UNIVERSITY COLLEGE LONDON

Department of Computer Science

Simulation of a Limit Order Book and Application to Optimal Placement

Author:
Takudzwa MTOMBENI

Academic Supervisor:
Dr. Paris PENNESI

Industry Supervisors:
Dr. Alexander BARZYKIN
Dr. James NEWBURY

This dissertation is submitted as partial requirement for the MSc Computational Finance degree at UCL. It is substantially the result of my own work except where explicitly indicated in the text. The dissertation may be freely copied and distributed provided the source is explicitly acknowledged

September 10, 2020

UNIVERSITY COLLEGE LONDON

Abstract

Faculty of Engineering Sciences
Department of Computer Science

Master of Science

Simulation of a Limit Order Book and Application to Optimal Placement

by Takudzwa MTOMBENI

The ubiquity of the order-driven market as the trading mechanism of choice for modern financial markets has generated significant academic and industrial interest in the modelling of limit order books. The majority of prior work has been focused on producing analytically tractable models in order to allow for the computation of various quantities of interest from the limit order book. This however means that certain simplifying assumptions must be made that may hinder the usefulness of the models in certain applications. The aim of this thesis is to implement a computationally efficient synthetic limit order book simulator capable of reproducing the empirical facts observed in real order-driven markets. We provide a flexible framework which can be used in the testing of optimal execution, optimal placement and market making strategies. We model the most granular form of the limit order book, where prices and volumes are discrete random variables, such a model is said to be a microscopic limit order book model. As an illustration of its use case we use the simulator to test an optimal placement strategy.

Contents

List of Figures

List of Tables

Acknowledgements

I would like to thank Dr. James Newbury and Dr. Alexander Barzykin for the guidance and support they offered during the writing of this thesis. I would also like to thank Dr. Paris Pennesi and Dr. Simone Righi whose market microstructure lectures inspired my interest in the topic.

Chapter 1

Introduction

1.1 Order-Driven Markets

The majority of modern financial markets now operate as order-driven markets. In such markets, each market participant may provide liquidity to the market by submitting limit orders or take liquidity from the market through the submission of market orders. Limit orders that have been submitted to the market and have yet to be executed are stored on the limit order book (LOB) and are referred to as active orders. Active orders may be cancelled at the discretion of the party who submitted the order.

1.1.1 Mathematical Description

Definition 1 (Order) *Let the the 4-tuple,*

$$z = (\epsilon_z, p_z, \nu_z, t_z)$$

define an order submitted at time t_z to buy ($\epsilon_z = 1$) or sell ($\epsilon_z = -1$) ν_z units of an asset for a price no greater than p_z (in the buy case) or less than p_z (in the sell case)

We refer to ϵ_z as the direction, p_z as the price, ν_z as the volume and t_z as the arrival time. Using this notation a buy market order is given by $z = (1, \infty, \nu_z, t_z)$ and a sell market order by $z = (-1, 0, \nu_z, t_z)$.

Definition 2 (Limit Order Book) *We define the limit order book $\mathcal{L}(t)$ as the set of all active orders for a particular asset at time t.*

We use the convention that although the components of an order z are known at time t_z, order z is not present in the LOB at t_z but immediately after t_z. i.e. $z \notin \mathcal{L}(t_z)$. Throughout this thesis we assume an underlying filtered probability space $(\Omega, \mathcal{F}, \mathbb{P}, \mathbb{F})$ where the filtration $\mathbb{F} = \{\mathcal{F}_t\}_{t \geq 0}$ is such that the limit order book process $\mathcal{L}(t)$ and an order $z = (\epsilon_z, p_z, \nu_z, t)$, placed at time t, are $\mathcal{F}_t -$ measurable.

Definition 3 (Tick Size) *The tick size τ of a LOB is the smallest permissible difference between different price levels on the LOB, i.e. for all orders z we must have that,*

$$p_z = k\tau \text{ for some } k \in \mathbb{N}$$

Definition 4 (Lot Size) *The lot size ω of a LOB is the smallest tradeable unit of the asset on the LOB, i.e. for all orders z we must have that,*

$$\nu_z = k\omega \text{ for some } k \in \mathbb{N}$$

The following definitions are for various quantities of interest that can be derived from the LOB and are commonly used in the literature.

Definition 5 (Available Depth) *The available depth at price level p at time t is the total volume of the limit orders residing at price level p at time t,*

$$V(p,t) = \sum_{z \in \mathcal{L}(t) | p_z = p} \nu_z$$

Definition 6 (best-bid) *The best-bid at time t is the highest price among buy limit orders on the LOB at time t.*

$$b(t) = \max_{z \in \mathcal{L}(t) | \epsilon_z = 1} p_z$$

Definition 7 (best-ask) *The best-ask at time t is the lowest price among sell limit orders on the LOB at time t.*

$$a(t) = \min_{z \in \mathcal{L}(t) | \epsilon_z = -1} p_z$$

If $a(t) \leq b(t)$ then there are buy and sell limit orders that can be immediately executed and removed from the LOB; thus we always have $a(t) > b(t)$. The orders queued at the best bid and ask are often referred to as *'the top of the book'*.

Definition 8 (best-ask spread) *The spread at time t is the difference between the best-ask and best-bid at time t.*

$$s(t) = a(t) - b(t)$$

The spread is often used as a crude measure of a market's liquidity. A narrow spread corresponds to a more liquid market and a wide spread corresponds to a more illiquid market. The spread can also be thought of as the cost of obtaining immediate execution.

Definition 9 (mid-price) *The mid-price at time t is an average of the best-bid and best-ask at time t.*

$$M(t) = \frac{1}{2}(a(t) + b(t))$$

Definition 10 (log-return) *From the mid-price we can define the log return at time t of an asset over a time period $\tau > 0$ as*

$$r_\tau(t) = \log(M(t+\tau)) - \log(M(t))$$

For the purpose of brevity, when referring to the log-returns we will sometimes drop the 'log' and simply refer to them as the returns.

Definition 11 (volatility) *By definition $r_\tau(t)$ is not $\mathcal{F}_t$-measurable (i.e. the return is a random variable at time t). We can therefore define the standard deviation of $r_\tau(t)$ as follows,*

$$\sigma_\tau(t) = \sqrt{\mathbb{E}[r_\tau(t)^2|\mathcal{F}_t] - (\mathbb{E}[r_\tau(t)|\mathcal{F}_t])^2}$$

We refer to the standard deviation of $r_\tau(t)$ as the τ period volatility of the asset at time t.

1.1.2 Trading in an Order-Driven Market

When a buy (respectively, sell) limit order z is submitted, the markets trade matching algorithm checks whether it is possible to match the incoming order to an existing sell (respectively, buy) limit order. A matching is possible if $p_z \geq a(t)$ (respectively, $p_z \leq b(t)$). It may be the case that only a partial matching is possible, this occurs when there is insufficient share volume at prices that may be executed by the incoming order. In this case the order is partially filled and the remaining portion becomes an active order on the LOB. If no matching is possible, the entire order becomes an active order on the LOB. When a market order is submitted, it executes immediately, matching with active orders according to the market's priority rules. The most common priority rule is price-time priority, under this rule the market order matches with active orders from best price to worst price, and matches orders at the same price in order of submission time (from earliest to latest)(Bouchaud et al., 2018).

In this thesis, we assume that the market operates under price-time priority. Each price level in the LOB is effectively a first-in-first-out queue. Active orders reside on the LOB until they are either matched, cancelled or expire. Figure 1.1 shows an illustration of the LOB.

1.1.3 Early History of Limit Order Book Modelling

Academic efforts at modelling the limit order book can be traced back to economic work on the modelling of auctions, see (Engelbrecht-Wiggans, 1980) for a comprehensive survey of early work on this topic. (Mendelson, 1982) presents a statistical model of a 'market clearing house'. Although no notion of a limit order book was present at the time, this trading mechanism is a periodic double auction and approximates the continuous double auction used in modern order-driven markets. (Madhavan, 1992) offers an analysis of price formation in a continuous double auction, a periodic double auction and a traditional quote driven market in the context of stock exchanges. The paper models trading using a game theoretic approach and its aim is to compare the efficiency of prices produced by each market mechanism. In (Parlour, 1998) the limit order book, as found in modern stock exchanges, is modelled explicitly. The modelling approach is economic unlike the statistical approach favoured in recent literature.

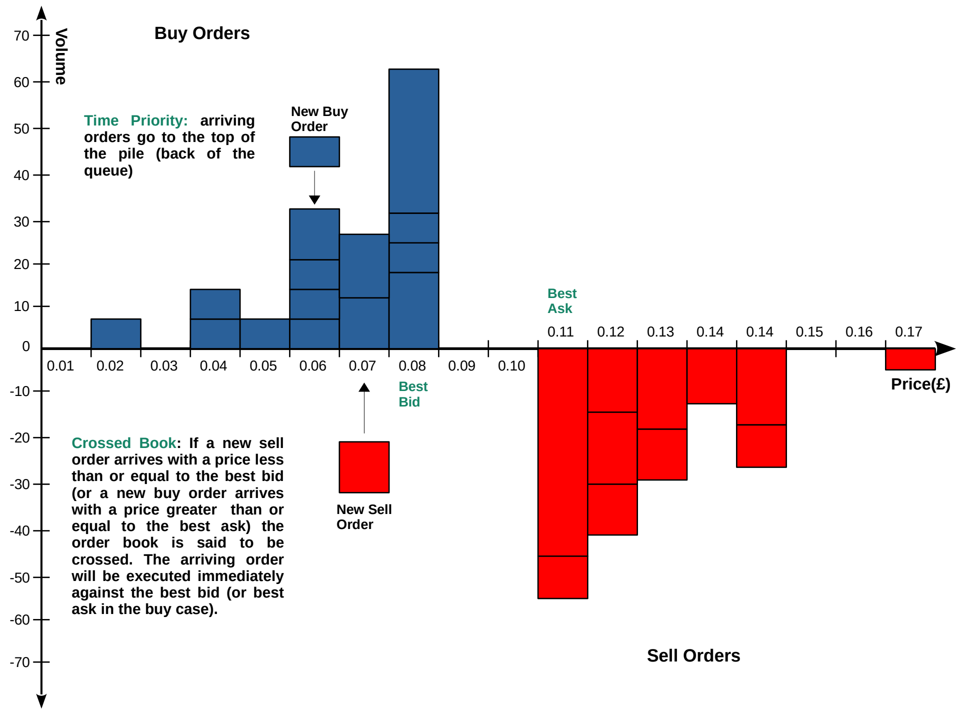


FIGURE 1.1: Limit Order Book Illustration
For illustrative purposes we depict sell limit orders as having negative volume and buy limit orders as having positive volume.

For a detailed review of recent literature, see chapter 2.

1.1.4 Motivations for Modelling the Limit Order Book

General Motivations for Limit Order Book Modelling Various academic and industrial applications exist for limit order book models. They may be used to study the dynamics of the order book and infer relationships between properties of the order-flow, liquidity and the observed price dynamics. Models may also be used to perform short-term price predictions. They are used in the derivation of optimal execution and optimal placement strategies. The application we focus on in this thesis is their use as a testing ground for various high frequency trading strategies.

Thesis Motivations Using limit order book simulation, we can construct a framework that can be used to test and compare optimal execution strategies, optimal placement strategies, market making strategies and other high frequency trading strategies. Simulation can be used in conjunction with paper trading which reproduces market conditions exactly but, unlike simulation, does not account for the market impact of the strategy being tested. Additionally, simulations can be re-run as often as is computationally feasible, allowing us to construct and compare the distributions of the costs incurred/profits earned by various strategies. Simulation also offers us the ability to adjust model parameters in order to examine the behaviour of strategies under different market conditions.

1.2 Thesis Structure and Summary of Contributions

This thesis is structured as follows: chapter 2 offers a review of recent literature on limit order book modelling. In chapter 3 we give a description of the our proposed model; chapter 4 describes the parameter estimation procedures used to calibrate our model. In chapter 5 we provide an empirical analysis of our simulated limit order book and contrast our results to empirical studies of real limit order book data. In chapter 6 we explore the effects of the model parameters on the volatility in the simulated limit order book. In chapter 7 we show how limit order book simulation can be used to test optimal order placement strategies by using our simulator to test the optimal placement solution presented in Cont and Kukanov, 2017 and chapter 8 concludes.

Contributions We have proposed a computationally tractable framework for use in the testing of optimal execution and placement strategies. Our framework lends itself well to customisation through its modular approach to model design. Within the context of this framework, we have presented a model that reasonably reproduces important stylised facts of financial time series. We have shown market impact within the simulated market and the effect of the model parameters on simulated volatility. Finally, we used the model to test an optimal placement strategy.

Chapter 2

Literature Review

Recent literature on limit order book modelling can generally be placed into one of two categories: agent based modelling or zero-intelligence modelling. In Agent based models, the approach is to divide market participants into several categories and model their behaviour separately. In such models, the order flow of the market is a convolution of the order flow generated by each category of market participant. This is in contrast to zero-intelligence modelling where the market order flow is modelled directly using stochastic methods. The zero-intelligence approach avoids making behavioural assumptions about market participants which often rely on unobservable parameters. Below we provide a selection of literature from both approaches.

2.1 Agent Based Modelling Literature

Parlour, 1998 provides a probabilistic agent based model of a limit order book. In the model, the asset possesses a fundamental value and agents are differentiated based on their subjective valuation of the asset. Some key assumptions of the model are that order sizes are constant; the market consists of only two prices, a bid price and an ask price, and agents arrive only once on the market (orders cannot be cancelled and the same trader cannot submit more than one order). Foucault, Kadan, and Kandel, 2005 present a model which incorporates expected time till execution into the trading behaviour of the agents. Their model is similar to Parlour, 1998 in that agents arrive only once on the market and order sizes are constant. Foucault, Kadan, and Kandel, 2005 however, present a model that allows for a range of limit order prices, though this range is bounded. Other key assumptions are that the arrival of buyers and sellers alternate with certainty and limit orders must narrow the spread by at-least 1 tick. Roşu, 2009 presents a model which, like Foucault, Kadan, and Kandel, 2005, incorporates waiting costs into agent behaviour but also allows agents to cancel their orders. Agents are differentiated by their patience and the tick size is assumed to be zero.

2.2 Zero-intelligence Modelling Literature

Some early literature in which the 'zero-intelligence' approach is used are, Bouchaud, Mézard, Potters, et al., 2002 where the model is inspired by empirical observations presented in the paper. The limit order prices are distributed as a power law around the best price. Smith et al., 2003 and Daniels et al., 2003 where order arrival times and cancellations times are modelled as Poisson processes. The order size is assumed constant (a random order size - half normally distributed - is also tested in simulation), order prices and order cancellations are distributed uniformly. Farmer, Patelli, and Zovko, 2005 present a model they term 'minimally intelligent'; this model has two types of agents with differing behaviour similar to the agent-based modelling approach, but these agents place trades at random as in the zero-intelligence approach.

