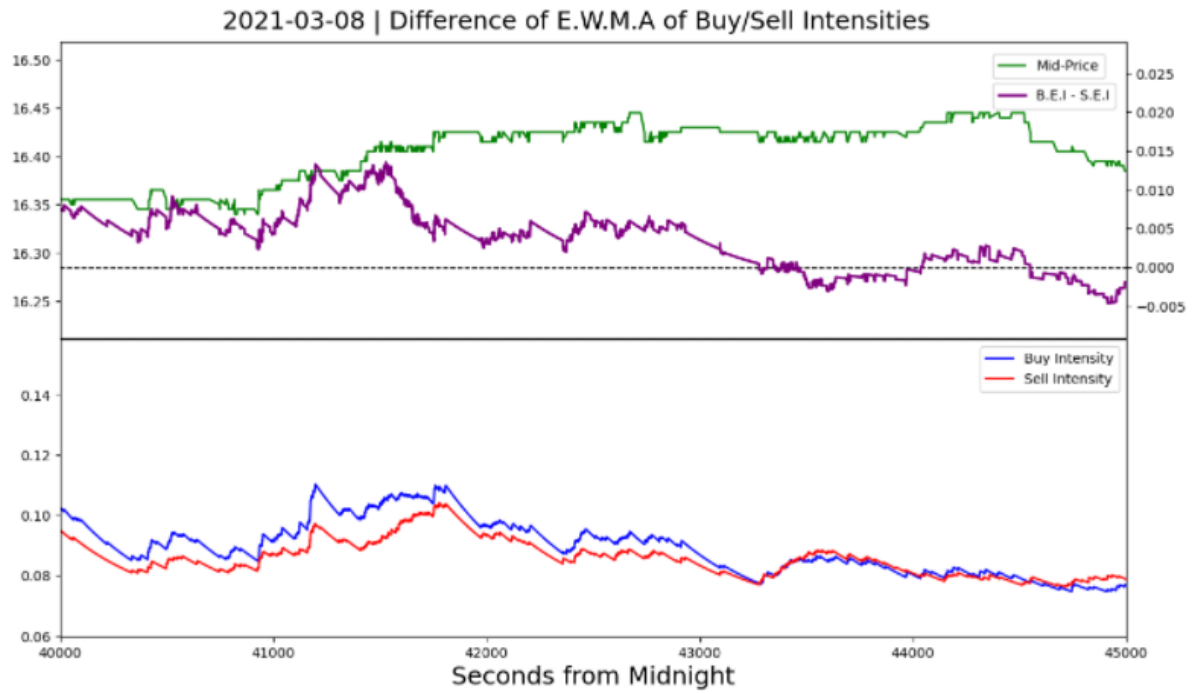


Figure 5.6: Exponentially weighted moving averages of buy and sell intensity created using order flow on ten fixed price levels for one hour of sugar data.

Looking closer at figure 5.6 it can be seen that over many short intervals the moving averages of buy and sell intensity are almost exactly correlated with each other. With this in mind focusing on the buy or sell intensity alone does not present an auspicious indicator to enter into a trade. This is further evidenced by the fact that it is not uncommon for a buy and a sell event to happen simultaneously. As such to create a more trade-able signal from these functions we

