Short-term Stock Price Prediction Based on Limit Order Book Dynamics

AN, Yang

A Thesis Submitted in Partial Fulfillment
of the Requirements for the Degree of
Doctor of Philosophy
in
Statistics

The Chinese University of Hong Kong
July 2015

Thesis Assessment Committee

Professor FAN, Xiaodan (Chair)

Professor CHAN, Ngai Hang (Thesis Supervisor)

Professor YAU, Chun Yip (Committee Member)

Professor CHEN, Kani (External Examiner)

Abstract

Interaction of capital market participants is a complicated dynamic process. A stochastic model is proposed to describe the dynamics to predict short-term stock price behaviors. Independent compound Poisson processes are introduced to describe the occurrences of market orders, limit orders, and cancellations of limit orders, respectively. Based on high-frequency observations of the limit order book, the modified maximum empirical likelihood estimator (MELE) is applied to estimate the parameters of the compound Poisson processes. A simulation study is conducted to demonstrate the accuracy and efficiency of the estimation procedure. Moreover, an analytical formula is derived to compute the probability distribution of the first-passage time of a compound Poisson process. Based on this formula, the conditional probability of price increase and the conditional distribution of the duration until the first change in mid-price are obtained. Finally, a novel approach of short-term stock price prediction is proposed and this methodology works reasonably well in the data analysis of Intel (INTC).

摘要

资本市场参与者的行为是一个极其复杂的动态过程。本文提出的随机模型可以用来描述这种动态变化，从而对短期股票价格的走势进行预测。这个模型包含了多个独立的复合泊松过程，用以描述限价单(limit orders)和市价单(market orders)的发生和取消情况。在参数估计方面，我们对原始的最大经验似然估计(MELE)进行了改进，以更好地适应当前的问题，并在模拟研究中展示了这种估计方法的准确及稳定性。同时，我们推导出了复合泊松过程的首达时间分布。依据这一公式，我们可以获得股票价格上升的条件概率，以及下一次股价变化时间的条件分布。最后，我们提出了一种新的股价预测方法，这种方法在实际数据分析中很好地展示了其有效性。

Acknowledgments

First and foremost, I would like to express my sincere gratitude to my supervisor, Professor Ngai Hang Chan, for his guidance, patience, enthusiasm and encouragement during the course of this study program. Without his continuous support, I could not have accomplished so much in my Ph.D. study. Besides my supervisor, I am also very grateful to the rest of my defense committee: Professor Xiao Dan Fan, Professor Chun Yip Yau, and Professor Kani Chen for their engagement and insightful comments. I would also like to thank Professor Ming Gao Gu, Professor Xin Yuan Song, and Professor Siu Hung Cheung for their teaching and help in many aspects. Last but not the least, my deepest thanks to my family and friends for being so supportive and considerate.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Limit Order Book Dynamics

Studies of financial trading have initially concentrated on *quote-driven* markets, where traders transact with dealers (market makers). Dealers maintain an inventory of financial assets, post bid and ask prices, and are required to transact at their quotes. The existence of dealers provides liquidity, while transparency in a quote-driven market is relatively low.

In recent years, developments of electronic communications networks (ECNs) have provided an alternative *order-driven* trading system, where traders directly transact with other traders and there is no intermediary dealer. There are mainly two types of order-driven markets: (1) continuous order-driven markets, where trades are executed continuously throughout the trading days; (2) periodic call auctions, where orders are gathered together to be executed simultaneously at predetermined times.

In either type of an order-driven electronic platform, traders post two kinds of orders. A *limit* order is an order to trade at a price level better than or equal to the limit price. In the case when market prices do not move to the limit, the trade will not be executed, so it has *execution uncertainty*. A *market* order is an order to complete the trade immediately at the best possible price. The emphasis in a market order is the speed of execution, but there is

price uncertainty. Other than submitting new limit orders and market orders, traders can also cancel the limit orders they have already posted. Outstanding limit orders are aggregated in a *limit order book* (LOB), which is available for all market participants. Figure 1.1 gives some examples of LOBs from different companies. A limit order stays in the order book until it is either executed against a market order or it is canceled, when the quantities available in the limit order book are updated accordingly.

This order-driven trading system has more competition and results in better prices, but it also involves more complexity. Unlike in a quote-driven market, where modeling the behaviors of a few market makers is sufficient, there are thousands of anonymous traders in an order-driven market. These traders arrive randomly, choose whether they want to execute immediately, determine how large their order sizes should be, and even cancel their orders at any time to exit the market strategically. Figure 1.2 gives an example of the dynamics of a limit order book driven by a limit order. The vertical line represents price and each square represents one unit of outstanding orders. As is shown in Figure 1.2(b), with an incoming limit sell order inside the spread, the best ask price becomes one price tick lower and the shape of the order book changes as well. In addition to limit orders, the evolution of an order book is also driven by incoming market orders and cancellations of limit orders.