More recently Cont, Stoikov, and Talreja, 2010 propose a Markovian model with constant order sizes where event arrivals are Poisson processes. They make use of the Laplace transform to compute probabilities for various events of interest conditional on the state of the LOB. Cont and De Larrard, 2012 where only the dynamics at the best bid and best ask are modelled, this is a so called reduced form model. Cont and Mueller, 2019 and Hambly, Kalsi, and Newbury, 2020 where the dynamics of the limit order book are modelled using stochastic partial differential equations.

Prior work places a heavy emphasis on analytical tractability, as the goal is to compute the probabilities of various events of interest. Our goal is to provide a framework that can be used to test optimal execution/placement algorithms and other high frequency trading algorithms, thus we focus on computational tractability and the reproduction of well known properties of LOB dynamics.

Chapter 3

Model Specification

Given the definition of an order as

$$z = (\epsilon_z, p_z, \nu_z, t_z)$$

our approach is to model each order type and each component of the order tuple separately. This modular approach allows us to produce complex models of the limit order book by specifying simple tractable models on the order components. This modularity also allows for models produced under this framework to be easily modified. This is a desirable property as one may slightly alter the configuration of a given model for different practical use cases.

We use the convention that although the components of an order z are known at time t_z, order z is not present in the LOB at t_z but immediately after t_z. i.e. $z \notin \mathcal{L}(t_z)$. We assume an underlying filtered probability space $(\Omega, \mathcal{F}, \mathbb{P}, \mathbb{F})$ where the filtration $\mathbb{F} = \{\mathcal{F}_t\}_{t\geq 0}$ is such that the limit order book process $\mathcal{L}(t)$ and an order $z = (\epsilon_z, p_z, \nu_z, t)$, placed at time t, are $\mathcal{F}_t -$ measurable.

3.1 Event Times

Definition 12 (Counting Process) *A stochastic process $\{N(t), t \geq 0\})$ is a counting process if it satisfies the following,*

(C1) $N(t) \geq 0 \ \forall t \geq 0$

(C2) $N(t) \in \mathbb{Z} \ \forall t \geq 0$

(C3) $N(s) \leq N(t) \ \forall\, 0 \leq s \leq t$

Definition 13 (Mutually exciting Hawkes process) *Let $\boldsymbol{N}(t) = \{N_i(t)\}_{i=1}^{n}$ be a collection of counting processes. Let $\{T_{ij} | i \in \{1, \cdots, n\}, j \in \mathbb{N}\}$ be the random arrival times associated with each process and t_{ij} the observed arrival times. A mutually exciting Hawkes process is a process $\boldsymbol{N}(t)$ such that for each $i \in \{1, \cdots, n\}$, $N_i(t)$ has a conditional intensity* *of the form,*

$$\lambda_i(t) = \mu_i + \sum_{j=1}^{n} \int_{-\infty}^{t} g_j(t-u) dN_j(u) \tag{3.1}$$

$\mu_i > 0$ *is termed the baseline intensity and* $g_j : (0, \infty) \to [0, \infty)$ *the excitation function.*

Let $[0, T]$ be the time interval on which the simulation takes place. We model the arrival times of market orders, limit order submissions and limit order cancellations using three mutually exciting Hawkes process with conditional intensities given by,

$$\forall i \in [1,2,3], \quad \lambda_i(t) = \mu_i + \sum_{j=1}^{3} \sum_{t_k^j < t} \phi_{ij}(t - t_k^j) \tag{3.2}$$

$$\phi_{ij}(t) = \alpha_{ij} \gamma \exp(-\gamma t) \tag{3.3}$$

The above intensity functions correspond to choosing exponential decay as the excitation functions in 3.1 (Laub, Taimre, and Pollett, 2015).

μ_i are the baseline intensities. $\{\alpha_{ij}\}$ are the adjacencies. γ is the decay. t_k^j are the arrival times of events for process j. The baseline intensities and adjacencies are estimated from the data, we fix the decay pre-estimation.

Estimation and simulation of the event arrival processes is done using the `tick` package in python (Bacry et al., 2017).

3.2 Market Orders

Definition 14 (Poisson Distribution) *A random variable X is said to have the Poisson distribution with rate* $\lambda > 0$ *if its probability mass function is given by,*

$$\mathbb{P}[X = x] = \begin{cases} \frac{\lambda^x \exp(-\lambda)}{x!} & x \in \{0, 1, 2, \cdots\} \\ 0 \textit{ otherwise} & \end{cases}$$

We write $X \sim \mathcal{P}(\lambda)$

Let $m_1, m_2, \cdots, m_{N_m}$ be the sequence of simulated market orders. Where

$$m_i = (\epsilon_{m_i}, \nu_{m_i}, t_{m_i})$$

we omit the price in our notation since for market orders this can be inferred from the direction. Given market order arrival times, to simulate a market order m we need only generate the direction ϵ_m and a volume ν_m.

Directions It is a well known empirical fact that market order directions are positively autocorrelated and that the autocorrelations decay slowly in time (Bouchaud et al., 2018). To reproduce the positive autocorrelation property, we model the market order trade sign process as a Markov chain with the following transition probabilities,

$$\mathbb{P}[\epsilon_{m_i} = 1 | \epsilon_{m_{i-1}} = 1] = \pi_{11}, \quad \forall i \in \{2, 3, \cdots, N_m\} \tag{3.4}$$

$$\mathbb{P}[\epsilon_{m_i} = -1 | \epsilon_{m_{i-1}} = -1] = \pi_{00}, \quad \forall i \in \{2, 3, \cdots, N_m\} \tag{3.5}$$

$$\mathbb{P}[\epsilon_{m_0} = 1] = \frac{1}{2} \tag{3.6}$$

In this thesis we set $\pi = \pi_{11} = \pi_{00}$ as we observed very little difference between the estimates of π_{11}, π_{00} in our available data.

Volume Empirical evidence suggests that the size of market orders is related to the available depths at the top of the book (Farmer et al., 2004). This phenomenon can be explained by selective liquidity taking, large market orders are submitted when available liquidity is high and small market orders are submitted when available liquidity is low (Bouchaud, Farmer, and Lillo, 2009). We thus model market order volume as follows, given the lot size ω

$$\begin{aligned} v_{m_i} &= (X_{m_i} + 1)\omega \\ X_{m_i} &\sim \mathcal{P}(\eta_{m_i}) \\ \text{where, } \eta_{m_i} &= \begin{cases} \beta\, V(a(t_{m_i}), t_{m_i}) & \text{for } \epsilon_{m_i} = 1 \\ \beta\, V(b(t_{m_i}), t_{m_i}) & \text{for } \epsilon_{m_i} = -1 \end{cases} \text{, and } \beta > 0 \end{aligned} \tag{3.7}$$

$V(p,t)$ is the available depth at price level p at time t and $a(t), b(t)$ are respectively the best ask and best bid at time t. The parameter β can be interpreted as the expected proportion of available liquidity at the top of the book that a market order will attempt to capture. This modelling choice imposes a linear relationship between the size of market orders and the volume at best; this is in contrast with empirically observations that this relationship is roughly concave (Farmer et al., 2004).

3.3 Limit Order Arrivals

Definition 15 (Geometric Distribution) *A random variable X is said to have the Geometric distribution with success probability $p \in [0,1]$ if its probability mass function is given by,*

$$\mathbb{P}[X = x] = \begin{cases} (1-p)^{x-1} p & x \in \{1, 2, 3, \cdots\} \\ 0 & \textit{otherwise} \end{cases}$$

We write $X \sim \mathcal{G}(p)$

Let $z_1, z_2, \cdots, z_{N_z}$ be the sequence of simulated limit order arrivals. Given limit order arrival times, to simulate a limit order z, we need to generate a direction ϵ_z, a price p_z, and a volume ν_z.

Definition 16 (Normalised Order Imbalance) *We define the normalised order imbalance in the LOB at time as the normalised difference between total volume of sell orders and the total volume of buy orders at time t.*

$$NOI(t) = \frac{\sum_{z\in\mathcal{L}(t)|\epsilon_z=-1}\nu_z - \sum_{z\in\mathcal{L}(t)|\epsilon_z=1}\nu_z}{\sum_{z\in\mathcal{L}(t)}\nu_z}$$

From the definition we have that $-1 < \text{NOI}(t) < 1$. In economic terms the NOI is simply the normalised excess supply at time t.

Direction We model the direction for the limit order z_i as follows,

$$\mathbb{P}[\epsilon = \epsilon_{z_i}] = \begin{cases} \psi_{\epsilon_{z_i}} & \text{for } \epsilon_{z_i} = 1 \\ 1 - \psi_{\epsilon_{z_i}} & \text{for } \epsilon_{z_i} = -1 \end{cases} \tag{3.8}$$

$$\psi_{\epsilon_{z_i}} = \frac{\exp(\rho \times \text{NOI}(t_{z_i}))}{1 + \exp(\rho \times \text{NOI}(t_{z_i}))} \tag{3.9}$$

This modelling choice is used to induce stationarity in the order imbalance process. The parameter ρ controls the mean reversion speed of the order imbalance process. Larger values of ρ reduce the variability of the normalized order imbalance.

Definition 17 (Delta) *We define the Delta of a limit order z as the difference (in number of ticks) between the price at which the limit order is submitted and the best price.*

$$\Delta_z = \begin{cases} (b(t_z) - p_z)\tau^{-1} & \textit{for } \epsilon_z = 1 \\ (p_z - a(t_z))\tau^{-1} & \textit{for } \epsilon_z = -1 \end{cases}$$

Price Our approach to modelling limit order prices is similar to Bouchaud, Mézard, Potters, et al., 2002, although we differ on the distributional assumptions on Δ_z. We model the price of the i^{th} limit order arrival z_i as

$$p_{z_i} = \begin{cases} b(t_{z_i}) - \Delta_{z_i}\tau & \text{for } \epsilon_{z_i} = 1 \\ a(t_{z_i}) + \Delta_{z_i}\tau & \text{for } \epsilon_{z_i} = -1 \end{cases} \tag{3.10}$$

We model Δ_{z_i} as a discrete random variable with no lower or upper bound. By allowing Δ_{z_i} to be negative, we allow limit orders to be placed inside the spread. By not lower bounding Δ_{z_i}, we allow for the placing of limit orders that may be at least partially executable, such orders an in effect a combination of a market order and a non executable limit order thus this approach introduces correlations between market orders and limit orders. An effect that is not captured in models such as

Smith et al., 2003 where limit order prices cannot cross the spread. We model Δ_{z_i} as,

$$\begin{aligned} \Delta_{z_i} =& X - Y, \text{ where } X \sim \mathcal{P}(\theta^X_{z_i}), \; Y \sim \mathcal{P}(\theta^Y_{z_i}) \\ \ln(\theta^X_{z_i}) =& \delta_{00} + \delta_{01}\tilde{s}(t_{z_i}) \\ \ln(\theta^Y_{z_i}) =& \delta_{10} + \delta_{11}\tilde{s}(t_{z_i}) \end{aligned} \tag{3.11}$$

Where $\tilde{s}(t_{z_i}) = \ln(s(t_{z_i})\tau^{-1})$ is the log spread in number of ticks.

Volume We model the volume of the i^{th} limit order z_i as,

$$\begin{aligned} \nu_{z_i} =& X_{z_i}\omega \\ X_{z_i} \sim& \mathcal{G}\left(\frac{1}{u}\right) \end{aligned} \tag{3.12}$$

Where ω is the lot size. The parameter u is the mean order size of a limit-order.

If desired, the model for limit order volume may be further complicated by making u dependent on the order flow. For example, there is empirical evidence that the order size of an incoming limit order is related to the distance from best at which it is placed. Larger orders tend to be placed close to the best and order size decreases the greater the distance from best at which the order is placed (Bouchaud, Mézard, Potters, et al., 2002). It is also known that the sizes of limit orders are autocorrelated (Gould et al., 2013). These facts can be used to motivate making u a function of prior order volumes and Δ.

3.4 Limit Order Cancellations

Let $\{t_{c_1}, t_{c_2}, \cdots, t_{c_{N_c}}\}$ be the sequence of limit order cancellation times generated according to the process described in 3.1. At time t_{c_i} we cancel an active limit order $z \in \mathcal{L}(t_{c_i})$. z is chosen uniformly from $\mathcal{L}(t_{c_i})$ (the set of all active orders at cancellation time). The specified model favours the submission of limit orders relatively close to the best, thus price levels close to the best will contain more limit orders than those far from the best. As a consequence, we have that relatively more cancellations will occur close to the best, which is consistent with empirical observations (Cont, Stoikov, and Talreja, 2010).

We make the additional restrictions that a cancellation only occurs at t_{c_i} if,

$$|\{z \in \mathcal{L}(t_{c_i}) | \epsilon_z = 1\}| > V_{\min} \quad \text{and} \quad |\{z \in \mathcal{L}(t_{c_i}) | \epsilon_z = -1\}| > V_{\min}$$

(i.e both sides of the book have more than $V_{\min}$ orders). This restriction is imposed to prevent the complete depletion of the volume on a side of the book via cancellations. We fix $V_{\min} = 10$ for the remainder of this thesis. The value used for $V_{\min}$

must be large enough to prevent frequent depletions of a side of the LOB whilst being small enough as to not overly restrict the number of cancellations that occur. In our model for $V_{\min} = 10$, whilst using the parameters given in section 4.4, we observed that approximately 98% of generated limit order cancellation times resulted in cancellations.

Chapter 4

Parameter Estimation

For parameter estimation, we will make use of the freely available AAPL-10 and AAPL-50 sample data sets as provided by LOBSTER (Huang and Polak, 2011). AAPL-10 consists of level 10 order book data for apple shares traded on the NASDAQ from 09:30:00 to 16:00:00 on 21-June-2012. AAPL-50 consists of level 50 order book data for the same exchange, share and day for the time period 09:30:00 to 10:30:00. Our model simulates all levels of the book so we will use the level 50 data to estimate parameters for our event time model. The level 10 data will be used for the rest our parameters. The market has a lot size of 1 share and a tick size of 1 cent.

Each LOBSTER sample dataset consists of an 'orderbook' file which stores the evolution of the limit order book, and a 'message' file which contains the details of events which caused a change in the state of the LOB. Limit order cancellations comprise of partial deletions (where the volume of an active order is reduced) and total deletions (where an entire limit order is removed). In this thesis we have ignored partial deletions of limit orders as these represent a very small proportion ($< 2\%$) of limit order cancellations in our dataset.

4.1 Event Times

We use the `tick` library to fit a mutually exciting hawks process to the data. We fix the decay $\gamma = 0.01$. We picked this value through trial and error because it produced event quantities that where consistent with the quantities we observed in our data. The adjacencies $\{\alpha_{ij}\}$ and the baseline intensities μ_i are estimated by minimizing the least square loss,

$$L = \sum_{i=1}^{3} \int_0^T \lambda_i(t)^2 dt - 2 \int_0^T \lambda_i(t) dN_i(t) \tag{4.1}$$

$$\lambda_i(t) = \mu_i + \sum_{j=1}^{3} \sum_{t_k^j < t} \alpha_{ij} \gamma \exp(-\gamma(t - t_k^j)) \tag{4.2}$$

The `tick` library solves the above problem using accelerated proximal gradient descent (Bacry et al., 2017).