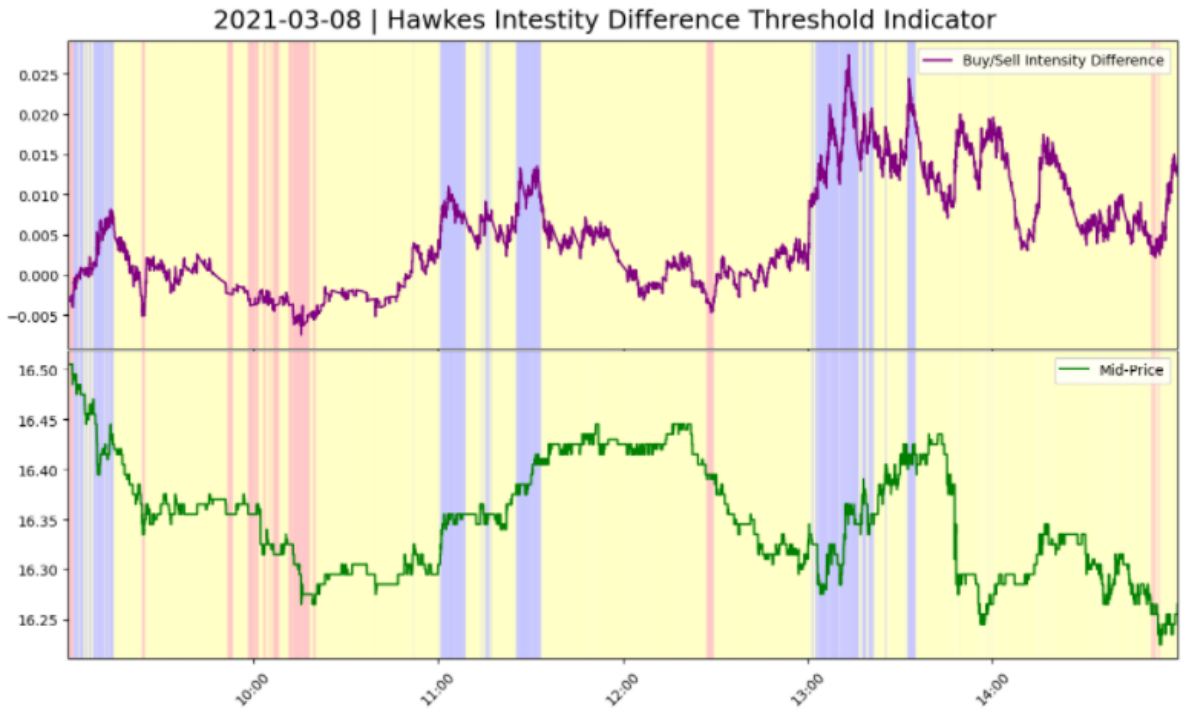


Figure 5.7: Hawkes intensity based directional indicator created using ten fixed price levels for one hour of sugar data. Upwards indicators are highlighted in blue, neutral in yellow and downwards in red.

instead look for intervals where the gap between them increases, decreases or in any way changes significantly relative to recent time. With the goal of implementing a simple back-testing strategy, it is then well considered to simplify down to a single directional indicator. To this end, we use extreme upper and lower quantiles of the difference between the moving averages of buy and sell intensity as thresholds for defining a three class indicator similar to as was defined in section 4.1. That is to say we monitor the relative high and low points of the intensity difference over a rolling time interval and assign indicators based on whether or not these upper and lower extremes are crossed. The upper and lower quantiles are then re-computed on a scheduled basis using incoming data with the purpose of effectively highlighting short term trends. This provides a means to define a three class directional indicator an example of which is shown in figure 5.7 as a contrasting background color coding to highlight the mid-price on a day of sugar data. Given the high frequency nature of our data it was not uncommon for hundreds or even thousands of oscillations in our indicator to occur over short periods of time resulting in a mixed signal that would be impossible to trade around without incurring considerable transaction costs. To alleviate this the thresholds for upwards or downwards quantiles in the intensity difference can be adjusted to create more consistent streams of common indicators. A good example of this adjustment is given in figure 5.7 which shows longer clusters of non-conflicting indicators that are often correlated with the desired mid-price movement. These clusters suggest clear placements for intra-day trades that reflect the idea of limited, opportunistic trading.

To test the feasibility of this signal a simple uni-directional trading strategy can be implemented using historical data. The following general rules were applied to orchestrate a trading strategy that is limited in how often it trades and only seeks to capitalize on the most extreme observed changes in the order-book:

- Enter a trade if a set number of non-conflicting indicators occur over an set interval of time.
- Hold a maximum of one position at a time.
- Exit a trade if the opposite indicator to the current position appears or the initial signal fails to reappear for an extended period of time.
- Once a trade is completed do not enter another for some fixed period of time.
- Do not trade the first hour of any given day.

The following algorithm details the strategy as a whole, in particular it describes the way in which the next recommended position is computed at some point in time given the sequence of indicators up to said point. The other parameters of this strategy assign number values to the previously stated conditions for entry and exit in accordance to the rate at which other function updates are computed.