Despite of its complexity, various research has been focusing on order book dynamics, because it is an issue of great importance given the popularity of the limit order book as a form of security market organization. Models of order book dynamics not only can be applied to optimize trade execution strategies (Alfonsi *et al.* (2010), Obizhaeva and Wang (2006), Predoiu *et al.* (2011), Guilbaud and Pham (2013), Goettler *et al.* (2005, 2009)), but these models can also provide insight into the relationship between supply and demand (Farmer *et al.* (2004), Foucault *et al.* (2005)). Empirical studies indicate that the formation of short-term price behavior depends on the evolution of limit

order books (Parlour (1998), Harris and Panchapagesan (2005), Rosu (2009)).

Statistical features of limit order book dynamics are studied in various empirical studies. Bouchaud *et al.* (2002) empirically investigate some interesting features concerning the statistics of incoming limit order prices and the humped shape of the average order book. Smith *et al.* (2003) and Gourieroux *et al.* (1999) provide an extensive list of statistical features of limit order book dynamics. However, it remains a challenging task to capture essential features into a single model. Cont and de Larrard (2011) propose a Markovian model of a limit order market, which captures certain main features of market orders and limit orders and their influence on price dynamics, but they only focus on the best bid and ask queues rather than the dynamics of the entire limit order book. Bouchaud *et al.* (2008), Bovier *et al.* (2006), Luckock (2003), Maslov and Mills (2001) propose stochastic models of order books, but concentrate only on unconditional/steady-state distributions of various quantities. In this thesis, probabilities conditional on the current state of the limit order book are investigated instead. Cont *et al.* (2010) propose a model that tracks the number of limit orders at each price level in the limit order book. These limit orders wait in a queue to be executed against market orders or to be canceled. Furthermore, they assume that the occurrences of market events – incoming limit orders, incoming market orders and cancellations of limit orders, follow independent Poisson processes.

To be specific, in Cont *et al.* (2010) model, limit orders are placed on a price grid $\{1, \ldots, n\}$, which denotes multiples of a price tick. The boundary n is selected to be large enough such that it is highly unlikely for the stock price in question to rise to a level higher than n within the time frame concerned. The state of the limit order book is described by a continuous-time Markov process $X(t) = (X_1(t), \ldots, X_n(t))_{t \geq 0}$, where $|X_p(t)|$ is the amount of outstanding limit orders at price $p, 1 \leq p \leq n$. If the outstanding orders at price p are buy orders, then $X_p(t)$ is negative; otherwise, $X_p(t)$ is positive.

The *ask price* $p_A(t)$ at time t is defined by

$$p_A(t) = \inf\{p = 1, \ldots, n,\ X_p(t) > 0\} \wedge (n+1).$$

Similarly, the *bid price* $p_B(t)$ at time t is defined by

$$p_B(t) = \sup\{p = 1, \ldots, n,\ X_p(t) < 0\} \vee 0.$$

According to the above definition, the ask price becomes $n+1$ when there are no ask orders in the limit order book, and the bid price becomes 0 when there are no bid orders in the limit book.

Moreover, since most of the trading activities happen within a certain distance from the current bid and ask prices, it is necessary to keep track of the amount of outstanding orders in the vicinity of the bid/ask. The number of outstanding buy orders at a distance i from the ask is defined by

$$Q_i^B(t) = \begin{cases} X_{p_A(t)-i}(t) & 0 < i < p_A(t), \\ 0 & p_A(t) \leq i < n, \end{cases}$$

and the number of outstanding sell orders at a distance i from the bid is defined by

$$Q_i^A(t) = \begin{cases} X_{p_B(t)+i}(t) & 0 < i < n - p_B(t), \\ 0 & n - p_B(t) \leq i < n. \end{cases}$$

The representation $(p_A(t), p_B(t), Q^A(t), Q^B(t))$ contains the same information as $X(t)$, but it highlights the shape of a limit order book relative to the best available quotes.

The evolution of an order book is driven by the incoming flow of limit orders, market orders, and cancellations of limit orders. To capture certain previously observed features in a limit order book, Cont *et al.* (2010) propose a stochastic model where the events outlined above are described using independent Poisson processes. The intensity parameter related to each process models the occurrence frequency of each market event and the size of each event

is assumed to be unit. Although the formulation of the model leads to an analytically tractable framework where parameters can be estimated and various conditional probabilities may be computed through Laplace transforms, they point out that heterogeneity of order sizes and correlation of order flows need to be incorporated for future exploration.

Zhao (2010) and Toke (2011) extend Cont *et al.* (2010) model via considering the interdependence of market event arrival rates. Zhao (2010) models the arrival rates of all market events by a Hawkes process, which is a point process with time-varying intensity parameter. In this way, time clustering effects of order flows are produced, where periods of high and low arrival rates are grouped together separately. Unlike Zhao (2010), Toke (2011) proposes mutually (asymmetrically) exciting Hawkes processes and each process is introduced for each type of market event. However, models of random order sizes are still lacking.