4.2 Market Orders

The LOBSTER dataset does not explicitly record market order arrivals, however these can be reconstructed using the recorded limit order execution data (by reasonably assuming that limit orders which execute at the same time are executed by the same market order).

4.2.1 Directions

From the LOBSTER data we can obtain a sequence of N_m observed market order directions $\{\epsilon_{m_1}, \epsilon_{m_2}, \cdots, \epsilon_{m_{N_m}}\}$. Our model for the market order direction process is a discrete time and state space Markov Chain. To estimate the transition probabilities we use the proportions observed in the data. Let state 1 represent a buy market order and state 0 a sell market order. Let n_{ij} be the observed number of transitions from state i to j. Then our estimate for the probability of an unsurprising market order direction is given by,

$$\hat{\pi} = \frac{n_{11} + n_{00}}{n_{10} + n_{11} + n_{01} + n_{00}}$$

4.2.2 Volume

Let $\{v_{m_1}, v_{m_2}, \cdots, v_{m_{N_m}}\}$ be the observed sequence of market order volumes corresponding to the directions $\{\epsilon_{m_1}, \epsilon_{m_2}, \cdots, \epsilon_{m_{N_m}}\}$. For each market order m_i we can obtain the volume at best $(V(b(t_{m_i}), t_{m_i}), V(a(t_{m_i}), t_{m_i}))$ just prior to the arrival of order m_i.

Set $y_i = v_{m_i} - 1$, and $x_i = \begin{cases} V(a(t_{m_i}), t_{m_i}) & \epsilon_{m_i} = 1 \\ V(b(t_{m_i}), t_{m_i}) & \epsilon_{m_i} = -1 \end{cases}$, then the MLE estimate for the parameter β can be obtained as follows,

$$\begin{aligned} L(\beta; \boldsymbol{y}, \boldsymbol{x}) &= \prod_{i=1}^{N_m} \frac{(\beta x_i)^{y_i} \exp(-\beta x_i)}{y_i!} \\ l = \log(L(\beta; \boldsymbol{y}, \boldsymbol{x})) &= \sum_{i=1}^{N_m} y_i \log(\beta x_i) - \beta x_i - \log(y_i!) \\ \frac{\partial l}{\partial \beta} = \frac{1}{\hat{\beta}} N_m \bar{y} - N_m \bar{x} = 0 & \\ \hat{\beta} &= \frac{\bar{y}}{\bar{x}} \end{aligned} \tag{4.3}$$

4.3 Limit Order Arrivals

4.3.1 Directions

If we relabel the direction of sell market order from $\epsilon_z = -1$ to $\epsilon_z = 0$ our proposed model for the limit order direction process becomes a logistic regression with no intercept term, estimation can then be done using maximum likelihood estimation. From the dataset we obtain a sequence of N_z limit order directions $\{\epsilon_{z_1}, \epsilon_{z_2}, \cdots, \epsilon_{z_{N_z}}\}$, for each limit order z we compute the normalized order imbalance observed just before the order was submitted and obtain a sequence $\{NOI(t_{z_1}), NOI(t_{z_2}), \cdots, NOI(t_{z_{N_z}})\}$ of normalized order imbalances.

Let $x_i = NOI(t_{z_i}), y_i = \epsilon_{z_i}$, from our model 3.8, we have that

$$\begin{aligned}\mathbb{P}[y = y_i] &= \begin{cases} \psi_i^{y_i}(1-\psi_i)^{1-y_i} & \text{for } y_i \in \{1, 0\} \\ 0 & \text{otherwise} \end{cases} \\ \psi_i &= \frac{\exp \rho x_i}{1 + \exp \rho x_i}\end{aligned}$$

and the likelihood function is given by,

$$L(\rho; \boldsymbol{x}) = \prod_{i=1}^{N_z} \psi_i^{y_i}(1-\psi_i)^{1-y_i}$$

Through simple algebraic manipulation, we can show that the log of the likelihood function is,

$$l(\rho; \boldsymbol{x}) = \log(L(\rho; \boldsymbol{x})) = \sum_{i=1}^{N_z} y_i x_i \rho - \log(1 + \exp \rho x_i)$$

To obtain the MLE estimate $\hat{\rho}$ for ρ, we differentiating the log-likelihood and equate to zero,

$$\begin{aligned}\frac{dl}{d\rho} = \sum_{i=1}^{N_z} \left[y_i x_i - \frac{x_i \exp \hat{\rho} x_i}{1 + \exp \hat{\rho} x_i} \right] &= 0 \\ \sum_{i=1}^{N_z} x_i \left(y_i - \frac{\exp \hat{\rho} x_i}{1 + \exp \hat{\rho} x_i} \right) &= 0 \qquad (4.4)\end{aligned}$$

Equation 4.4 can be solved numerically to obtain an estimate for ρ.

4.3.2 Price

Our proposed model for the limit order deltas 3.11 is a Skellam regression. Estimation of the parameters $\delta = \{\delta_{00}, \delta_{01}, \delta_{10}, \delta_{11}\}$ is done via maximum likelihood estimation.

Definition 18 (Modified Bessel Function of the First Kind of Real Order) *We define $I_k : \mathbb{R} \to \mathbb{R}$ as the modified Bessel function of the first kind and order k, it is given by,*

$$I_k(x) = \sum_{m=0}^{\infty} [m!\Gamma(m+k+1)]^{-1} \left(\frac{x}{2}\right)^{2m+k} \qquad x, k \in \mathbb{R}$$

where $\Gamma(z) = \int_0^\infty x^{z-1}\exp(-x)dx \quad for \quad z \in (0,\infty)$ *is the gamma function*

Theorem 4.3.1 (Distribution of Difference of Independent Possions) *if* $K = X - Y$ *where* $X \sim \mathcal{P}(\theta_x), \;\; Y \sim \mathcal{P}(\theta_y)$*, then the probability mass function of K is given by (Skellam, 1946),*

$$\mathbb{P}[K = k] = e^{-(\theta_x+\theta_y)}\left(\frac{\theta_x}{\theta_y}\right)^{\frac{k}{2}} I_k(2\sqrt{\theta_x,\theta_y}) \tag{4.5}$$

where $I_k(x)$ *is the k-order modified Bessel function of the first kind of real order.*

From our data we obtain a sequence $\{k_{z_1}, k_{z_2}, \cdots, K_{z_N}\}$ of N observed deltas and a sequence $\{\tilde{s}(t_{z_1}), \tilde{s}(t_{z_2}), \cdots, \tilde{s}(t_{z_N})\}$ of N observed log spreads for each limit order submission. Using the 4.5 we construct a log-likelihood function for the δ parameters in our proposed model.

Where $\theta^X_{z_i} = \exp\left[\delta_{00} + \delta_{01}\tilde{s}(t_{z_i})\right]$, $\theta^Y_{z_i} = \exp\left[\delta_{10} + \delta_{11}\tilde{s}(t_{z_i})\right]$

$$\begin{aligned} 4.5, 3.11 \implies \mathbb{P}[\Delta_{z_i} = k_{z_i}] =& e^{-(\theta^X_{z_i}+\theta^Y_{z_i})}\left(\frac{\theta^X_{z_i}}{\theta^Y_{z_i}}\right)^{\frac{k_{z_i}}{2}} I_{k_{z_i}}\left(2\sqrt{\theta^X_{z_i},\theta^Y_{z_i}}\right) \\ \implies \log\left(\mathbb{P}[\Delta_{z_i} = k_{z_i}]\right) =& -\left(\theta^X_{z_i} + \theta^Y_{z_i}\right) + \frac{k_{z_i}}{2}\left[\log\left(\theta^X_{z_i}\right) - \log\left(\theta^Y_{z_i}\right)\right] + \log\left(I_{k_{z_i}}\left(2\sqrt{\theta^X_{z_i},\theta^Y_{z_i}}\right)\right) \end{aligned}$$

The log likelihood function is thus give by

$$l(\delta; \boldsymbol{k}, \boldsymbol{\tilde{s}}) = \sum_{i=1}^{N} \log\left(\mathbb{P}[\Delta_{z_i} = k_{z_i}]\right) \tag{4.6}$$

To obtain estimates for the δ parameters we maximize 4.6 numerically using the l-BFGS-B implementation in the python library `scipy` (Virtanen et al., 2020). For this estimation we would ideally use level 50 order book data for an entire trading day. Due to insufficient data points in APPL-50 we are forced to use APPL-10, resulting estimates will thus produce a Δ distribution which underestimates the number of orders that are placed deep in the book.

4.3.3 Volume

Let $\{v_{m_1}, v_{m_2}, \cdots, v_{m_{N_m}}\}$ be our sequence of observed market order sizes. We estimate u using maximum likelihood estimation and obtain,

$$\hat{u} = \frac{1}{N_m}\sum_{i=1}^{N_m} v_{m_i}$$

4.4 Parameter Estimates

The estimated adjacency matrix is,

$$A = \begin{bmatrix} 0.5885 & 0 & 0 \\ 1.5343 & 0.1460 & 0.6644 \\ 1.4183 & 0.0880 & 0.6706 \end{bmatrix} \tag{4.7}$$

The estimated limited order price parameters are,

$$D = \{\delta_{ij}\} = \begin{bmatrix} 3.1228 & 0.1618 \\ 2..5644 & 0.2954 \end{bmatrix} \tag{4.8}$$

Estimates for the remaining parameters are given in Table 4.1.

Parameter	Estimate
π	0.7955
β	0.5896
ρ	0.6907
u	89.9719
μ_1	0.5650
μ_2	1.1360
μ_3	1.0417

TABLE 4.1: Estimated Parameters

4.5 Initializing the Limit Order Book

For simulation purposes we require an initial limit order book state $\mathcal{L}(t_0)$. Various approaches can be used to obtain an initial state.$\mathcal{L}(t_0)$ could be user specified if, for example, one wishes to study the effect of different initial states on some outcome. $\mathcal{L}(t_0)$ could also be randomly generated. In this thesis, unless stated other wise we will use the initial state of the AAPL-50 dataset as our initial state. Table 4.2 shows the values of various LOB quantities for this initial state.

Quantity	Value
Mid-Price	58563.5 cents
Best Bid	58533.0 cents
Best Ask	58594.0 cents
Spread	61 cents
Volume at Best Bid	18 shares
Volume at Best Ask	200 shares
Buy Side Volume	8771 shares
Sell Side Volume	12659 shares

TABLE 4.2: Quantities of Interest in AAPL-50 initial state

Chapter 5

Empirical Analysis of The Simulated Limit Order Book

In this chapter, we examine the empirical properties of realisations of the model described in Chapter 3. We compare these properties to empirical observations from the literature.

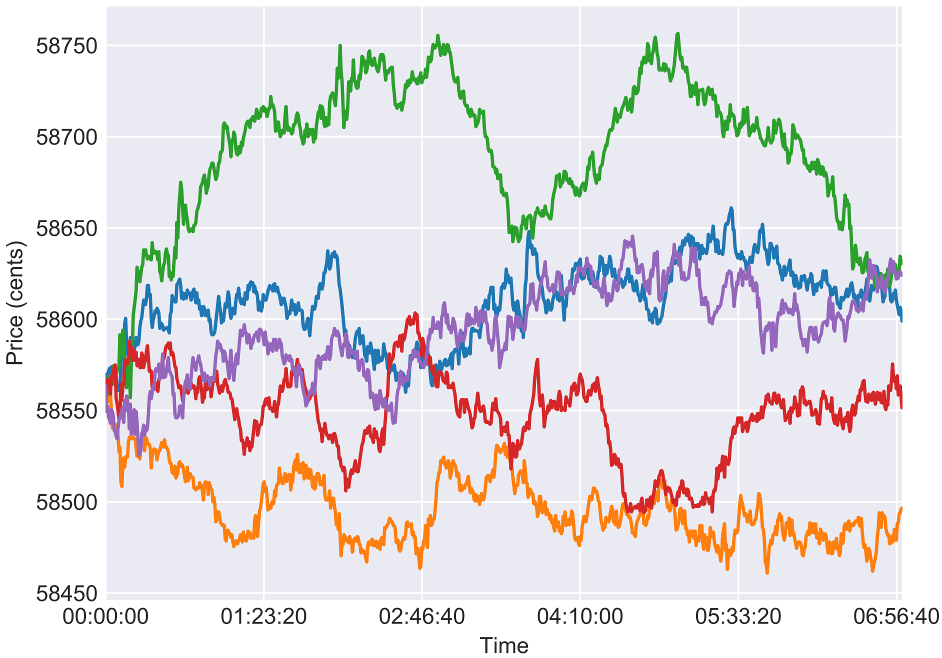
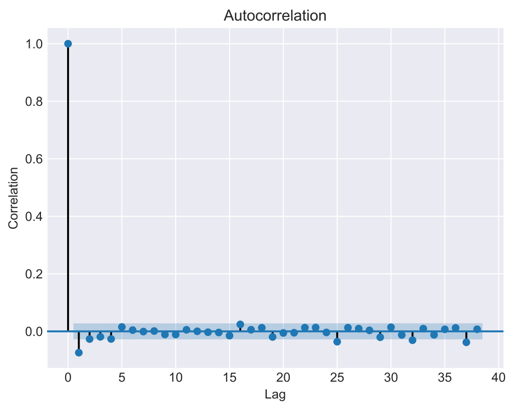


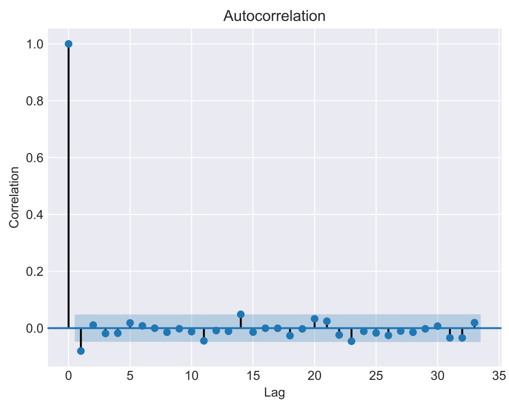
FIGURE 5.1: Simulated Mid-price Paths - 100ms Sampling

5.1 Returns

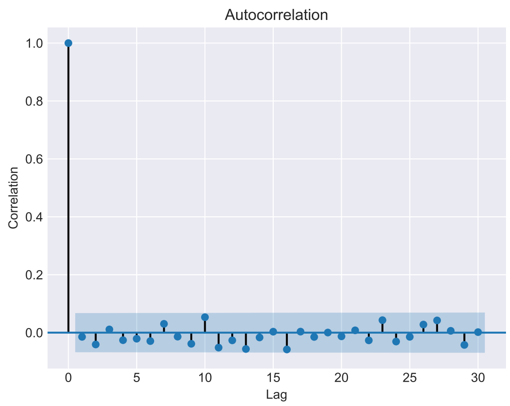
Absence of Autocorrelations Mid-price returns are known to exhibit weak negative autocorrelation at very short timescales but no autocorrelations at longer time scales (Chakraborti et al., 2011; Gould et al., 2013). Figure 5.2 shows autocorrelation plots for the mid-price log-returns at various time scales of a realisation of our model. A few statistically significant negative autocorrelations can be seen (most notably at lag 1) in the 5 second returns. At the 15 second time scale a statistically significant negative autocorrelation can be seen at lag 1. There are no statistically significant returns at the 30 second or 60 second time scale.