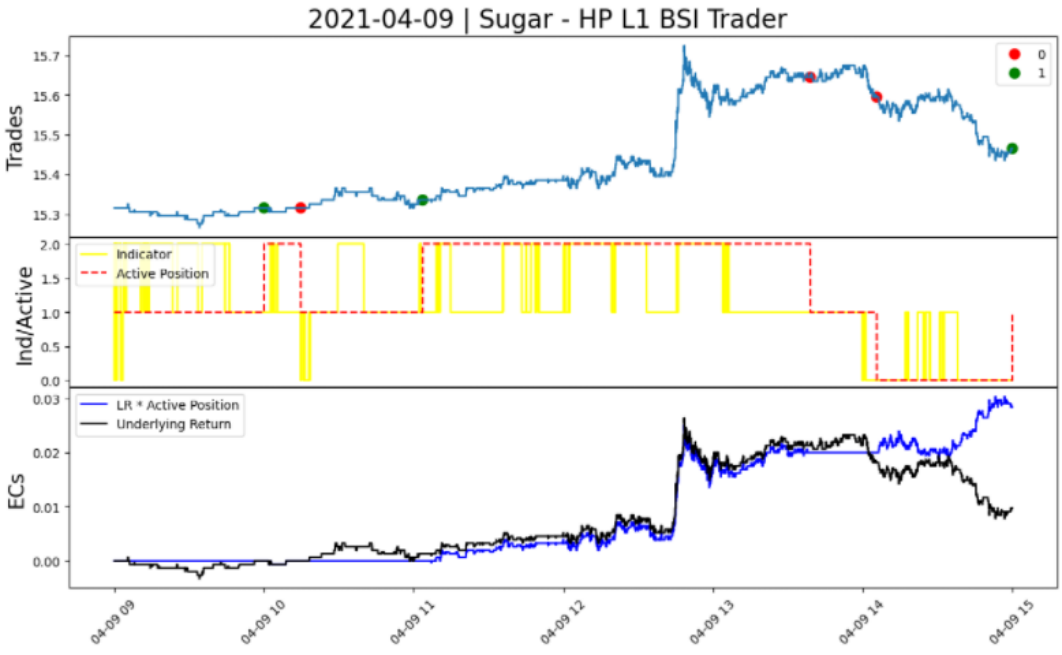


Figure 5.8: Example trading day for Hawkes intensity trader based on level one order queues.

For simplicity we assume that we only trade single lots and as such any impact on the order-book would be minimal. An example trading day is shown in figure 5.8 in which this strategy only traded threes times, but still captured periods of extended, consistent movement. This is in contrast to the actual indicator which oscillates rapidly

Algorithm 2 Hawkes Intensity Trading Algorithm

```
procedure TRADER(ind, refwindow, entrycs, exitcs, curpos, curhold, newholdsize)
    newpos ← curpos , curholdupdate ← 0
    if length of ind > 100 then
        prev = ind[−1 ∗ refwindow : ]
        if curpos = 0 then
            if |{prev | prev = −1}| <= 1 and |{prev | prevx = 1}| >= entrycs and curhold <= 0 then
                newpos = 1
            else if |{prev | prev = 1}| <= 1 and |{prev | prevx = −1}| >= entrycs and curhold <= 0 then
                newpos = −1
            else
                newpos = curpos
                curholdupdate = curhold − 1
            end if
        else
            if curpos = 1 then
                if |{prev | prev = 1}| <= exitcs or |{prev | prev = −1}| >= 1 then
                    newpos = 0
                    curholdupdate = newholdsize
                else
                    newpos = curpos
                end if
            else if curpos = −1 then
                if |{prev | prev = −1}| <= exitcs or |{prev | prev = 1}| >= 1 then
                    newpos = 0
                    curholdupdate = newholdsize
                else
                    newpos = curpos
                end if
            end if
        end if
    else
        newpos = curpos
        curholdupdate = curhold − 1
    end if
end procedure
```

between upwards (or downwards) and neutral indicators. The most apparent practical limitation of this strategy is the assumption that it trades immediately at the mid-price without delay. This is of course impossible as the mid-price is a factitious price, but given the context of historical data this assumption creates a straightforward framework for back-testing the viability of this strategy without having to simulate the order filling process. Another consideration is that this strategy assumes that any order would be filled immediately regardless of the current queue and any intervening price movement. Consequently, this strategy would likely be hindered by crossing the spread and other implicit transaction costs. As such, the following goals were set to give allowances for the aforementioned challenges present in real-time situations. Firstly, average over one tick gain per trade and secondly, trade less than ten times a day on average. Meeting these goals in out of sample testing provides an affirmation of the underlying theory and sets up future work on expanding the strategy to real time situations.

The trading strategy outlined throughout this section is dependent on several important parameters with effects on computation times, the amount data considered relevant at any given time, trade decisions and more. These parameters are as follows:

- NPE: number of previous events used in intensity functions.
- TPS: number of times per second the intensity functions are calculated.
- EST: minimum change in data considered an event for order flow model(event size threshold).
- Q: quantile for defining for defining extreme events.
- RM: length of running mean applied to intensity function.
- LBW: look-back windows for computing quantiles and Hawkes parameters.
- Ref Window: maximum number of past indicators considered in trade entry and exit decisions.
- Entry cs: number of non-conflicted indicators within the ref window required to enter a trade.
- Exit cs: number of indicators within the ref window required to stay in a trade.
- HLD: after trade holding period.