In this work, a new stochastic model is proposed that describes the limit order book dynamics by assuming the order flows are governed by independent compound Poisson processes. This model considers both the arrival frequency of each order and the heterogeneity of order sizes, which is more realistic and reflects the complexities in the aforementioned order-driven market. The model also enables a wide range of pricing quantities to be computed and the dynamic shape of a limit order book to be predicted.

1.2 Stock Price Prediction

Stock price forecasting is a widely studied topic in various fields. There are traditionally two main techniques to analyze stocks – fundamental analysis and technical analysis.

1.2.1 Fundamental Analysis

Fundamental analysts attempt to study both macroeconomic factors (like overall economy and industry conditions) and company-specific factors (like internal management and financial conditions) with the goal of producing a fundamental value which investors can compare with the current stock price. There are typically two categories of valuation models: absolute and relative valuation models.

Absolute valuation models focus on fundamentals such as dividends, free cash flows, and residue incomes of a single company and do not get any other companies involved. Valuation models that fall into this category include the dividend discount models (DDM), discounted free cash flow models, residual income models and asset-based models. Take the DDM, which William (1938) propose, as an example. Suppose that the holding period is extended indefinitely, then the fundamental value of a given security at time 0 is

$$V_0 = \sum_{t=1}^{\infty} \frac{D_t}{(1+r)^t}, \tag{1.1}$$

where D_t is the dividend expected to receive at the end of the time period t and r is the risk-free rate. Although the DDM is theoretically correct, its application in practice requires accurate dividend forecasts for many periods, which is a task we can rarely expect because of insufficient information. Thus the following growth models are developed:

(i) Gordon growth model (GGM). It assumes that the company grows at a constant rate g indefinitely (see Figure 1.3(a)). Then Equation (1.1) reduces to a simple formula

$$V_0 = \frac{D_0(1+g)}{r-g} = \frac{D_1}{r-g}; \tag{1.2}$$

(ii) Two-stage growth model. It assumes that dividends increase at a relatively high rate g_1 for a short period of time, and then the company

reverts to a long-run lower growth rate perpetually (see Figure 1.3(b));

(iii) H-model. The unrealistic aspect of the basic two-stage DDM is that it assumes that the company growth rate falls immediately back to its long-run level at the beginning of the second stage. The H-model solves this problem and utilizes a more realistic assumption: the growth rate starts out high at g_1 level and then declines linearly over the first stage until it reaches the lower level g_2 (see Figure 1.3(c));

(iv) Three-stage DDM. It assumes that the company experiences three distinct stages of earnings growth (see Figure 1.3(d)).

The fundamental value V_0 is compared with the market price of a stock. If a stock is trading at a price higher than its fundamental value (the value implied by one of the absolute valuation models), the stock is considered to be *overvalued*. Similarly, if the stock price is lower than the model price, the stock is considered to be *undervalued*, and if the market price and the model price are equal, the stock is considered to be *fairly valued*.

In contrast to absolute valuation models, relative valuation models compare the company concerned with other similar companies and various price multiples (such as price-to-earnings ratio, price-to-book ratio, price-to-sales ratio, price-to-cash flow ratio and dividend yield) are used for comparison. The rationale for this method is that two similar assets should sell at comparable price multiples under Law of One Price. For instance, if the P/E (price-to-earnings ratio) of the firm in question is higher than the average P/E of other comparable companies, the firm is considered to be overvalued. Similarly, "undervalued" and "fairly valued" can be defined.

Fundamental analysis involves lots of research to investigate the true value of a company, and sometimes enables investors to identify potential or bad companies before the general market. It is clear that fundamental analysis alone works well in the long term, but it may not provide enough basis for

short-term trading. Analysis of intrinsic value is one thing, but trading relies on the behaviors of all market participants, and fundamental analysis has little concern about this. Even if fundamental analysis suggests a company is undervalued, it does not guarantee that the price will increase soon, and a trader can even lose by longing the stock if he closes his position before seeing any gains. Technical analysis, on the other hand, deals with market actions without much concern about the underlying causes. This is discussed in the following section.

1.2.2 Technical Analysis

Technical analysis analyzes stock price trends based on past market data, primarily price and volume (Caginalp and Laurent (1998), Brock *et al.* (1992)). The key assumptions underlying technical analysis include:

(i) market prices reflect behaviors of both rational and irrational market participants;

(ii) the actual changes in supply and demand can be observed in market prices, although the causes of those shifts are unclear;

(iii) market prices exhibit trends and patterns that tend to repeat and can be identified for prediction.

Technical analysts identify some patterns that indicate the end of trends (reversal patterns), and other patterns that imply a trend is likely to continue (continuation patterns). A well-known example of reversal patterns is the head-and-shoulders pattern, as shown in Figure 1.4(a). This pattern implies that the demand pressure driving the uptrend is disappearing, especially if each of the highs in the pattern are with decreasing trading volume. On the other hand, continuation patterns indicate a pause in a trend instead of a reversal. Triangles form when prices reach lower highs and higher lows over

a certain period of time, as shown in Figure 1.4(b). The triangle patterns imply the current supply and demand forces have been roughly the same for the moment, but they do not indicate a change in direction of the prevailing trend.