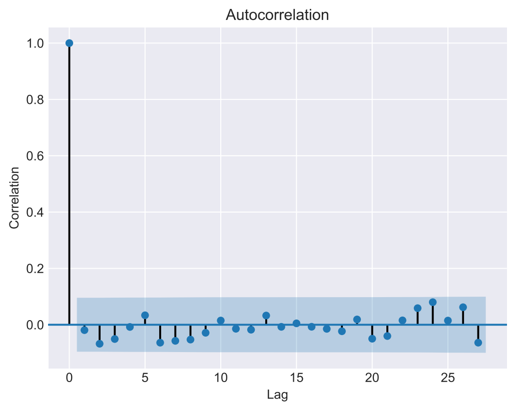
(A) 5 Second Returns



(B) 15 Second Returns



(C) 30 Second Returns



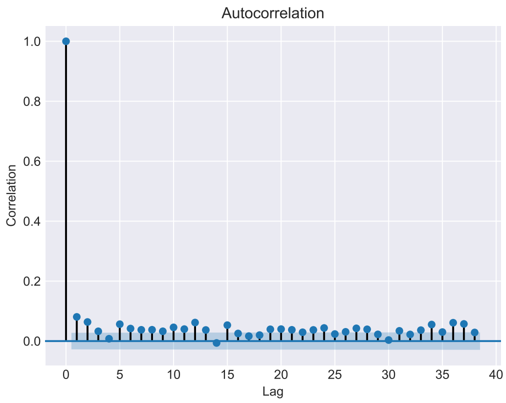
(D) 60 Second Returns

FIGURE 5.2: Autocorrelation Plots of Mid-Price Returns
Translucent blue area represents 95% confidence intervals computed via Bartlett's formula

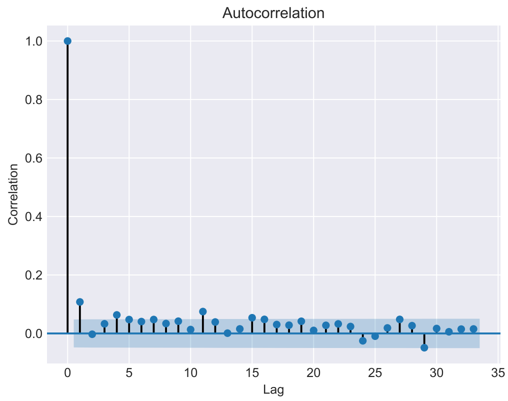
Autocorrelation of Absolute Returns (Volatility Clustering) The absolute value of returns and the squared returns processes are known to posses positive long-range dependencies (Chakraborti et al., 2011). Figure 5.3 shows autocorrelation plots for absolute returns of a realisation of our model measured at varying time scales. We observe long range positive dependencies for 5-second returns. We observe statistically significant positive dependencies for 15 and 30-second returns and for 60-second returns we observe a significant positive autocorrelation only at lag 2.

Leptokurtic Distribution The empirical distribution of returns is known to be 'fat-tailed' in that it has greater kurtosis than the Gaussian distribution. This is seen at all time-scales (Gould et al., 2013). Extreme returns occur with greater frequency than would be expected under a Gaussian distribution. Figure 5.4 shows qq-plots of mid-price returns measured at four time scales for a realisation of our model against the Gaussian distribution. At all four time scales the tails of the returns distribution are heavier than the Gaussian distribution.

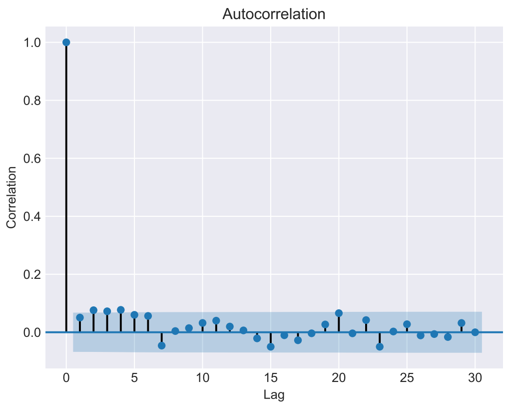
Aggregational Gaussianity Empirical observations suggest that as the time scale over which returns are computed is increased, the return distribution becomes less leptokurtic and more closely resembles a Gaussian distribution. In our simulated results (Figure 5.4) we noted that the tails of the return distribution appear to fit the Gaussian distribution better as we increase the length of the time interval over which returns are computed. This is consistent with empirical observations.



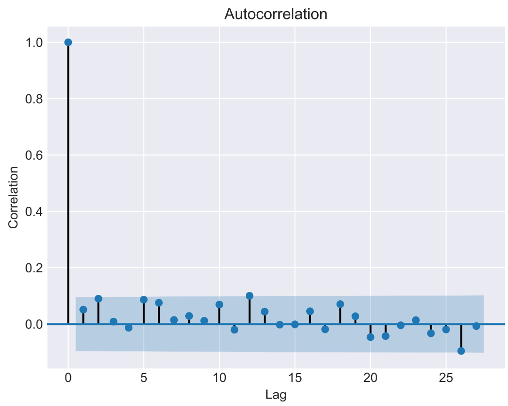
(A) 5 Second Absolute Returns



(B) 15 Second Absolute Returns

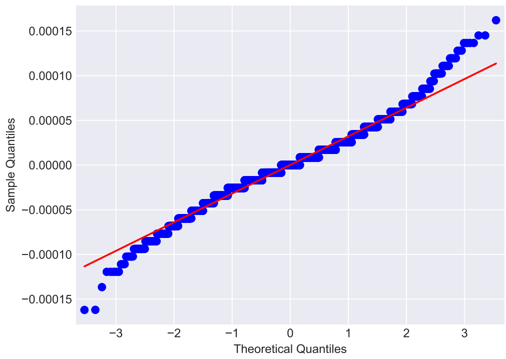


(C) 30 Second Absolute Returns

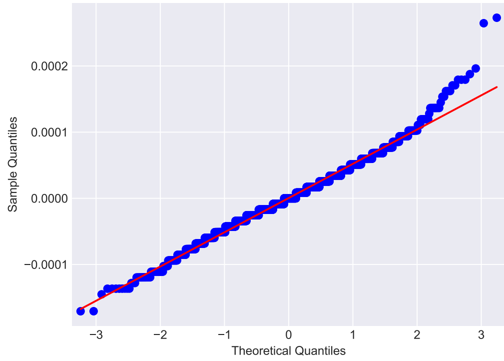


(D) 60 Second Absolute Returns

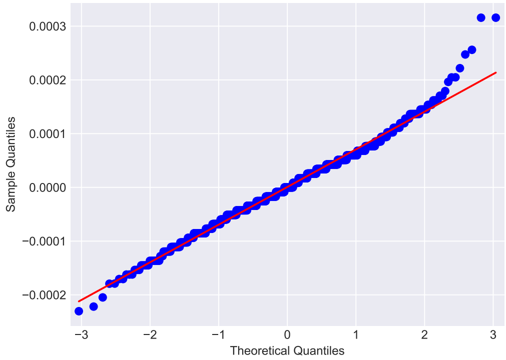
FIGURE 5.3: Autocorrelation Plots of Mid-Price Absolute Returns
Translucent blue area represents 95% confidence intervals computed via Bartlett's formula



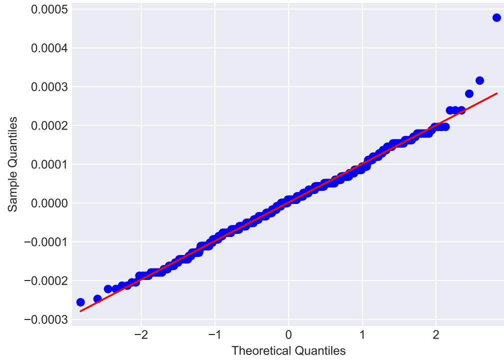
(A) 5 Second Returns



(B) 15 Second Returns



(C) 30 Second Returns

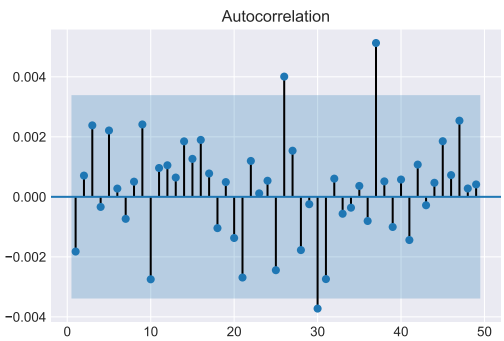


(D) 60 Second Returns

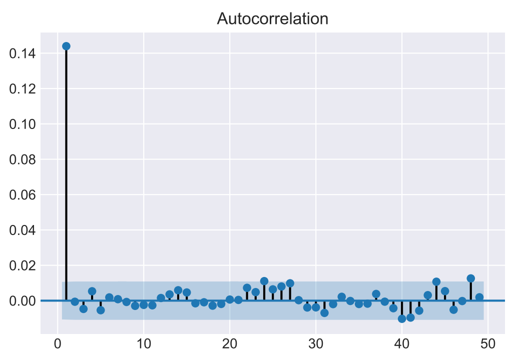
FIGURE 5.4: QQ-Plots of Log-Returns vs Gaussian Distribution

5.2 Order Flow

Market and Limit Order Size Figure 5.5 shows the autocorrelation functions for market order and limit order sizes. Empirical studies suggest that order sizes should display a long range positive autocorrelation structure (Gould et al., 2013). From figure 5.5 we note that our simulation fails to reproduce this order flow property. Limit order sizes display no autocorrelation structure, figure 5.5a. For market order sizes a significant correlation is present only at lag 1, figure 5.5b. This autocorrelation is due to the dependence of market orders on the depth at best.



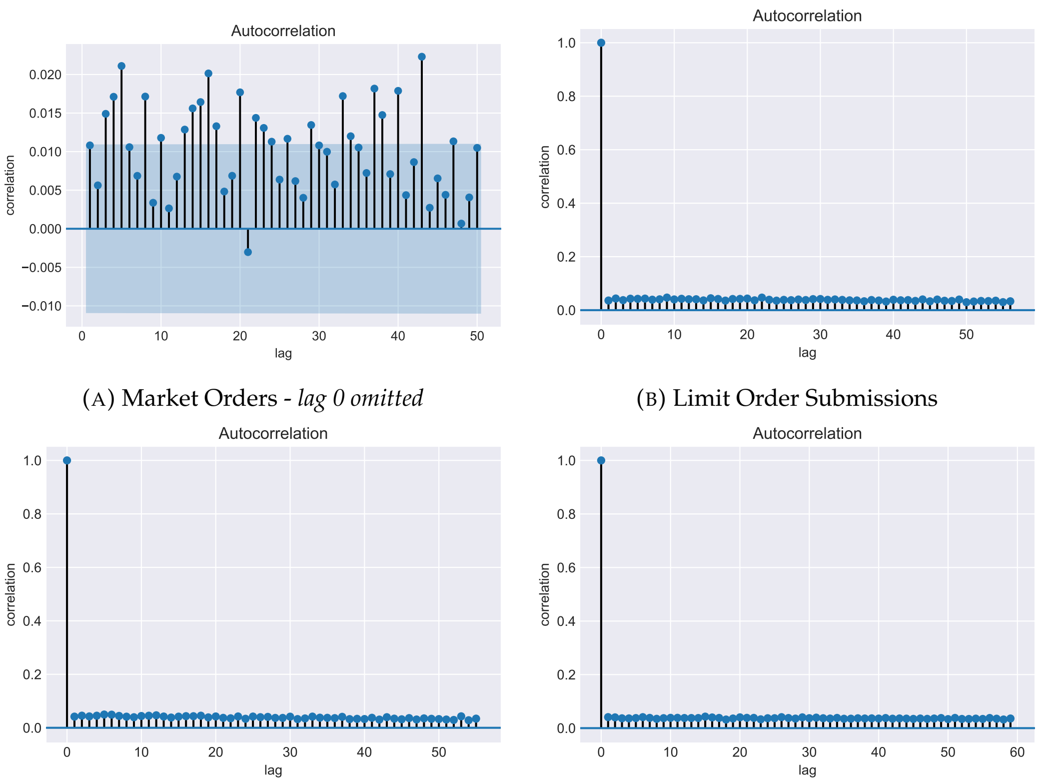
(A) Limit Order Size - *lag 0 omitted*



(B) Market Order Size - *lag 0 omitted*

FIGURE 5.5: Autocorrelation Plots for Simulated Order Sizes

Arrival Times Figure 5.6 shows autocorrelation functions for the simulated inter-arrival times of all three event types and the combined event arrival process. Limit order submissions and cancellations display long-range statistically significant positive auto-correlations. Market orders exhibit a positive autocorrelation structure but the autocorrelations are weaker than those of the other order types. These results are expected given that our event times are generated by hawks processes, and are consistent with various empirical studies of real data (Gould et al., 2013).



(A) Market Orders - *lag 0 omitted*

(B) Limit Order Submissions

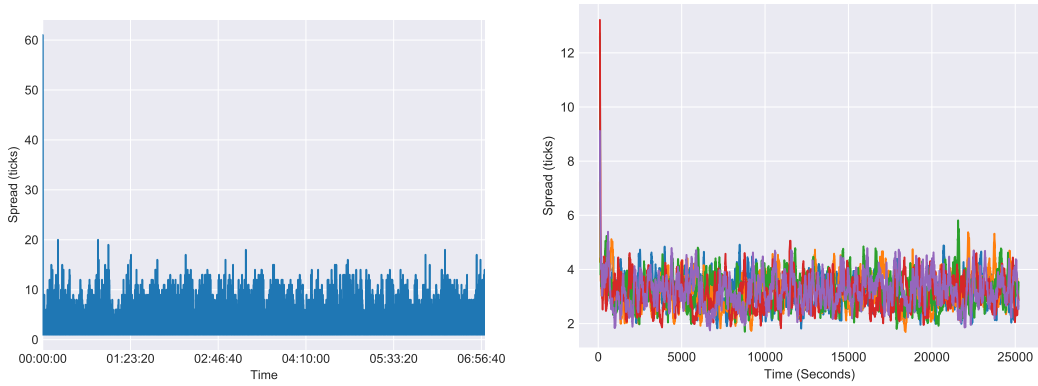
(C) Limit Order Cancellations

(D) All Event Arrivals

FIGURE 5.6: Autocorrelations of Inter-arrival times

Spread Figure 5.7 shows a single realisation of the spread as well as the 100 second rolling averages of 5 realisations of the spread. In our chosen initial state the LOB has a spread of 61 ticks $s(t_0) = 61$. From the figure we observe that the simulated spread falls rapidly from the initial state. We observed mean spreads of approximately 3 - 3.5 ticks and median spreads of approximately 3 ticks in our simulations. In our data set we observed mean and median spreads of approximately 15 ticks. Our models underestimation of the spread is caused by data limitations when estimating the parameters $\delta_{00}, \delta_{01}, \delta_{10}, \delta_{11}$ as discussed in section 4.3.2. We where able to achieve more realistic spread realisations by heuristically adjusting these parameters. However, these adjustments had no material effect on any of the results presented in this thesis so, for the sake of reproducibility, we have opted to omit any adjustments and use the estimated spread parameters as is.