It was generally found that these parameters are best tuned to localized values that reflect recent market conditions present over a limited span of time typically not more than ten days. To optimize out of sample performance cross-validation was used in conjunction with a grid search of a range of manually curated values for each of the described parameters. It was also found that a rolling cross validation window preformed better on average than an expanding window.

5.4 Results

The results reported here are summary statistics emphasizing individual trade performance for the in sample and out of sample cross validation folds over the sugar data-set. Table 5.1 contains results for the level one queue strategy using a five day to five day train to test split and a day rolling window. Table 5.2 shows in sample results and table 5.3 shows out of sample results. Comparing the two there is a noticeable drop off in performance noted by an average profit per trade in sample to out of sample. In table 5.4 it can seen that generalization is a consistent challenge, but in spite of this the algorithm maintained an average of over one tick per trade across a couple hundred trades. The lack of generalization of the strategy can be attributed to over-fitting as the statistics used to optimize have a tendency to favour one-off successes over repetitious smaller trades. This could also be combated by using a larger grid of parameters at the cost of a proportional increase computational costs. Also of note is the range of the number of trades

Table 5.1: Cross validation results for Hawkes Process level one order queue trader

Fold	Mean Res	Med Res	PnL	Trades
0	0.03233	0.0050	0.485	15.0
1	0.03250	0.0325	0.065	2.0
2	0.03125	0.0000	0.125	4.0
3	0.02967	0.0150	0.445	15.0
4	0.02022	0.0000	0.465	23.0
5	0.01240	-0.0050	0.310	25.0
6	0.01413	0.0100	0.325	23.0
7	0.04500	-0.0100	0.225	5.0
8	0.06833	0.0200	0.615	9.0
9	0.03583	0.0000	0.215	6.0
10	0.03667	0.0250	0.220	6.0
11	0.04917	-0.0075	0.295	6.0
12	0.08833	0.0050	0.265	3.0
13	0.01286	-0.0075	0.180	14.0
14	0.00500	0.0000	0.015	3.0
15	0.07333	0.0550	0.440	6.0
16	0.07500	0.0750	0.150	2.0
17	0.21250	0.2125	0.425	2.0
18	0.01000	0.0050	0.220	22.0

Table 5.2: Level one algorithm in sample results.

Fold	Mean Res	Med Res	PnL	Trades
0	-0.00342	0.0050	-0.065	19.0
1	0.01583	0.0025	0.095	6.0
2	0.02917	-0.0075	0.175	6.0
3	0.00739	0.0000	0.170	23.0
4	-0.00500	-0.0150	-0.125	25.0
5	0.00310	-0.0100	0.065	21.0
6	0.01974	-0.0100	0.375	19.0
7	0.03045	0.0100	0.335	11.0
8	-0.02375	-0.0225	-0.095	4.0
9	-0.04056	-0.0100	-0.365	9.0
10	0.04917	-0.0075	0.295	6.0
11	0.02833	-0.0150	0.170	6.0
12	-0.02333	-0.0300	-0.070	3.0
13	-0.01667	-0.0100	-0.300	18.0
14	0.03000	0.0300	0.030	1.0
15	0.00364	0.0100	0.040	11.0
16	0.21250	0.2125	0.425	2.0
17	-0.00000	-0.0000	0.000	2.0
18	-0.00042	0.0000	-0.010	24.0

Table 5.3: Level one algorithm out of sample results.

	Avg Pnl/Fold	Avg PnL/Trade	Total PnL	Trades
In	0.28868	0.04655	5.485	191
Out	0.06026	0.01664	1.145	216

Table 5.4: Level one algorithm summary statistics.

made in different folds given each fold is reflective of a disjoint five day window. In some folds the algorithm would trade upwards of five times a day while in other folds the optimal parameters were found to give rise to situations where some days saw no suitable conditions to trigger trades. The design of this strategy in emphasising a search for

more certain trades also had the effect of prioritizing mean results over win rates. In fact there are situations where it lost the majority of trades, yet came out ahead due a couple of individual big winners.

The next results shown in table 5.5 are for the ten fixed level order flow strategy over the same data. In this case a ten day to five day train to test cross validation set-up showed improved performance statistics over a shorter training window as was used for the level one strategy. This causes overlapping in consecutive training windows, hence the summary statistics for total gains and number of trades were excluded in table 5.8. When compared to the level one strategy the fixed level strategy showed many of the same tendencies, but it also showed a tendency to trade more often which created more cumulative gains at the cost of average profits per trade. The end result of this was a less than one tick per trade average, but a higher cumulative profit as compared to the level one version of this strategy. The generalization gap was again present, although evidently smaller than the level one strategy due to the higher overall cumulative number of trades made.