Techniques in artificial intelligence such as evolutionary algorithms and artificial neural networks have subsequently been combined with technical analysis to develop advanced models for stock price prediction (Leigh *et al.* (2002), Larsen (2010)). A major concern associated with all the techniques is the assumption of technical analysis suggests that the weak form of efficient market hypothesis (EMH) does not hold. However, various historical studies show virtually zero serial correlations in security prices, which is consistent with the weak form of EMH.

Although technical analysis assumes investor supply/demand relationship is reflected in prices, a model of stock price evolution directly based on buying and selling pressure is still missing. One of the important consequences of this thesis is that a novel methodology of price prediction based on supply/demand information from the limit order book is proposed. In this approach, short-term stock price trends are predicted based on the interactions of buyers and sellers, which are reflected in the limit order book dynamics. The data analysis in Chapter 5 shows that this method works reasonably well.

VOD.L VODAFONE GROUP PLC

4		71006	162.9	-162.95	79959		9
261		6,825,863	155.42523	-167.88226	7,839,197		432
Cumul	Maker	Size	Bid	Ask	Size	Maker	Cumul
4		71006	162.90	162.95	79959		9
10		110436	162.85	163.00	165547		11
11		194292	162.80	163.05	95435		15
14		165796	162.75	163.10	246286		18
16		319872	162.70	163.15	237244		14
10		224002	162.65	163.20	229145		13
7		163907	162.60	163.25	304053		13
4		108296	162.55	163.30	266717		13
3		90365	162.50	163.35	169815		8
25		165282	162.45	163.40	177534		7
1		30702	162.40	163.45	173809		5

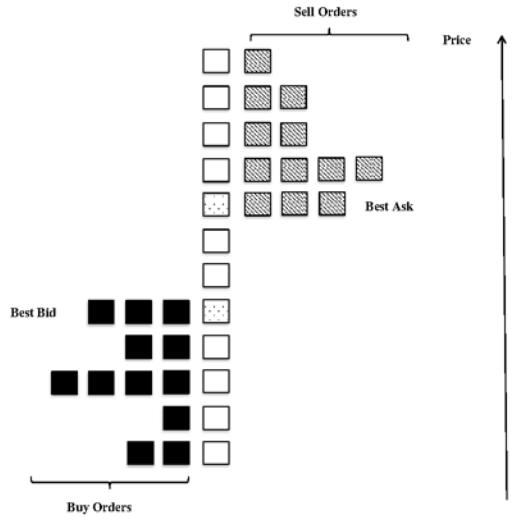
(a) Limit order book of Vodafone Group Plc.

GOOG GOOGLE INC

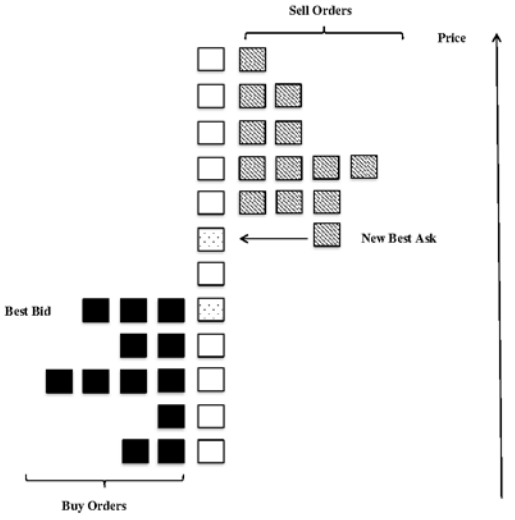
LAST MATCH		TODAY'S ACTIVITY	
Price	384.9000	Orders	1,295,622
Time	15:18:56	Volume	2,791,809
BUY ORDERS		**SELL ORDERS**	
SHARES	PRICE	SHARES	PRICE
50	384.8200	93	384.9500
100	384.8200	100	385.0300
100	384.8100	100	385.0600
300	384.8100	100	385.0700
100	384.8000	200	385.0900
500	384.7900	100	385.1800
200	384.7700	100	385.2400
500	384.7600	25	385.2500
100	384.7100	100	385.3500
100	384.6900	15	385.5000
200	384.6800	200	385.5500
300	384.5900	200	385.6000
100	384.5000	360	385.6300
50	384.0000	100	385.6800
100	384.0000	100	385.7100
(209 more)		(283 more)	

(b) Limit order book of Google Inc.

Figure 1.1: Examples of limit order books.