Furthermore we explored the addition of time dependencies to the spread parameters which produced time dependent spread realisations. However given our data set, sound estimation of the additional parameters introduced by this complication where not possible.



(A) Single Spread Realisation - 100ms Sampling (B) Spread Realisations - 100 sec rolling average

FIGURE 5.7: Simulated Spread

5.3 Market Impact

5.3.1 Impact Function of a Market Order

Let $m = (\epsilon_m, v_m, t_m)$ be a deterministic market order, i.e. each component of the triple is known and fixed. Let $M(t)$ be the mid-price process. We define the market impact at time $t > t_m$ due to order m as,

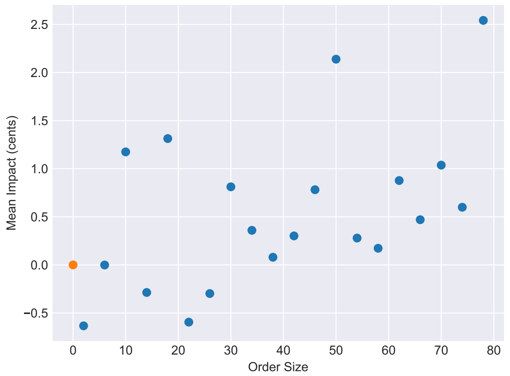
$$\mathcal{I}_m(t) = \mathbb{E}[M(t)|\mathcal{F}_{t_m}, m] - \mathbb{E}[M(t)|\mathcal{F}_{t_m}] \ \ \forall t > t_m \tag{5.1}$$

The first expectation term in 5.1 is the expected mid price in the case where market order m is introduced into the market and the second term is the expected mid price in the case that m does not occur. This definition is similar to that of the *reaction impact* given by Bouchaud et al., 2018. In real data this impact function cannot be observed as only one realisation of the mid-price process is available. The function can, however, be easily computed in a simulated market and we do so here.

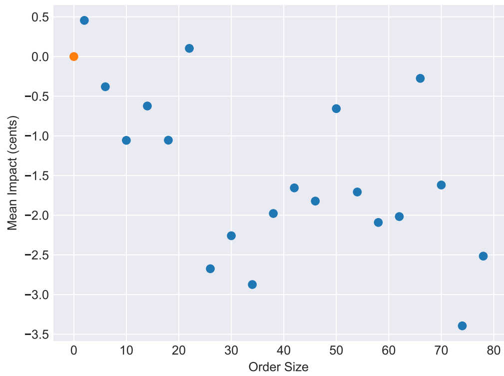
In this section, we explore how the above impact function varies in time and for different values of v_m within our simulated LOB.

5.3.2 Order Size Dependence

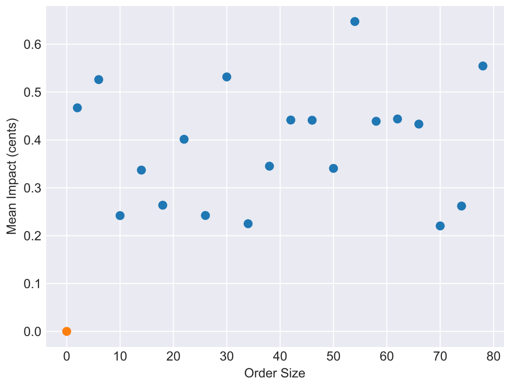
Figure 5.8 shows the mean market impact of a market order observed in our simulation 10 minutes and 1 minute after its arrival in both the buy and sell case. At the time of arrival of the market orders the queue at the best ask consists of 200 shares whilst only 18 shares are present at the best bid. As a result, sell market orders have greater impact than buy market orders. This is illustrated in figure 5.8d where a large difference can be seen at the 1 minute time horizon between the sell orders of size 14 and 18. From figure 5.8 we note that there is generally a positive relationship between the order size and the magnitude of the market impact. However, from figure 5.8c we note that at short time scales when the orders do not have instantaneous impact there is virtually no dependence on size and even small orders have meaningful impact.



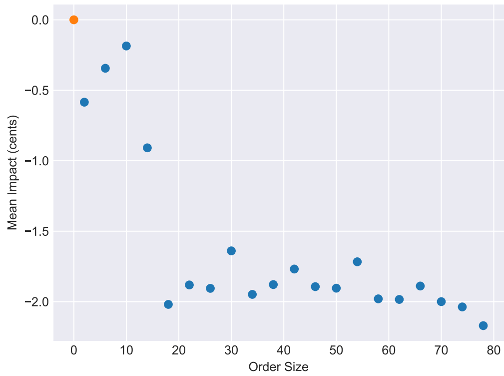
(A) Buy Market Order Impact - 10 Minutes Since Arrival, 2000 simulations per point



(B) Sell Market Order Impact - 10 Minutes Since Arrival, 2000 simulations per point



(C) Buy Market Order Impact - 1 Minute Since Arrival, 20000 simulations per point



(D) Sell Market Order Impact - 1 Minute Since Arrival, 20000 simulations per point

FIGURE 5.8: Market Impact - Size Dependence, the orange 0 point is fixed and not obtained from simulation

It is worth noting that a significant amount of noise exists in the plots given in figure 5.8. Each point in the plots is an average over several realisations; at 1 minute there is a standard deviation of ≈ 20 for each order size, at 10 minutes we observed a standard deviation of ≈ 40. Confidence intervals for each point are extremely wide, especially in the 10 minute case. Such a low signal to noise ratio for this relationship is consistent with what has been observed empirically in real data (Almgren, 2020).

5.3.3 Time Evolution

Figure 5.9a shows 4 realisations of the impact function for 4 different orders. Each impact function was generated using 500 simulations. The state of the limit order book at the time of market order arrival is given in table 4.2. The asymmetry between the impact of buy and sell orders is a reflection of the dependence of market impact on the state of the limit order book at the order arrival time. Our initial state has greater sell volume than buy volume and greater volume at the best ask than the best sell. Sell orders of size 18 or greater trigger an immediate price move. Such orders are said to have instantaneous market impact. The queue at the best ask contains 200 shares thus none of the buy market orders shown have instantaneous market impact.

(A) Market Order Impact - Various Order Sizes/Directions, 1 Hour Since Arrival. 500 simulations per realisation

(B) Market Order Impact - Buy Size 1 Realisations, 15 minutes from arrival, 5000 simulations per realisation

FIGURE 5.9: Market Impact - Time Evolution

From figure 5.9a we observe that market impact seems to peak sometime after the order arrival then begins to decay. Larger orders seem to take longer to attain their peak impact. In the figure the impact functions of the buy orders of size 50 and 100 cross 0 at around 10 and 20 minutes respectively. This does not imply the impact is completely diminished at this point, but rather the effect is too small to be detected with the sample size used. This is made clear by Figure 5.9b which shows 4 impact function realisations of buy market orders of volume 1 up to 15 minutes after order arrival. In this figure we used 5000 simulations to generate each impact function realisation. We note again that impact appears to build to a peak value over time then begin to decay, but even for an order of size 1 impact is still present 15 minutes after order arrival.

Chapter 6

Relating Order Flow to Volatility

In this chapter we introduce Generalised Additive Models (GAMs). To study the effect of the model parameters on the simulated market volatility, we construct and conduct a simulation study. We fit a GAM with observed volatility from our model as the dependent variable and our model parameters as independent variables.

6.1 Generalized Additive Models

Suppose our data set consists of n pairs $(\boldsymbol{x}_i, y_i)$, where $\boldsymbol{x}_i \in \mathbb{R}^p$ is the vector of independent variables, and y_i is termed the dependent variable. A Generalized Additive Model assumes that $y_i \sim EF(\mu_i, \phi)$, that is the dependent variable y_i is distributed according to an exponential family distribution with mean μ_i and scale parameter ϕ. The mean of y_i is related to the explanatory variables via,

$$g(\mu_i) = \eta_i = c + f_1(x_{1i}) + f_2(x_{2i}) + \cdots + f_p(x_{pi}) \tag{6.1}$$

g is called the link function and is usually chosen such that it is invertible and its inverse g^{-1} maps $\eta_i \in \mathbb{R}$ to the domain of μ_i. $c \in \mathbb{R}$ is the intercept term. Additionally 6.1 can be extended to include interaction terms, e.g.

$$g(\mu_i) = \eta_i = c + f_1(x_{1i}) + f_2(x_{2i}) + \cdots + f_p(x_{pi}) + f_{p+1}(x_{1i}, x_{2i}) \tag{6.2}$$

The functions f_j are smooth functions of the independent variables. There are several ways in which the f_j maybe be constructed. We describe here the smoothing approach known as P-splines as this is the method used in the `pygam` python library (Servén and Brummitt, 2018). For a detailed treatment of the subject see Wood, 2017.

6.1.1 Constructing Smooth Functions

Identifiability Problem The presence of multiple functions in the GAM equation 6.1 results in an identifiability problem. The f_j can only be estimated up to an additive constant since a constant may be added to one of the f_j and subtracted from another and this would have no effect on η_i (Wood, 2017). A common solution to this problem is the introduction of a sum-to-zero constraints on the f_j, i.e.,

$$\sum_{i=1}^{n} f_j(x_{ji}) = 0$$

these constraints allow the f_j to maintain their shape but they shift the f_j vertically so that they have a mean value of zero (Wood, 2017). This should be kept in mind when interpreting the partial dependence plots given in subsection 6.2.2.

P-Splines and B-Splines One method of constructing the f_j would be through the use of B-splines. Let $x_1 \leq x_2 \leq \cdots x_{k+m+2}$ be a sequence of knots. An $(m+1)^{\text{th}}$ order spline with a k parameter B-spline basis is given by

$$f(x) = \sum_{i=1}^{k} B_i^m(x)\beta_i \tag{6.3}$$

$$B_i^m(x) = \frac{x - x_i}{x_{i+m+1} - x_i} B_i^{m-1}(x) + \frac{x_{i+m+2} - x}{x_{i+m+2} - x_{i+1}} B_{i+1}^{m-1}(x) \quad \forall i \in \{1, 2, \cdots, k\} \tag{6.4}$$

$$B_i^{-1}(x) = \begin{cases} 1 & x_i \leq x \leq x_{i+1} \\ 0 & \text{otherwise} \end{cases} \tag{6.5}$$

P-splines (Penalized B-splines) are smoothers based on B-splines, they have a quadratic smoothing penalty applied directly to the β_i parameters to produce a reduced rank smoother. It is common practice to select the value of the smoothing coefficient using a cross-validation procedure.

Model Assumptions The generalized additive model assumptions are nearly identical to those made by linear regression. The key difference is that GAMs relax the assumption of covariates being linearly related to mean of the independent variable. Thus when fitting a GAM, as in linear regression, we need to check that the model assumptions are reasonable, this is done in section 6.2.3.

6.2 Simulation Study

6.2.1 Study Description

We fix the decay, $\gamma = 0.01$ and the adjacencies $\{\alpha^{ij}\}$ to the values estimated from the data (see section 4.4). We simulate N = 1000 realizations of our model, each with a run time of T = 60 minutes. With varying order flow parameters. The non-fixed order flow parameters and our interpretations as given in 3 are,

π - the probability of an unsurprising market order sign

β - average proportion of top of the book liquidity captured by market orders

ρ - mean reversion speed of the order imbalance

u - the average volume of a limit order

$\delta_{00}, \delta_{01}, \delta_{10}, \delta_{11}$ - the limit order price parameters

μ_1 - base rate of market order arrivals

μ_2 - base rate of limit order submissions

μ_3 - base rate of limit order cancellations

In each simulation, for each order flow parameter θ we sample a parameter value $\tilde{\theta}$ for use in the simulation from $\tilde{\theta} \sim U[LB_\theta, UB_\theta]$. From each simulation we compute the realized mid-price volatility. The bounds LB_θ, UB_θ used for each model parameter can be deduced from the x-axis ranges given in the partial dependence plots (section 6.2.2). In order to infer the functional relationships between the order flow parameters and the volatility, we fit a gamma generalized additive model with the realized volatility as the dependent variable and the order flow parameters as independent variables. The model formulation is as follows, using the mean parametrization of the gamma distribution,

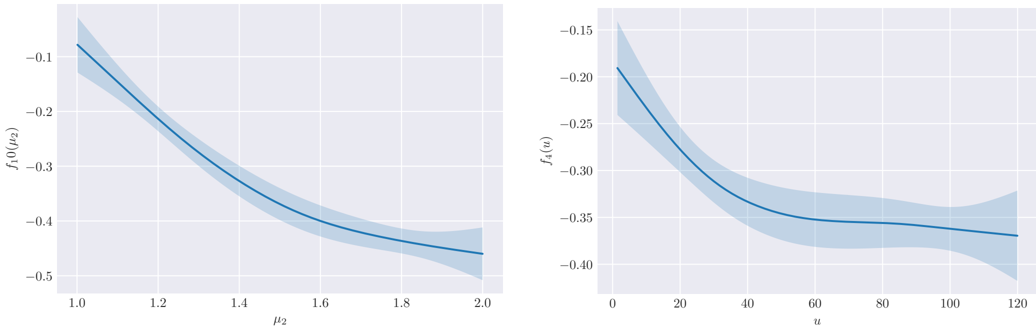
$$\begin{aligned} \sigma_i &\sim \Gamma(\mu_i, \nu) \\ \log(\mu_i) &= c + f_1(\pi) + f_2(\beta) + \cdots + f_{11}(\mu_3) + f_{12}(\mu_2, \mu_3) + f_{13}(\beta, u) \end{aligned} \tag{6.6}$$

We include terms to capture possible μ_2, μ_3 and β, u interactions as we would reasonably expect an interaction effect between the rates of limit order submissions and limit order cancellations. We also wish to explore the interaction between the average limit order size and the average proportion of liquidity at best consumed by market orders.

6.2.2 Partial Dependence Plots

In the following plots, we examine the shapes of the smooth functions f_j corresponding to each of the model parameters. As mentioned previously, due to the identifiability problem inherent in GAMs, the f_j are only estimated up to an additive constant. The blue translucent areas represent 95% confidence intervals.

Negatively Related to Volatility The parameters μ_2 and u are negatively related to the volatility. Increases in these parameters increase the queue sizes at each price level and the number of occupied price levels (decreases the sparsity of the book) which diminishes the ability of incoming market orders to cause price changes. From the scale of the y-axis in figure 6.1b we note that the effect of the mean limit order size on the volatility is much weaker than the effect of the rate of limit order submissions.



(A) μ_2 - base rate of limit order submissions (B) u - mean limit order size

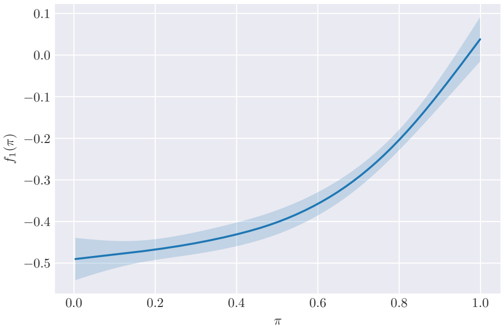
FIGURE 6.1: Model Parameters Negatively Related to Volatility

Positively Related to Volatility The parameters π, β, δ_{10}, δ_{11}, μ_1, μ_3 are positively related to volatility.