Table 5.5: Cross validation results for Hawkes process fixed ten level order flow trader

Fold	Mean Res	Med Res	PnL	Trades
0	0.03658	0.0200	0.695	19.0
1	0.02813	0.0175	1.125	40.0
2	0.02971	0.0125	1.010	34.0
3	0.02906	0.0225	0.930	32.0
4	0.02487	0.0125	0.995	40.0
5	0.01800	0.0000	0.630	35.0
6	0.01696	-0.0100	0.475	28.0
7	0.03000	0.0150	1.020	34.0
8	0.04167	0.0200	1.125	27.0
9	0.01968	0.0100	0.610	31.0
10	0.01312	0.0050	0.420	32.0
11	0.01444	0.0100	0.520	36.0
12	0.00512	-0.0050	0.215	42.0
13	0.00548	0.0000	0.400	73.0
14	0.00983	0.0075	0.295	30.0
15	0.01145	-0.0025	0.435	38.0
16	0.01342	0.0150	0.510	38.0
17	0.02303	0.0100	0.760	33.0
18	0.01350	-0.0075	0.540	40.0

Table 5.6: Order flow algorithm in sample results.

Fold	Mean Res	Med Res	PnL	Trades
0	0.01474	0.0050	0.280	19.0
1	0.01133	-0.0050	0.170	15.0
2	0.03531	0.0175	0.565	16.0
3	-0.00107	0.0025	-0.015	14.0
4	0.01643	0.0000	0.345	21.0
5	0.00971	0.0050	0.165	17.0
6	0.05167	0.0300	0.620	12.0
7	0.03441	0.0100	0.585	17.0
8	0.00735	0.0000	0.125	17.0
9	0.00969	-0.0075	0.155	16.0
10	-0.02765	-0.0100	-0.470	17.0
11	-0.01087	-0.0100	-0.250	23.0
12	-0.00559	0.0000	-0.095	17.0
13	-0.00206	0.0000	-0.070	34.0
14	-0.00433	-0.0100	-0.065	15.0
15	-0.01118	-0.0100	-0.190	17.0
16	0.00553	0.0100	0.105	19.0
17	0.01000	0.0075	0.160	16.0
18	0.01583	0.0050	0.285	18.0

Table 5.7: Order flow algorithm out of sample results.

	Avg Pnl/Fold	Avg PnL/Trade	Total PnL	Trades
In	0.66895	0.02021	–	–
Out	0.12658	0.00838	2.405	340

Table 5.8: Order flow algorithm summary statistics.

The performance of both versions of the strategy was found to be heavily dependent on computationally expensive hyper-parameter optimization over which the range of values trialed was dictated by a manually curated grid search.

This cost would likely be prohibitive to the performance of any real-time adaptation, but also allots for potential improvements in performance metrics through further refinement of the standard ranges for localized parameters. In particular a broader range for time dependent parameters such as the length of moving averages used and the reference window for trade decisions could be more finely tuned to help with generalization. Overall the strategy met the initial goals of limiting the number trades and having a mean profit high enough to afford the luxury of losing the spread in all situations.

Chapter 6

Conclusions and Future Work

In this work we have demonstrated the predictive value of Hawkes processes through their direct application to high-frequency financial data. By applying existing theory the general compound Hawkes process was reconstructed and shown to be an apt model for previously untested limit order book data in section 3.2.3. Further, in section 4.2 an examination of several predictive methods for the general compound Hawkes process based on diffusive limit theorems was conducted and found that all the proposed methods produced similar if not exactly the same results. It was exceedingly rare for the different methods to produce opposing directional predictions and hence it can be concluded that each diffusive limit method is relatively interchangeable. On the other hand it was also shown in section 4.3 that a Monte-Carlo-esque approach to generating predictions can also produce positive results which differ from those produced by the diffusive limit methods. Overall, given fixed hyper-parameters, the classification statistics for the general compound Hawkes process were encouraging across whole data-sets, but we also showed that catering hyper-parameters to more localized values could improve performance statistics. In section 4.5 we then showed the viability of a simple trading strategy based on these predictive methods.

A practical application of the Hawkes intensity function was investigated in section 5 as a means to apply the order-flow within an order-book to a momentum based trading strategy. In section 5.2 two different methods for classifying order book activity into buy and sell events using differing amounts of data were contrasted, this highlighted the influence of the outer levels of the order-book vs. that of the most central level. In section 5.3 it was shown that the intensity functions for a series of these buy or sell events can be crafted into a directional indicator for the mid-price and that a viable trading strategy can be formed by aggregating many repeated indicators. It was shown in section 5.4 that the incorporation of more levels of the order-book into the defined trading strategy made it more generalizable resulting in higher cumulative profits at the cost of lower average profits per trade. The main challenge to this work in general was the computational cost of fitting models and subsequently running experiments across substantial amounts of

limit order book data. The challenge is compounded by the necessity of hyper-parameter tuning in producing positive results and suggests compromises when considering the real time application of this work.