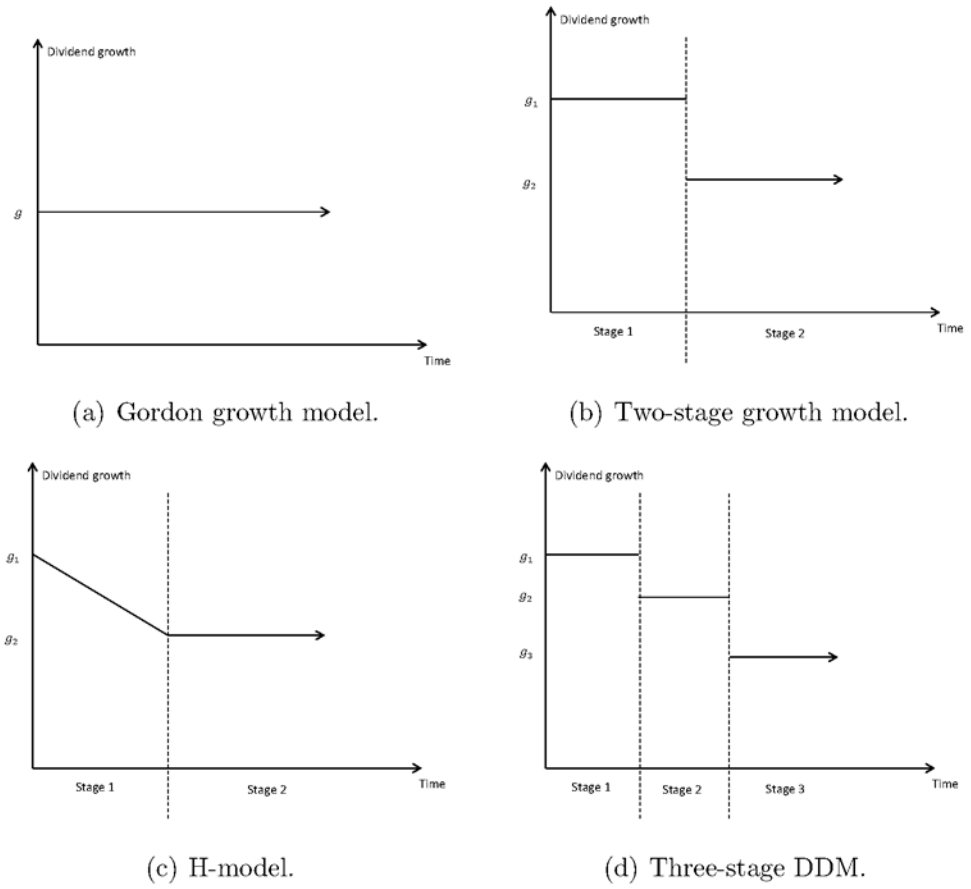


(a) Initial limit order book status.



(b) Updated status with one incoming limit sell order within the spread.

Figure 1.2: An example of limit order book dynamics.



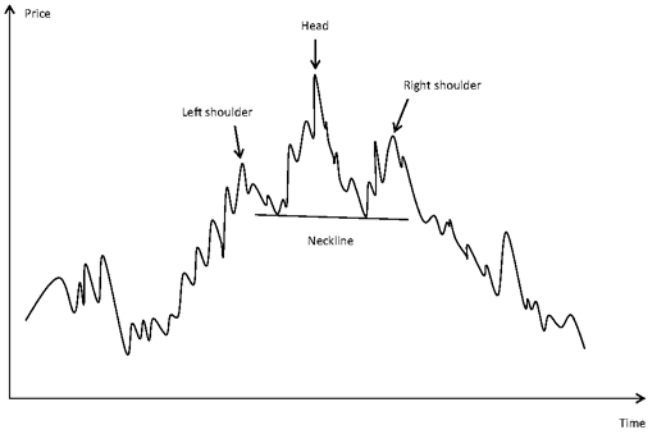
(a) Gordon growth model.

(b) Two-stage growth model.

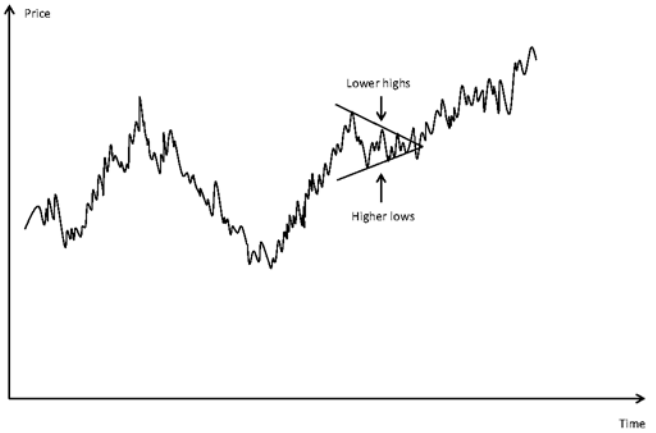
(c) H-model.

(d) Three-stage DDM.

Figure 1.3: Dividend growth rate assumptions in different dividend discount models.



(a) Head-and-shoulders pattern as a reversal pattern in an uptrend.



(b) Triangle pattern as a continuation pattern in an uptrend.

Figure 1.4: Examples of reversal and continuation patterns.