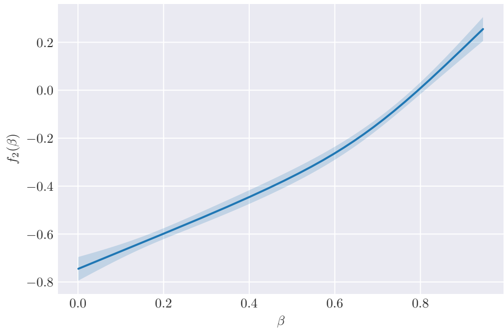
Increases in π make it more likely for markets orders of the same sign to occur consecutively which results in more depletions of the queue at best, and thus more price changes. Increases in β increase the proportion of the top of the book captured by incoming market orders. This increases the probability of an incoming market order triggering a price change. From figures 6.2a and 6.2b we note that there appears to be a convex relationship between π, β and the observed volatility.

Since the log-spread is non-negative, increases in the model parameters δ_{10}, δ_{11} increase the probability of a limit order being placed inside the spread. Such orders establish a new best and cause a price change contributing to volatility. Figures 6.2c and 6.2c suggest that this relationship may be linear or well approximated by a linear function.

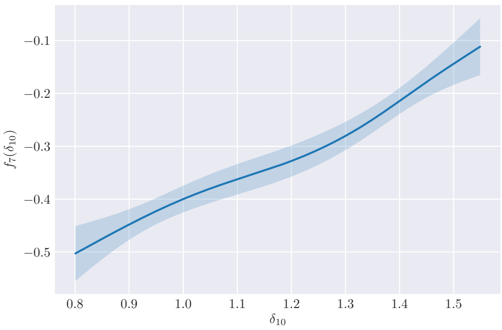
Increases in μ_1 cause an increase in market orders arriving into the market resulting in increased price movements. Increases in μ_3 increase the number of limit order cancellations. Cancellations can trigger price changes directly by depleting the queue at best. They also have an indirect effect in that they increase the sparsity of the book making large price moves more likely. From figures 6.2e and 6.2f we observe that the relationships between μ_1, μ_3 and the volatility appears to be concave and convex respectively.



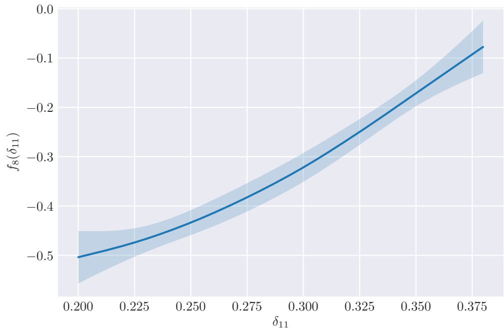
(A) π - probability of an unsurprising market order sign



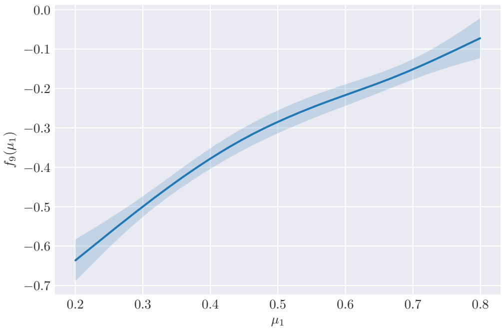
(B) β - mean market order size as a proportion of volume at best



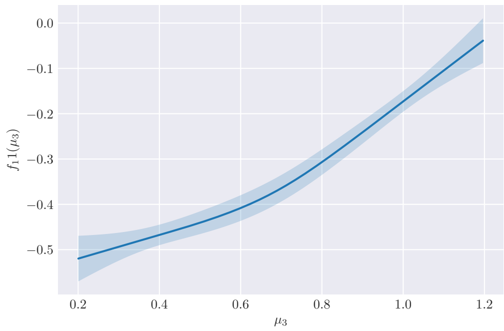
(C) δ_{10}



(D) δ_{11}



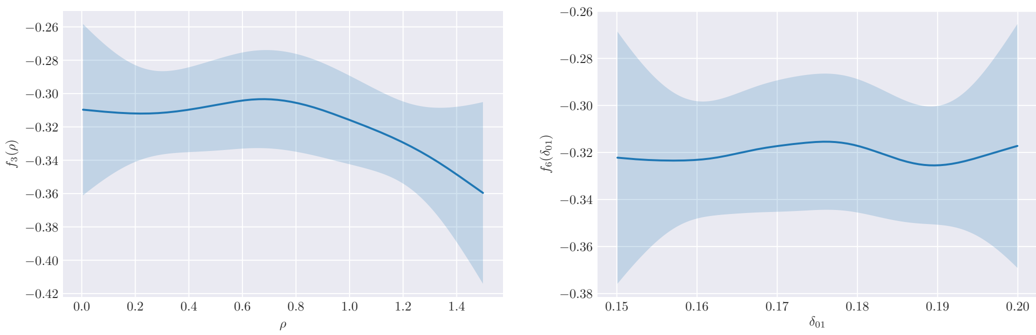
(E) μ_1 - base rate of market order arrivals



(F) μ_3 - base rate of limit order cancellations

FIGURE 6.2: Parameters Positively Related to Volatility

Not Related to Volatility We found no statistically significant relationship between the model parameters δ_{01}, ρ and the volatility, as can be seen from figures 6.3b and 6.3a, the partial dependence plot for the smooth function corresponding to these parameters may well be horizontal lines.



(A) ρ - mean reversion speed of NOI

(B) δ_{01}

FIGURE 6.3: Model Parameters Unrelated to Volatility

Non-Monotonic Relationships The relationship between the volatility and the model parameter δ_{00} appears parabolic, figure 6.4. A possible explanation for this is that when δ_{00} increases initially, it decreases the probability of an order being placed inside the spread, thus volatility initially falls but as δ_{00} increases further an increasingly large proportion of limit orders are being placed deep in the book resulting in lower volumes and greater sparsity at the top of the book.

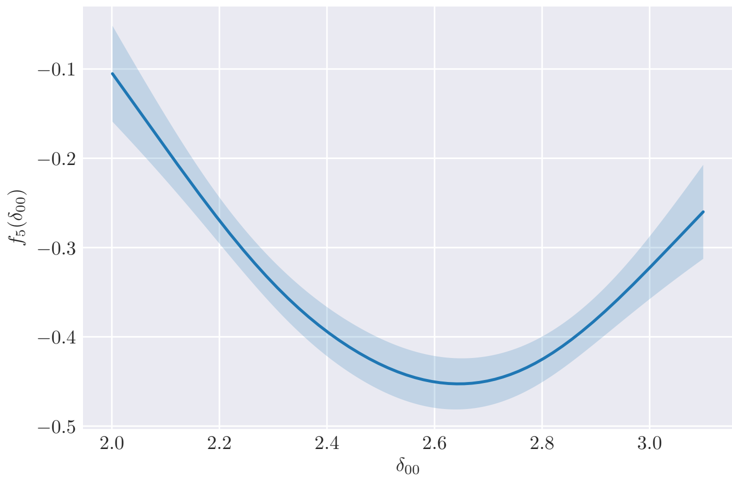
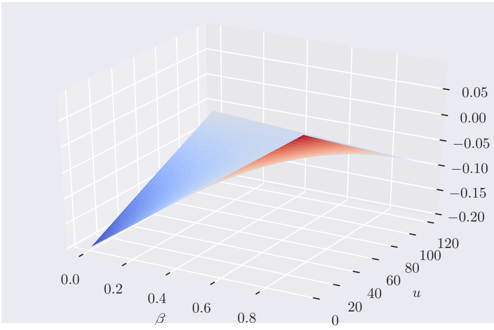


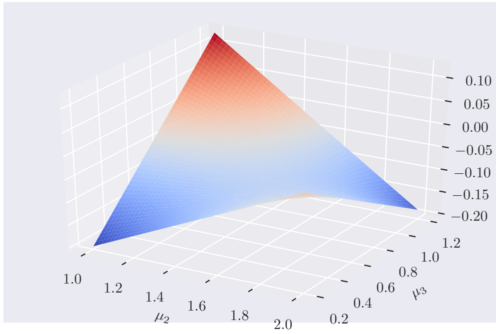
FIGURE 6.4: δ_{00} Partial Dependence Plot

Interaction Effects Figure 6.5a shows the partial dependence plot of the $\beta - u$ interaction. We observe that when both parameters are decreasing (or both are increasing) there is a decrease in volatility. Increasing β while decreasing the mean limit order size increases the volatility.

Figure 6.5b shows the partial dependence plot of the $\mu_2 - \mu_3$ interaction. When the arrival rate of limit orders is increasing and cancellations are decreasing (or vice-versa) the result is an increase in volatility. If both are increasing (or decreasing) the is a decrease in volatility.



(A) β and u interaction

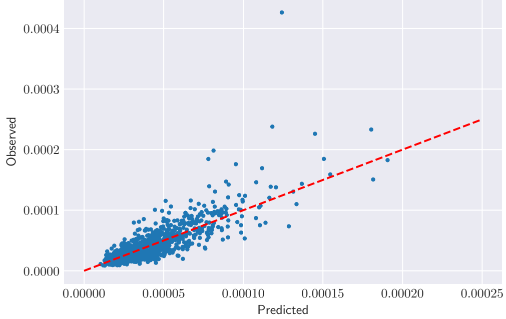


(B) μ_2 and μ_3 interaction

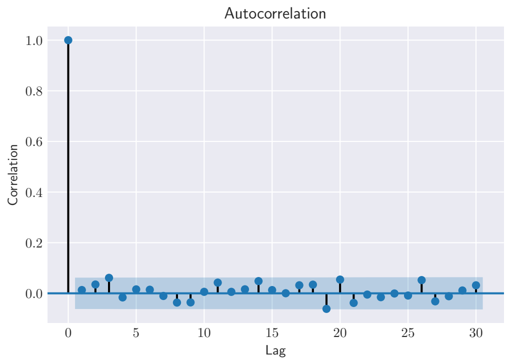
FIGURE 6.5: Interaction Effects

6.2.3 Model Diagnostics

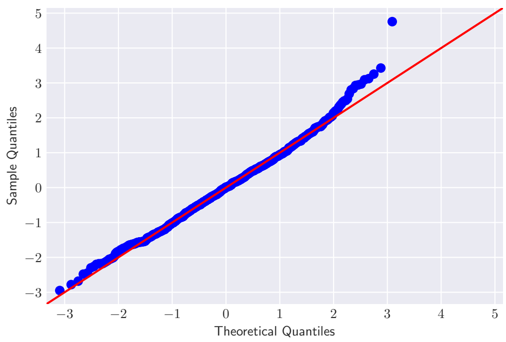
We produce model diagnostic plots in order to check that our model reasonably satisfies the GAM assumptions.



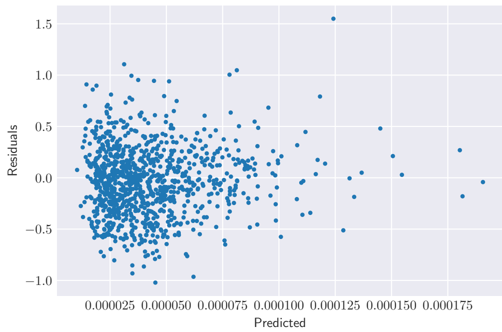
(A) Observed vs. Predicted



(B) Autocorrelation of Deviance Residuals



(C) Deviance Residuals QQ-plot



(D) Deviance Residuals vs. Predicted

FIGURE 6.6: Model Diagnostic Plots

Autocorrelation of Residuals We check the assumption of no autocorrelation of the residuals by examining the autocorrelation plot of the residuals, figure 6.6b. There are no statistically significant autocorrelations.

Normality of Residuals GAMs assume residuals are normally distribution with zero mean. We examine this assumption by plotting the quantiles of the deviance residuals against those of the normal distribution, figure 6.6c. Although we observe deviations from normality in the tails, this assumption appears to hold reasonably.

Homoscedasticity To check the assumption that the residuals have constant variance, we plot the deviance residuals against the predicted values, figure 6.6d. If this assumption held, we would expect a random scatterer of residuals around zero. Our plot indicates that this is not the case.

Linearity The GAM assumes a linear relationship between the log-mean of the volatility and the smooth functions of the model parameters. The validity of this assumption is related to the Homoscedasticity assumption and can be checked using figure 6.6d. Although the residuals are distributed around zero, we observe a relationship between the size of the predicted value and the variability of the residuals. This indicates that the true relationship between the smooth functions and the log-mean is non-linear. However, given the lack of severe violations we conclude that the model approximates the true relationship sufficiently well for the results in 6.2.2 to be valid.

Chapter 7

Application: Testing Optimal Order Placement

The trading process consists of several stages, each stage is characterised by a key trading decision. These stages occur at different timescales. Broadly, they can be described as: The portfolio allocation stage - where investment managers decide on their desired portfolio composition and which assets will be bought or sold to achieve it - this stage occurs on a timescale of days to weeks. The optimal execution stage - where the decision of how a large order should be split in order to minimise market impact - this occurs on a timescale ranging from hours to several days. The optimal placement stage - where the decision of how the slices of a large order will be placed is made, this involves deciding on what types of orders will be used, their size and the exchange to place the order on - this occurs on a time scale of fractions of seconds to minutes.

The optimal execution and placement problems are the focus of the following literature: Bertsimas and Lo, 1998; Almgren and Chriss, 2001; Alfonsi, Fruth, and Schied, 2010 focus on solving the optimal execution problem. Where as Cont and Kukanov, 2017; Guo, De Larrard, and Ruan, 2017; Figueroa-López, Lee, and Pasupathy, 2018 provide solutions to the optimal placement problem.

The possibility of testing optimal execution and placement strategies is a motivating application for the simulator we have proposed. In this chapter, we detail how the simulator can be used for the testing of optimal placement strategies by using it to test the solution presented in Cont and Kukanov, 2017.

7.1 The Optimal Placement Problem

7.1.1 The Objective Function

The following is a summary of the optimal placement problem and its solution as described in Cont and Kukanov, 2017, for the case where the asset is traded on a single exchange.

Suppose we wish to purchase S units of a given asset on the time interval $[t, t + dt]$. When placing the order we can choose between two order types, a market order or a limit order placed at the time t best $b(t)$. The optimal placement problem is to determine the optimal value of the decision variable $X = (M, L)$ where M is the amount that should be placed as a market order and L is the amount that should be placed as a limit order at the best.

Let $Q := V(b(t), t)$ be the length of the queue at best that order L will join. Let ξ be the total outflow from the front of this queue during time $[t, t + dt]$. At $t + dt$ the filled amount of the limit order for amount L is thus

$$\max(\xi - Q, 0) - \max(\xi - Q - L, 0)$$

Define $A(X, \xi)$ as the total amount of shares bought during $[t, t + dt]$, thus

$$A(X, \xi) = M + \max(\xi - Q, 0) - \max(\xi - Q - L, 0)$$

ξ is a random variable, i.e. it is not known at time t when X must be chosen, thus $A(X, \xi)$ is also unknown at t. We assume that if $A(X, \xi) < S$ the trader submits a market order for $S - A(X, \xi)$ at time $t + dt$ in order to reach their target.