Future work on this topic could entail any of the following modifications to the proposed experimental framework or areas of study in general.

- Investigating the use of other excitation functions in any instance of the Hawkes process such as the power law function could produce a better fit and subsequently better predictions.
- Investigating the use of variations of the Hawkes process such as the non-linear, Marked or State-dependent Hawkes processes in place of the linear exponential Hawkes process. In particular with a marked Hawkes process marks could be used to incorporate the size of buy and sell events into the level one order-flow strategy.
- A multivariate Hawkes process could be applied to fit the joint sequence of buy and sell events or the mid-price processes of correlated assets. The relationship between mutual excitation and self-excitation and theirs effects on the proposed strategies could then be explored.
- A further investigation of the effect of the size of the spread within an order-book and the size of the average mid-price change on the outlined model fitting process.
- Looking at other stochastic dynamics within an order-book such as the time to the next event or the renewal of depleted queues could help enhance any high-frequency strategy.

Appendix A

Buy and Sell Event Classification Algorithms

Algorithm 3 Algorithm for classifying and sizing level one order queue buy events

procedure L1 EVENT CLASSIFIER($newrow = nr,\ lastrow = lr$)
  if $nr(BP) - lr(BP) > 0$ **then** ▷ L1 ask price = BP, L1 bid size = BQ
    $BQM = |nr(BQ)|$ ▷ BQM = bid queue move
  else if $nr(BP) - lr(BP) < 0$ **then**
    $BQM = |lr(BQ)|$
  else if $nr(BP) - lr(BP) = 0$ **then**
    $BQM = nr(BQ) - lr(BQ)$
  end if

  if $nr(AP) - lr(AP) > 0$ **then** ▷ L1 ask price = BP, L1 ask size = AQ
    $AQM = |lr(AQ)|$ ▷ AQM = ask queue move
  else if $nr(AP) - lr(AP) < 0$ **then**
    $AQM = |nr(AQ)|$
  else if $nr(AP) - lr(AP) = 0$ **then**
    $AQM = nr(AQ) - lr(AQ)$
  end if

  if $BQM > 0 \mid AQM < 0$ **then**
    if $BQM > 0\ \&\ AQM < 0$ **then**
      $BSE = |BQM - AQM|$ ▷ BSE = Buy Event Size
    else if $BQM > 0\ \&\ AQM > 0$ **then**
      $BSE = BQM$
    else if $BQM < 0\ \&\ AQM < 0$ **then**
      $BSE = -AQM$
    end if
  end if

  if $BES > 0$ **then**
    $BuyEvent = True$
  else
    $BuyEvent = False$
  end if
end procedure

Algorithm 4 Algorithm for classifying buy and sell events based on order-flow

procedure ORDER FLOW EVENT CLASSIFIER($newrow = nr,\ lastrow = lr, \alpha$)
$Cls \leftarrow \{nr(BPs), nr(APs)\}\ ,\ Pls \leftarrow \{lr(BPs), lr(BPs)\}$
$CQs \leftarrow \{nr(BQs), nr(AQs)\}\ ,\ PQs \leftarrow \{lr(BQs), lr(BQs)\}$
$buysize \leftarrow 0\ \ ,\ \ sellsize \leftarrow 0$

for $l \in Cls$ **do** ▷ for each price level in new row
$cs = CQs(l)$
if $l \notin Pls$ **then**
if $\min Pls < l < \max Pls$ **then** ▷ if new levels appears in between old levels
$ls = 0$
end if
else
$ls = PQs(l)$
end if
$sc = cs - ls$
if $sc > 0$ **then**
$buysize+ = sc$
else if $sc < 0$ **then**
$sellsize- = sc$ ▷ ask sizes are assumed negative
end if
end for

$lo = Pls \setminus Cls$ ▷ looks for unused levels in previous row
for $l \in lo$ **do**
if $\min Cls < l < \max Cls$ **then**
$sc = -PQs(l)$
if $sc > 0$ **then**
$buysize+ = sc$
else if $sc < 0$ **then**
$sellsize- = sc$
end if
end if
end for

if $buysize > \alpha$ **then** ▷ α is threshold for size of order flow viewed as an event
$buyevent = True$
end if
if $sellsize > \alpha$ **then**
$sellevent = True$
end if
end procedure

Appendix B

GCHP Model Fitting Results for Stock Data

	Best Fit's	% of Time Best Fit	Mean Error Rate	Best Error Rate	Worst Error Rate
GCHPDO	0	0	N/A	N/A	N/A
GCHP2sDO	3	16.7	4.16	1.93	6.91
GCHP4DO	0	0	N/A	N/A	N/A
GCHPnSDO	15	83.3	16.21	1.79	39.91

Table B.1: Summary statistics for GCHP model fitting process over all AMZN data.