Chapter 2

Models of Limit Order Books

2.1 Limit Order Books

In an order-driven market, traders post two types of orders. A limit order is an order to buy or sell a certain number of shares at the specified or better price. A limit order ensures the investor does not pay more than he/she is willing to, although the execution is not guaranteed. To avoid execution uncertainty, a market order can be used instead of a limit order. A market order instructs the broker to complete the trade immediately at the best possible price and it is often appropriate when a trader believes he/she has information advantages and wants to execute quickly. However, a market order exposes the trader to the risk of executing at unfavorable prices. All outstanding limit orders are aggregated in a limit order book that is available to market participants. As demonstrated in Figure 2.1, there are different types of limit orders depending on their relative positions in the limit order book:

(i) a limit sell order below the best available bid price or a limit buy order above the best available ask price is said to be *marketable* or *aggressively priced* because the order is likely to be filled at least partially;

(ii) a limit order is said to be *making a new market* or *inside the market* if it is between the best bid and the best ask;

(iii) a limit sell order at the best ask or a limit buy order at the best bid is said to *make the market*;

(iv) a limit sell order above the best ask or a limit buy order below the best bid is said to be behind the market because it is likely that the order will not be executed until market prices move towards the limit price;

(v) a limit sell order considerably higher than the best ask or a limit buy order significantly lower than the best bid is said to be far from the market.

In this work, limit orders are placed on a price grid $\{n_1, \ldots, n_2\}$, which denotes multiples of a price tick. Since 2001, the minimum tick size for stock trading above 1 dollar is 0.01. The range between n_1 and n_2 is selected to be large enough such that it is highly unlikely for the stock price under consideration to fluctuate outside the range within the time frame concerned. Since we focus on short-term stock price prediction in this thesis, the model is intended to be used in the time scale of minutes and thus this finite price assumption is reasonable. The state of the limit order book is tracked by a continuous process $X(t) = (X_1(t), \ldots, X_n(t))_{t\geq 0}$, where $n = n_2 - n_1 + 1$ and $|X_l(t)|$ is the amount of outstanding limit orders at price $p = l + n_1 - 1,\ 1 \leq l \leq n$. If the outstanding orders at the l-th entry are buy orders, then $X_l(t)$ is negative; otherwise, $X_l(t)$ is positive.

The location of the *ask price* $l_A(t)$ at time t is defined by

$$l_A(t) = \inf\{l = 1, \ldots, n,\ X_l(t) > 0\} \wedge (n+1),$$

and the corresponding ask price $p_A(t) = l_A(t) + n_1 - 1$. Similarly, the location of the *bid price* $l_B(t)$ at time t is defined by

$$l_B(t) = \sup\{l = 1, \ldots, n,\ X_l(t) < 0\} \vee 0,$$

and the corresponding bid price $p_B(t) = l_B(t) + n_1 - 1$. According to the above definition, the location of the ask price becomes $n+1$ when there are no ask orders in the limit order book, and the location of the bid price becomes 0 when there are no bid orders in the limit book. To this end, define the *midprice* $p_M(t)$ and *bid-ask spread* $p_S(t)$ as

$$p_M(t) = \frac{p_B(t) + p_A(t)}{2} \quad \text{and} \quad p_S(t) = p_A(t) - p_B(t),$$

which are the average of and the difference between the current bid and ask prices being quoted respectively.

We use the limit order book data in R *orderbook* package as an illustrative example. This package is developed by Kane *et al.* (2011) and the current order book status is given in Table 2.1. We choose $n_1 = 100$ and $n_2 = 10,000$ so that the range $[1, 100]$ is wide enough to cover the stock price within the time frame of our analysis. Then the order book status at time t=21:35:02 can be described by an n-dimensional lattice $X_n(t)$, and $n = n_2 - n_1 + 1 = 9901$. The 1033-th to the 1037-th entries of $X_n(t)$ are 700, 1,600, 1,100, 1,100, 2,700, and the 1039-th to the 1043-th entries are 400, 1,600, 1,205, 1,400, 900, respectively. All the other entries of $X_n(t)$ are 0. Moreover, the ask price is $p_A(t) = 11.38$, the bid price is $p_B(t) = 11.36$, the midprice is $p_M(t) = 11.37$, and the bid-ask spread is $p_S(t) = 0.02$.

2.2 Dynamics

2.2.1 Market Events

Other than submitting new limit or market orders, a trader can also cancel their orders to exit the market. Every limit order in the order book is assigned a unique order ID so that cancellation messages can identify the corresponding order.

Current time is 21:35:02		
	Price	Ask size
	11.42	900
	11.41	1,400
	11.40	1,205
	11.39	1,600
	11.38	400
2,700	11.36	
1,100	11.35	
1,100	11.34	
1,600	11.33	
700	11.32	
Bid size	Price	

Table 2.1: The current limit order book status.