Market impact is assumed to be linear in both the volume of limit and market orders, with the same slope θ for both order types. Cont and Kukanov, 2017 specify the optimal placement problem as,

$$\min_X \mathbb{E}[\nu(X, \xi)] \tag{7.1}$$

where the cost function $\nu(X, \xi)$ is given by,

$$\begin{aligned} \nu(X, \xi) =& (h + f)M - (h + r)[\max(\xi - Q, 0) - \max(\xi - Q - L, 0)] \\ &+ \theta[M + L + \max(S - A(X, \xi), 0)] + \lambda_u[\max(S - A(X, \xi), 0)] \\ &+ \lambda_o[\max(A(X, \xi) - S, 0)] \end{aligned} \tag{7.2}$$

where $h = \frac{1}{2}S(t)$ is the half spread at the beginning of the interval. f, r are respectively the market order fee and the limit order rebate; these are set by the exchange.

The $\theta[M + L + \max(S - A(X, \xi), 0)]$ term represents the cost due to market impact and the term $\lambda_u[\max(S - A(X, \xi), 0)] + \lambda_o[\max(A(X, \xi) - S, 0)]$ is a penalty term which penalizes under or over-filling the required trade size S over the interval. Where $\lambda_u \geq 0, \lambda_o \geq 0$ are the associated marginal penalties in dollars per share.

7.1.2 Solution

Cont and Kukanov, 2017 show that if the following assumptions hold,
A1) $r + h > 0$
A2) $\lambda_o > h + r, \lambda_o > -(h + f)$
A3) $\lambda_u > h + f$

Then the following is a global minimizers for 7.1,
Denote by F, F^{-1} the density and quantile functions respectively of the random variable ξ. let $\underline{\lambda}_u = \frac{2h+f+r}{F(Q+S)} - (h + r + \theta)$ and $\overline{\lambda}_u = \frac{2h+f+r}{F(Q)} - (h + r + \theta)$,

$$(M^*, L^*) = \begin{cases} (0, S) & \lambda_u \leq \underline{\lambda}_u \\ (S, 0) & \lambda_u \geq \overline{\lambda}_u \\ \left(S - F^{-1}\left(\frac{2h+f+r}{\lambda_u+h+r+\theta}\right) + Q, F^{-1}\left(\frac{2h+f+r}{\lambda_u+h+r+\theta}\right) - Q\right) & \lambda_u \in (\overline{\lambda}_u, \underline{\lambda}_u) \end{cases} \tag{7.3}$$

7.2 Simulation Results

We use our simulated limit order book to test solution 7.3 to the optimal placement problem. We also test two benchmark strategies for comparison.

Benchmarks The market order benchmark strategy submits the entire volume S as a market order at time t. The limit order benchmark strategy submits the entire volume S as a limit order placed at best at time t, it then submits a market order for any unfilled amount at $t + dt$.

Simulation Parameters For the purposes of the simulation, we use the following parameters: we set the market order fee to $f = 0.29$ cents per share and the limit order rebate to $r = 0.25$ cents per share, these are the same parameters used in Cont and Kukanov, 2017. In their formulation Cont and Kukanov, 2017 assume a constant half spread h and require $\lambda_u > h + f$. The spread however, is endogenous in our model so we set $\lambda_u = h + f + c$ where $h = \frac{1}{2}s(t)$ is the half spread at the start of the time interval. The parameter c effectively becomes the user specified penalty parameter. In our tests we set c to be very small (we used 2.22×10^{-16} our machine epsilon) to enable the optimal placement solution to achieve its lowest possible execution costs.

We will test the optimal placement solution under three different market conditions. We generate the three market conditions by varying the model parameter β (average proportion of the top of the book captured by incoming market orders). For the other model parameters, we use the estimated values in chapter 4 for all three market conditions. Our three market conditions are:

$\beta = \beta_{\text{low}} = 0.0500$ - This market exhibits low volatility, large queues form at the top of the book.

$\beta = \beta_{\text{mid}} = 0.5896$ - This market exhibits moderate volatility and is prone to occasional explosions in the spread. This is the β value estimated from the data and the market studied in 5.

$\beta = \beta_{\text{high}} = 0.9000$ - This market exhibits high volatility, frequent explosions in the spread occur and smaller queues at the top of the book.

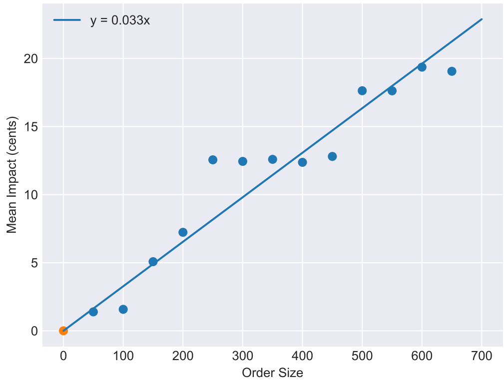
Quantity	**Value**
Mid-Price	58640.5 cents
Best Bid	58631.0 cents
Best Ask	58650.0 cents
Spread	19.0 cents
Volume at Best Bid	100 shares
Volume at Best Ask	18 shares
Buy Side Volume	19042 shares
Sell Side Volume	19430 shares

TABLE 7.1: Quantities of Interest in AAPL-50 - row 5000

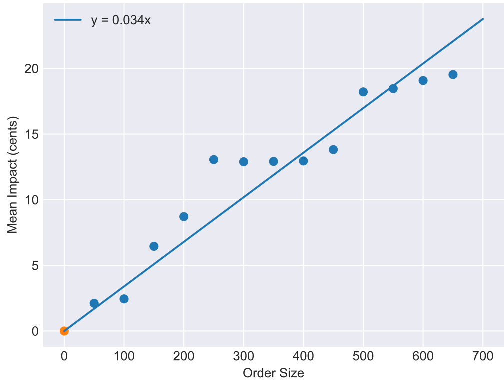
Initial Condition As the initial condition in the placement tests we used the 5000th row of the AAPL-50 LOBSTER dataset. We use this initial condition because we wish to test placement later on in the trading day when spreads have narrowed compared to the open and greater volumes are available on the LOB. LOB quantities of interest at this initial condition are given in table 7.1. We estimate the market impact coefficients $\theta_{\text{low}}, \theta_{\text{mid}}, \theta_{\text{high}}$ using the same procedure described in section 5.3. The solution presented in Cont and Kukanov, 2017 assumes linear market impact at the 1 minute time scale. Figure 7.1 shows mean market impact at different order sizes for the three different market conditions as well as the estimated linear market impact curve (estimation is done assuming no intercept term as is assumed in Cont and Kukanov, 2017).

Parameter	**Estimate**
θ_{low}	0.033
θ_{mid}	0.034
θ_{high}	0.038

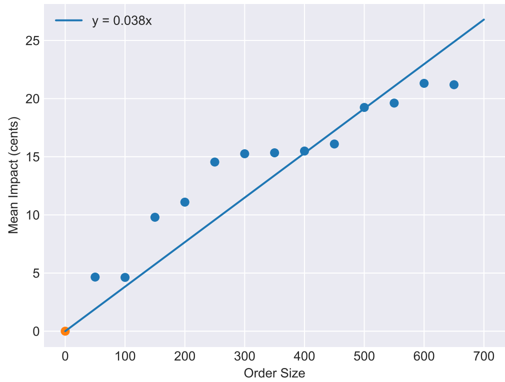
TABLE 7.2: Estimated Market Order Impact Coefficients



(A) $\beta = 0.0500$



(B) $\beta = 0.5896$



(C) $\beta = 0.9000$

FIGURE 7.1: Market Order Impact - 1 minute after order arrival, each point is an average of 10000 simulations

Table 7.2 shows the estimated market impact coefficients for our model. The impact coefficient appears to increase with the model parameter β. Our model had a lower impact coefficient than the $\theta = 0.05$ used in Cont and Kukanov, 2017.

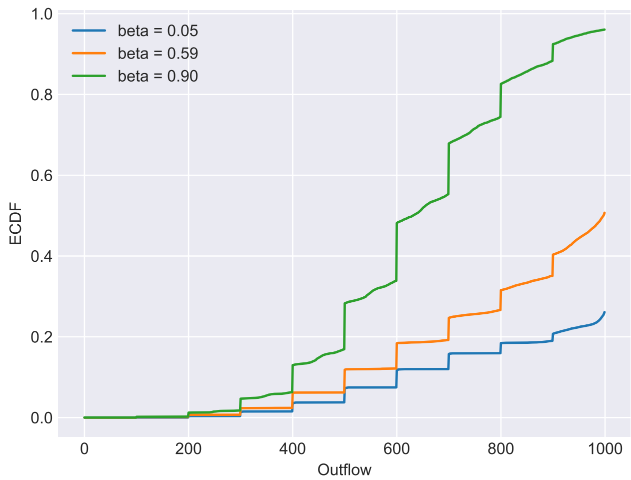


FIGURE 7.2: Conditional Empirical Cumulative Distributions Functions (ECDF) of Outflows - ECDF of $\xi | Q = 1000$, each ECDF was estimated using 10000 realizations of $\xi | Q = 1000$

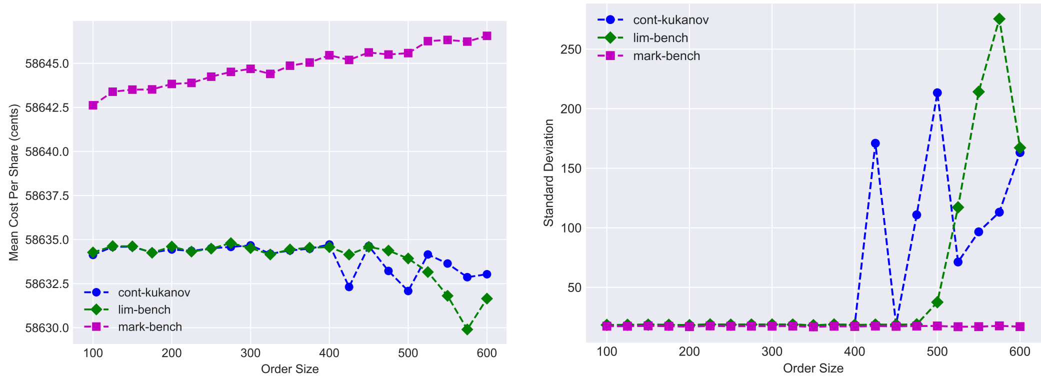
Figure 7.2 shows the empirical distributions of the outflows for the three market conditions. Given the dependence of incoming market order size on the size of the queue at best in our model, we give the distribution conditional on the size of the queue at best.

Results Table 7.3 shows the average cost per share and the standard deviation of the optimal placement solution and the two benchmarks in each of the three market conditions. Each method bought 1000 shares in 60 seconds. 50 000 Simulations where used in each case. The time interval on which the placement was done began at the initial condition described in table 7.1 (i.e. orders are being placed on [0, 60] in seconds). Hence, there is no variance in the achieved market order price. In all three market conditions, the Market Order Benchmark is the most expensive. There is no meaningful difference between the performance of the optimal solution and the Limit Order Benchmark.

	Average Cost Per Share	**Standard Deviation**
$\beta = 0.0500$		
Cont-Kukanov Solution	58636.9688	12.9756
Benchmark - Market Order	58680.6170	0
Benchmark - Limit Order	58636.9148	12.8984
$\beta = 0.5896$		
Cont-Kukanov Solution	58638.5517	10.4895
Benchmark - Market Order	58680.6170	0
Benchmark - Limit Order	58638.5800	10.4829
$\beta = 0.9000$		
Cont-Kukanov Solution	58645.4642	7.3163
Benchmark - Market Order	58680.6170	0
Benchmark - Limit Order	58645.4147	7.3196

TABLE 7.3: Placement Test Results - S = 1000 with $dt = 60$, $t = 0$

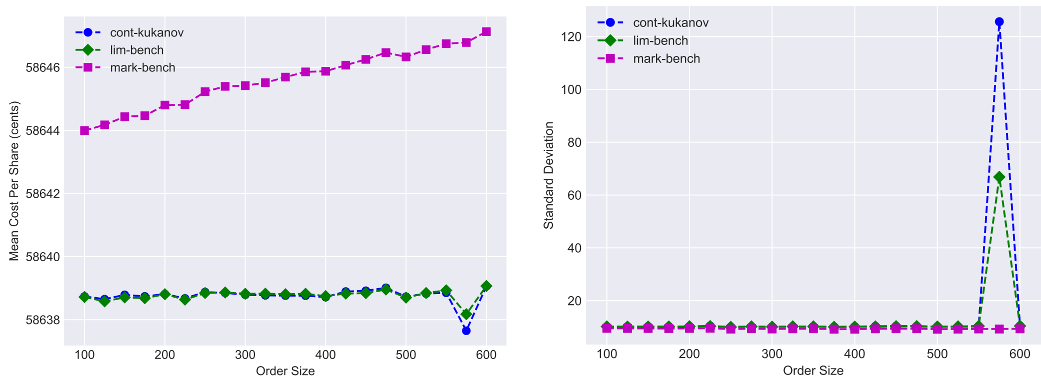
Figures 7.3, 7.4 and 7.5 show the average cost per share and standard deviation of each placement method for various order sizes in each of the three market conditions. Each point is an average over 10000 simulations. The shares are bought on the time interval [60, 120] (in seconds).



(A) Mean Cost Per Share (cents) (B) Standard Deviation

FIGURE 7.3: Placement Test Results - $\beta = 0.05$, $dt = 60$, $t = 50$

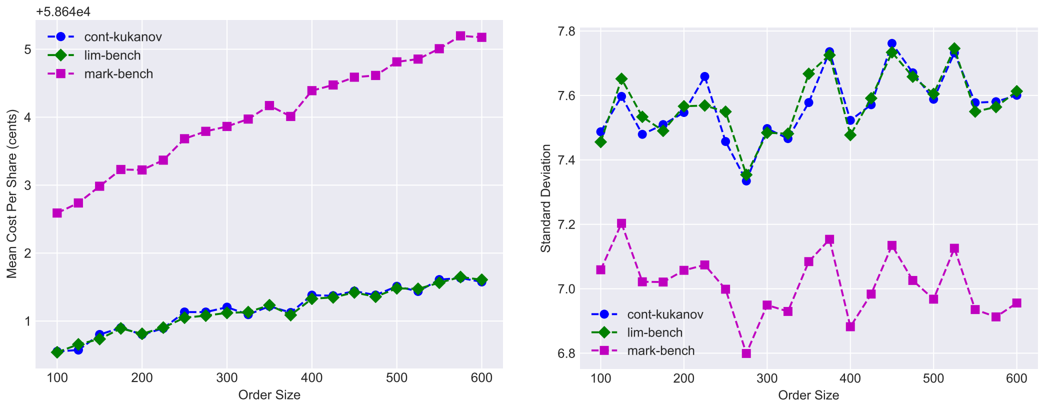
In the low volatility market condition, figure 7.3 we observe that both the optimal solution and the limit order benchmark achieve significantly lower costs than the market order benchmark. The market order benchmarks cost per share is increasing in order size where as the optimal solution and the limit order benchmark are not. Although we see decreased costs at larger order sizes(>400) for the optimal solution and the limit order benchmark, these points are very noisy and correspond with large spikes in the standard deviation. We see no real advantage of the optimal solution over the limit order benchmark in this case.