	Best Fit's	% of Time Best Fit	Mean Error Rate	Best Error Rate	Worst Error Rate
GCHPDO	0	0	N/A	N/A	N/A
GCHP2sDO	1	5.6	2.13	2.06	2.06
GCHP4DO	9	50.0	7.97	0.57	24.02
GCHPnSDO	8	44.4	10.66	0.36	25.06

Table B.2: Summary statistics for GCHP model fitting process over all EBAY data.

	Best Fit's	% of Time Best Fit	Mean Error Rate	Best Error Rate	Worst Error Rate
GCHPDO	0	0	N/A	N/A	N/A
GCHP2sDO	3	16.7	1.61	0.30	3.51
GCHP4DO	7	38.9	6.77	0.06	23.98
GCHPnSDO	8	44.4	6.78	2.14	17.23

Table B.3: Summary statistics for GCHP model fitting process over all FB data.

	Best Fit's	% of Time Best Fit	Mean Error Rate	Best Error Rate	Worst Error Rate
GCHPDO	2	11.1	23.055	22.45	23.66
GCHP2sDO	8	44.4	10.34	0.22	18.60
GCHP4DO	5	27.8	7.32	0.78	28.33
GCHPnSDO	3	16.7	7.12	1.07	10.64

Table B.4: Summary statistics for GCHP model fitting process over all MSFT data.

Bibliography

[1] F. Alexis and A. T. Ciprian. Modeling First Line Of An Order Book With Multivariate Marked Point Processes. Papers 1211.4157, arXiv.org, November 2012.

[2] E. Bacry, S. Delattre, M. Hoffmann, and J. Muzy. Some limit theorems for hawkes processes and application to financial statistics. *Stochastic Processes and their Applications*, 123(7):2475–2499, 2013.

[3] E. Bacry, I. Mastromatteo, and J. Muzy. Hawkes processes in finance. *Market Microstructure and Liquidity*, 01(01):1550005, 2015.

[4] E. Bacry and J. Muzy. Hawkes model for price and trades high-frequency dynamics. *Quantitative Finance*, 14(7):1147–1166, 2014.

[5] A. Berchtold and A. Raftery. The Mixture Transition Distribution Model for High-Order Markov Chains and Non-Gaussian Time Series. *Statistical Science*, 17(3):328 – 356, 2002.

[6] Y. Bilodeau. Deep limit order book events dynamics. *SSRN Electronic Journal*, 2020.

[7] T. Bjork. *Arbitrage Theory in Continuous Time*. Oxford University Press, 3 edition, 2009.

[8] F. Black. Equilibrium exchanges. *Financial Analysts Journal*, 51(3):23–29, 1995.

[9] D. Boos and L. A. Stefanski. *Essential statistical inference theory and methods*. Springer, 2013.

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

[11] Á. Cartea, S. Jaimungal, and J. Penalva. *Algorithmic and High-Frequency Trading*. Cambridge University Pres., 2015.

[12] V Chatalbashev, Y Liang, A Officer, and N Trichaki. Stanford University MS&E 444 group project submission, 2007. retrieved on 1 Aug 2021. http://users.iems.northwestern.edu/ armbruster/2007msande444/report1a.pdf,.

[13] R. Cont, S. Stoikov, and R. Talreja. A stochastic model for order book dynamics. *Operations Research*, 58(3):549–563, 2010.

[14] J. Da-Fonseca and R. Zaatour. Hawkes process: Fast calibration, application to trade clustering and diffusive limit. *SSRN Electronic Journal*, 2013.

[15] D. J. Daley and D. Vere-Jones. *An introduction to the theory of Point Processes*. Springer, 2003.

[16] D. J. Daley and D. Vere-Jones. *An introduction to the theory of point processes. Vol. II*. Probability and its Applications (New York). Springer, New York, second edition, 2008. General theory and structure.

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

[18] L. Glosten. Is the electronic open limit order book inevitable? *The Journal of Finance*, 49(4):1127–1161, 1994.