In this section, we continue using the dataset from Kane *et al.* (2011) to demonstrate how the limit order book is updated with the incoming flows of limit orders, market orders and cancellations. The initial limit order book status is shown in Table 2.1 and Figure 2.2(a). Figure 2.2(a) is a graphical representation of the order book, where price levels are displayed on the y-axis and the corresponding order sizes are displayed on the x-axis. The maximum and minimum price levels in the figure are set by default as 10% above and below the midpoint.

The following three examples of order flows are designed to show their effects on the order book status, respectively.

(i) An incoming limit buy order with 30,000 shares at the initial bid price of 11.36. This leads to an increase of the number of outstanding shares at 11.36 from 2,700 to 32,700. All the other quantities in the order book remain unchanged.

(ii) An incoming market buy order with 3,000 shares. The outstanding limit sell orders at price levels 11.38 (400 shares) and 11.39 (1,600 shares)

Updated LOB with new limit orders			Updated LOB with new market orders			Updated LOB with cancellations		
	Price	Size		Price	Size		Price	Size
	11.42	900		11.44	700		11.43	4,800
	11.41	1,400		11.43	4,800		11.41	1,400
	11.40	1,205		11.42	9,00		11.40	1,205
	11.39	1,600		11.41	1,400		11.39	1,600
	11.38	400		11.40	205		11.38	400
32,700	11.36		2,700	11.36		2,700	11.36	
1,100	11.35		1,100	11.35		1,100	11.35	
1,100	11.34		1,100	11.34		1,100	11.34	
1,600	11.33		1,600	11.33		1,600	11.33	
700	11.32		700	11.32		700	11.32	
Size	Price		Size	Price		Size	Price	

Table 2.2: The updated limit order book (LOB) status with occurrences of market events.

are fully filled. Since there are no more shares available at those two prices, the remaining 1,000 shares are executed against the ask orders at a higher price level of 11.40.

(iii) All orders at the price 11.42 are canceled. In the data file, there are three limit sell orders at this price level with a total of 900 shares. The trade IDs are 8754861, 8536979 and 8816887 respectively. These market events have no effects on the quantities of the bid side.

The updated limit order books are summarized in Table 2.2 and Figures 2.2(b), 2.2(c) and 2.2(d). As one may see, with these incoming market events, the best available ask/bid price and the shape of the order book may change accordingly. Our purpose is to propose a stochastic model which outlines the occurrences of these market events so that the dynamic shape of a limit order book may be predicted and various pricing quantities may be computed.

2.2.2 Models of Order Book Dynamics

For a state $\tilde{x} \in \mathcal{Z}^n$ (n-dimensional lattice), an order size of m units and a price level $n_1 \leq p \leq n_2$, define

$$\tilde{x}^{p\pm m} = \tilde{x} \pm (0, \ldots, 0, m, 0, \ldots, 0),$$

where the m is located at the l-th component and $l = p - n_1 + 1$. The occurrences of market events are reflected in $\tilde{x}$, detailed as follows:

(i) a limit buy order of size m at price level $p < p_A(t)$ increases the outstanding amount at price p: $\tilde{x} \to \tilde{x}^{p-m}$;

(ii) a limit sell order of size m at price level $p > p_B(t)$ increases the outstanding amount at price p: $\tilde{x} \to \tilde{x}^{p+m}$;

(iii) a market buy order of size m decreases the outstanding amount at the ask price $p_A(t)$: $\tilde{x} \to \tilde{x}^{p_A(t)-m}$;

(iv) a market sell order of size m decreases the outstanding amount at the bid price $p_B(t)$: $\tilde{x} \to \tilde{x}^{p_B(t)+m}$;

(v) a cancellation of size m of limit buy orders at price level $p \leq p_B(t)$ decreases the outstanding amount at price p: $\tilde{x} \to \tilde{x}^{p+m}$;

(vi) a cancellation of size m of limit sell orders at price level $p \geq p_A(t)$ decreases the outstanding amount at price p: $\tilde{x} \to \tilde{x}^{p-m}$.

Bouchaud *et al.* (2002) observe that incoming orders arrive more frequently in the vicinity of the current bid/ask price and the rate of occurrences depends on the distance to the bid/ask. To capture these empirical features, Cont *et al.* (2010) model the incoming market events by independent simple Poisson processes, which is analytically tractable. But they point out that the heterogeneity of order sizes need to be incorporated for future exploration.

Based on these studies, a new stochastic model is proposed in this thesis. The market events are modeled by independent compound Poisson processes with different rates and jump size distributions, where the rate outlines the frequency of occurrences of market events and the jump distribution models the size of each event. Specifically, we have

(i) the initial bid-ask spread $S = p_S(0) = p_A(0) - p_B(0) \leq 5$. Although there are several drivers for the bid-ask spread (Wyart *et al.* (2008)), spread is one of the main indicators for liquidity. In this thesis, we only focus on very liquid stocks with small bid-ask spreads for the sake of building a quantitatively tractable model.