(A) Mean Cost Per Share (cents) (B) Standard Deviation

FIGURE 7.4: Placement Test Results - $\beta = 0.5896$, $dt = 60$, $t = t0$

Observations in the moderate volatility market condition, figure 7.4 are similar to those observed in the low volatility market condition. There are, however, fewer spikes in standard deviation.



(A) Mean Cost Per Share (cents) (B) Standard Deviation

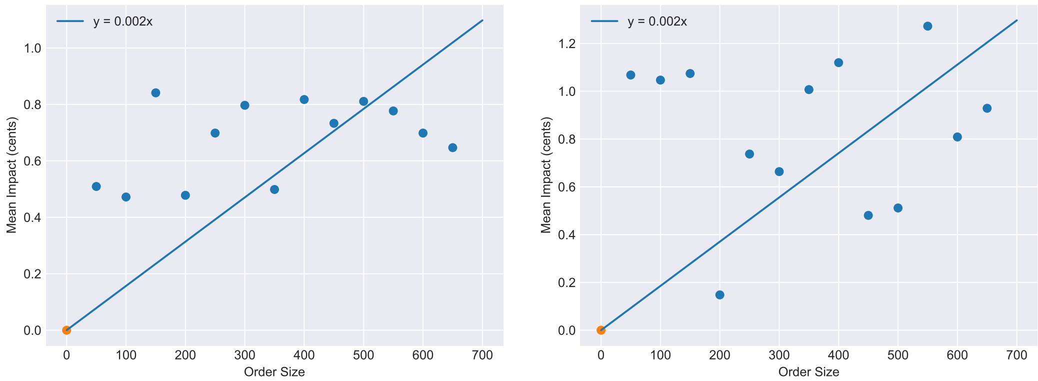
FIGURE 7.5: Placement Test Results - $\beta = 0.9000$, $dt = 60$, $t = 60$

In the low volatility market condition, figure 7.5, the optimal solution and the limit order benchmark again achieve significantly lower costs than the market order benchmark. However, in this condition their mean cost per share is increasing in the order size. No spikes in their standard deviation are observed at higher order sizes.

Discussion Tests in our simulated market did not reveal any advantage of the optimal placement solution presented in Cont and Kukanov, 2017 over the limit order benchmark strategy which submits the entire volume S as a limit order placed at best at time t, it then submits a market order for any unfilled amount at $t + dt$.

A possible reason for this is that the optimal placement solution assumes that market orders and limit orders have equal impact. This is not the case at short time scales within our simulated market. In the moderate volatility case, we observed a 1 minute market order impact coefficient of 0.034. Whereas for limit orders over the same time period we observed in impact coefficient of 0.002. This difference is due to the fact that in our model, market orders exert impact through two mechanisms: firstly, they alter the order flow by ensuring that the next arriving market order will have the same sign with probability π and secondly, they remove liquidity from the top of the book which increases the probability of a market order depleting the queue at best. The former dominates at short time scales and ensures that even small market orders have significant impact, it also results in market orders with a surprising trade sign having greater impact than market orders with an unsurprising trade sign. This effect has been observed empirically in real data (Bouchaud, Farmer, and Lillo, 2009). Limit orders in our model exert impact only through the latter mechanism and thus at short time scales have impact which is comparatively smaller with no significant size dependence. Though at longer timescales there appears to be a dependence on the order size (figure 7.6b).

Linear market impact of market orders at the 1 minute time scale seems a plausible assumption in our simulated market (figure 7.1). However, at short time scales market impact is not linear in our simulated market (figure 7.6a).



(A) 1 minute from order placement, each point is an average over 10000 simulations

(B) 10 minutes from order placement, each point is an average over 2000 simulations

FIGURE 7.6: Buy Limit Order Impact - $\beta = 0.5896$

Chapter 8

Conclusion

We have proposed a computationally tractable framework for use in the testing of optimal execution and placement strategies. Our framework lends itself well to customisation through its modular approach to model design. Within the context of this framework, we have presented a model that reasonably reproduces important stylised facts of financial time series. We have shown market impact within the simulated market and the effect of the model parameters on simulated volatility. Finally, we used the model to test an optimal placement strategy.

8.1 Further Challenges

8.1.1 Improvements in Implementation

Improved implementations of the framework and the model we have described may be explored (see Appendix A for implementation details). This could be done, for example, through the use of better algorithms or by incorporating parallel processing. Improved run times would improve the simulators capacity as a Monte Carlo tool. Studies of order flow dynamics in the simulated market, such as the time evolution of market impact (section 5.3), could be conducted for longer time periods whilst obtaining more less noisy and thus more meaningful results.

8.1.2 Simulation of Multiple Markets of The Same Asset

Many financial instruments are traded on multiple exchanges. Extending this framework to one that allows for the parallel simulation of multiple markets of the same asset would allow for the testing of optimal placement solutions in the case of multiple exchanges (or smart order routing).

8.1.3 Further Complications to the Model

Our presented model shows that there is sufficient flexibility in our framework to produce a model that is able to reproduce some of the properties of real limit order books. More complex models than the one presented here may be produced under the proposed framework. For example, models that incorporate time dependencies, more complex dependencies amongst the order components and more complex dependencies between the order components and the history or state of the limit order book.

8.1.4 Testing Other Intended Use Cases

In this thesis, we have used the proposed framework and model to test an optimal placement solution. Further use cases such as the testing of optimal execution algorithms, market making algorithms and other HFT strategies could be explored.

8.1.5 Short Term Price Prediction

It is worth exploring whether a model can be constructed under this framework that may, through Monte Carlo, be used to predict price movements in short time periods to satisfactory levels of accuracy.

Bibliography

Alfonsi, Aurélien, Antje Fruth, and Alexander Schied (2010). "Optimal execution strategies in limit order books with general shape functions". In: *Quantitative finance* 10.2, pp. 143–157.

Almgren, Robert (2020). *Transaction Cost Analysis for Futures Trading*. URL: https://www.maths.ox.ac.uk/system/files/media/Slides%20Oxford-Jan2020.pdf.

Almgren, Robert and Neil Chriss (2001). "Optimal execution of portfolio transactions". In: *Journal of Risk* 3, pp. 5–40.

Bacry, Emmanuel et al. (2017). *Tick: a Python library for statistical learning, with a particular emphasis on time-dependent modelling*. arXiv: 1707.03003 [stat.ML].

Bertsimas, Dimitris and Andrew W Lo (1998). "Optimal control of execution costs". In: *Journal of Financial Markets* 1.1, pp. 1–50.

Bouchaud, Jean-Philippe, J Doyne Farmer, and Fabrizio Lillo (2009). "How markets slowly digest changes in supply and demand". In: *Handbook of financial markets: dynamics and evolution*. Elsevier, pp. 57–160.

Bouchaud, Jean-Philippe, Marc Mézard, Marc Potters, et al. (2002). "Statistical properties of stock order books: empirical results and models". In: *Quantitative finance* 2.4, pp. 251–256.

Bouchaud, Jean-Philippe et al. (2018). *Trades, quotes and prices: financial markets under the microscope*. Cambridge University Press.

Chakraborti, Anirban et al. (2011). "Econophysics review: I. Empirical facts". In: *Quantitative Finance* 11.7, pp. 991–1012.

Cont, Rama and Adrien De Larrard (2012). "Order book dynamics in liquid markets: limit theorems and diffusion approximations". In: *Available at SSRN 1757861*.

Cont, Rama and Arseniy Kukanov (2017). "Optimal order placement in limit order markets". In: *Quantitative Finance* 17.1, pp. 21–39.

Cont, Rama and Marvin S Mueller (2019). "A stochastic PDE model for limit order book dynamics". In: *arXiv preprint arXiv:1904.03058*.

Cont, Rama, Sasha Stoikov, and Rishi Talreja (2010). "A stochastic model for order book dynamics". In: *Operations research* 58.3, pp. 549–563.

Daniels, Marcus G et al. (2003). "Quantitative model of price diffusion and market friction based on trading as a mechanistic random process". In: *Physical review letters* 90.10, p. 108102.

Engelbrecht-Wiggans, Richard (1980). "State of the art—auctions and bidding models: a survey". In: *Management Science* 26.2, pp. 119–142.

Farmer, J Doyne, Paolo Patelli, and Ilija I Zovko (2005). "The predictive power of zero intelligence in financial markets". In: *Proceedings of the National Academy of Sciences* 102.6, pp. 2254–2259.

Farmer, J Doyne et al. (2004). "What really causes large price changes?" In: *Quantitative Finance* 4.4, pp. 383–397.

Figueroa-López, José E, Hyoeun Lee, and Raghu Pasupathy (2018). "Optimal placement of a small order in a diffusive limit order book". In: *High Frequency* 1.2, pp. 87–116.

Foucault, Thierry, Ohad Kadan, and Eugene Kandel (2005). "Limit order book as a market for liquidity". In: *The review of financial studies* 18.4, pp. 1171–1217.

Gould, Martin D et al. (2013). "Limit order books". In: *Quantitative Finance* 13.11, pp. 1709–1742.

Guo, Xin, Adrien De Larrard, and Zhao Ruan (2017). "Optimal placement in a limit order book: an analytical approach". In: *Mathematics and Financial Economics* 11.2, pp. 189–213.

Hambly, Ben, Jasdeep Kalsi, and James Newbury (2020). "Limit order books, diffusion approximations and reflected SPDEs: from microscopic to macroscopic models". In: *Applied Mathematical Finance*, pp. 1–39.

Huang, Ruihong and Tomas Polak (2011). "LOBSTER: Limit order book reconstruction system". In: *Available at SSRN 1977207*. URL: https://lobsterdata.com/.

Laub, Patrick J., Thomas Taimre, and Philip K. Pollett (2015). *Hawkes Processes*. arXiv: 1507.02822 [math.PR].

Madhavan, Ananth (1992). "Trading mechanisms in securities markets". In: *the Journal of Finance* 47.2, pp. 607–641.

Mendelson, Haim (1982). "Market Behavior in a Clearing House". In: *Econometrica* 50.6, pp. 1505–1524. ISSN: 00129682, 14680262. URL: http://www.jstor.org/stable/1913393.

Parlour, Christine A (1998). "Price dynamics in limit order markets". In: *The Review of Financial Studies* 11.4, pp. 789–816.

Roşu, Ioanid (2009). "A dynamic model of the limit order book". In: *The Review of Financial Studies* 22.11, pp. 4601–4641.

Servén, Daniel and Charlie Brummitt (2018). *pyGAM: Generalized Additive Models in Python*. DOI: 10.5281/zenodo.1208723. URL: https://doi.org/10.5281/zenodo.1208723.

Skellam, J. G. (1946). "The Frequency Distribution of the Difference Between Two Poisson Variates Belonging to Different Populations". In: *Journal of the Royal Statistical Society* 109.3, pp. 296–296. ISSN: 09528385. URL: http://www.jstor.org/stable/2981372.

Smith, Eric et al. (2003). "Statistical theory of the continuous double auction". In: *Quantitative finance* 3.6, pp. 481–514.

Virtanen, Pauli et al. (2020). "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python". In: *Nature Methods* 17, pp. 261–272. DOI: https://doi.org/10.1038/s41592-019-0686-2.

Wood, Simon N (2017). *Generalized additive models: an introduction with R*. CRC press.

Appendix A

Implementation Details

We describe here the computational details of our implementation of the proposed framework. All source code for this thesis can be found at the following repository:

https://github.com/TakuMtombeni/Limit-Order-Book-Simulator-public.git

A.1 Simulation Framework

A.1.1 Algorithm

Algorithm 1 Sequential LOB Simulation

Require: model : The simulation model, an object with members `gen_arrival_times`, `gen_market_order`, `gen_limit_order`, `gen_cancellation`

Require: lob : The initial LOB state, an object with members `submit_market_order`, `submit_limit_order`, `cancel_limit_order`

```
function SIMULATE(model, lob, num_sample_paths, obs_freq, run_time)
    for k ← 1 to num_sample_paths do
        event_times ← model.gen_arrival_times(obs_freq, run_time)
        for event in event_times do
            if event.type is 'limit_order_arrival' then
                order = model.gen_limit_order(lob)
                lob.submit_limit_order(order, event.time)
            else if event.type is 'limit_order_cancellation' then
                cancel_id = model.gen_cancellation(lob)
                lob.cancel_limit_order(cancel_id, event.time)
            else if event.type is 'market_order_arrival' then
                order = model.gen_market_order(lob)
                lob.submit_market_order(order, event.time)
            else if event.type is 'observation_time' then
                midprice_process.append(lob.mid_price())
                spread_process.append(lob.spread())
                order_imbalance_process.append(lob.order_imbalance())
        output[k] ← (midprice_process, spread_process, order_imbalance_process)
    return output
```

A.1.2 Performance

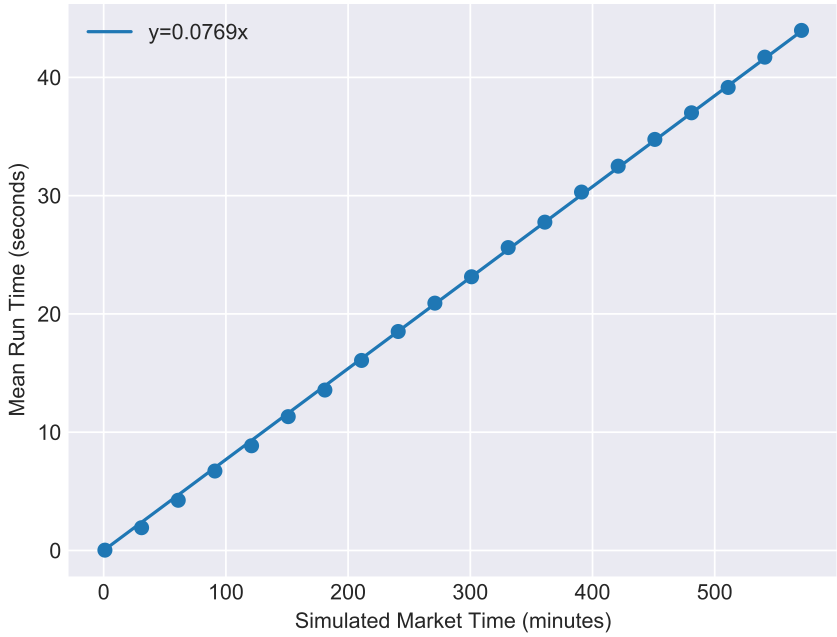


FIGURE A.1: Simulation Run Time - this figure shows the recorded runtime of our Algorithm 1 implementation. Each point is an average over 100 simulations. The algorithm appears to scale linearly in time with scaling coefficient 0.0769