[19] M. Gould, M. Porter, S. Williams, D. McDonald, D. Fenn, and S. Howison. Limit order books. *Quantitative Finance*, 13(11):1709–1742, 2013.

[20] Q. Guo, B. Remillard, and A. Swishchuk. Multivariate general compound point processes in limit order books. *Risks*, 8(3), 2020.

[21] S. J. Hardiman, N. Bercot, and J. Bouchaud. Critical reflexivity in financial markets: a hawkes process analysis. *The European Physical Journal B*, 86(10), Oct 2013.

[22] J. Hasbrouck. Measuring the information content of stock trades. *The Journal of Finance*, 46(1):179–207, 1991.

[23] A. G. Hawkes. Point spectra of some mutually exciting point processes. *Journal of the Royal Statistical Society. Series B (Methodological)*, 33(3):438–443, 1971.

[24] Q. He and A. Swishchuk. Quantitative and comparative analyses of limit order books with general compound hawkes processes. *Risks*, 7(4), 2019.

[25] M. Höschler. Limit order book models and optimal trading strategies. Ph.D. thesis Technical University of Berlin, 2011.

[26] J. F. C. Kingman. *Poisson Processes.* Number Vol. 3 in Oxford Studies in Probability. Clarendon Press, 1993.

[27] S.G. Kou. Chapter 2 jump-diffusion models for asset pricing in financial engineering. *Financial Engineering*, page 73–116, 2007.

[28] P. J. Laub, T. Taimre, and P. K. Pollett. Hawkes processes, 2015. *aarXiv*. arXiv:1507.02822.

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

[30] M. Morariu-Patrichi and M. S. Pakkanen. State-dependent hawkes processes and their application to limit order book modelling. *Quantitative Finance*, 22(3):563–583, 2021.

[31] N. Nadagouda and M. A. Davenport. Switched hawkes processes. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[32] A. Ntakaris, M. Magris, J. Kanniainen, M. Gabbouj, and A. Iosifidis. Benchmark dataset for mid-price forecasting of limit order book data with machine learning methods. *Journal of Forecasting*, 37(8):852–866, Aug 2018.

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

[34] Y. Ogata. On lewis' simulation method for point processes. *IEEE Transactions on Information Theory*, 27(1):23–31, 1981.

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

[36] T. Ozaki. Maximum likelihood estimation of hawkes' self-exciting point processes. *Annals of the Institute of Statistical Mathematics*, 31:145–155, 1979.

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

[38] M. Sjogren and T. DeLise. General compound hawkes processes for mid-price prediction, 2021. *aarXiv*. arXiv:2110.07075.

[39] J.M. Steele. *Stochastic Calculus and Financial Applications*. Stochastic Modelling and Applied Probability. Springer New York, 2012.

[40] A Swishchuk. General compound hawkes processes in limit order books. *SSRN Electronic Journal*, 2017.

[41] A. Swishchuk. Merton investment problems in finance and insurance for the hawkes-based models. *Risks*, 9(6), 2021.

[42] A Swishchuk. Modelling of limit order books by general compound hawkes processes with implementations. *Methodology and Computing in Applied Probability*, 23(1):399–428, 2021.

[43] A. Swishchuk. Stochastic modelling of big data in finance. *Methodology and Computing in Applied Probability*, 22(4):1613–1630, 2021.

[44] A. Swishchuk, T. Hofmeister, J. Schmidt, and K. Cera. General Semi-Markov Model for Limit Order Books: Theory, Implementation and Numerics. *International Journal of Theoretical and Applied Finance*, 20(3):1–21, 2016.

[45] A. Swishchuk and A. Huffman. General compound hawkes processes in limit order books. *Risks*, 8(1), 2020.

[46] A. Swishchuk and N. Vadori. Semi-Markovian modelling of limit order markets. *SIAM J. on Financial Mathematics*, 8(1):240–273, 2017.

[47] I. M. Toke and N. Yoshida. Marked point processes and intensity ratios for limit order book modeling. *Japanese Journal of Statistics and Data Science*, 5(1):1–39, 2022.

[48] S. Vyetrenko, D. Byrd, N. Petosa, M. Mahfouz, D. Dervovic, M. Veloso, and T. Balch. Get real: Realism metrics for robust limit order book market simulations. *Proceedings of the First ACM International Conference on AI in Finance*, 2020.

[49] X Zhang. *Empirical Risk Minimization*, pages 312–312. Springer US, Boston, MA, 2010.

[50] Z. Zhang, S. Zohren, and S. Roberts. Deeplob: Deep convolutional neural networks for limit order books. *IEEE Transactions on Signal Processing*, 67(11):3001–3012, Jun 2019.