(ii) for $1 \leq j_0 \leq 5$, the arrivals of limit buy (respectively sell) orders at a distance of j_0 ticks from the opposite best quote follow independent compound Poisson processes with rate $\lambda(j_0)$ and jump size distribution F_{j_0}; for a distance of $j_0 \geq 6$, $\lambda(j_0) = 0$ because the incoming limit orders come at a very low rate when they are far away from the current bid/ask price.

(iii) the arrivals of market buy (respectively sell) orders follow independent compound Poisson processes with rate μ and jump size distribution Q.

(iv) for $1 \leq j_0 \leq 5$, the occurrences of cancellations of limit orders at a distance of j_0 ticks from the opposite best quote follow independent compound Poisson processes with rate $\theta(j_0)x$ and jump size distribution G_{j_0}, where x is the number of outstanding orders at the respective level. This assumption indicates that each outstanding order can be canceled at an exponential time with parameter $\theta(j_0)$, and thus with a batch of x outstanding orders, the overall cancellation rate is $\theta(j_0)x$. Similarly as limit orders, for a distance of $j_0 \geq 6$, $\theta(j_0) = 0$.

(v) the above events are mutually independent given the current state of the limit order book.

(vi) the occurrences of limit orders and cancellations of limit orders are not simultaneous. This assumption is reasonable because our observations are of high frequency and the unit sampling interval is no more than one second.

According to these assumptions, there are 11 mutually independent pairs of compound Poisson processes driving the dynamics of a limit order book. For each pair, two independent compound Poisson processes are involved sharing the same set of unknown parameters. Next, we have to examine 11 pairs of compound Poisson processes from the limit order book data, and conduct parameter estimation for each pair, respectively. This is discussed in the following chapter.

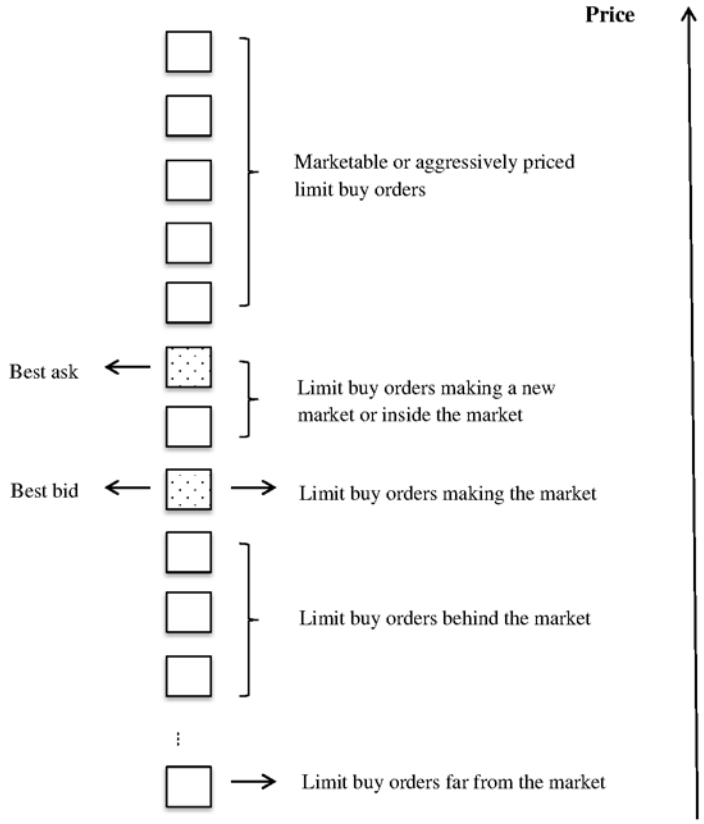
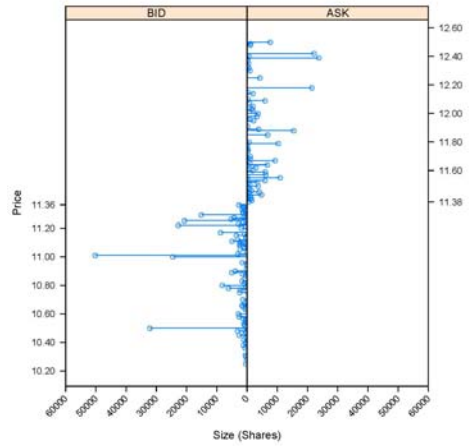
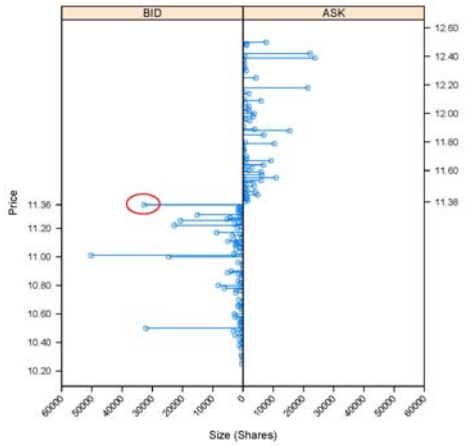


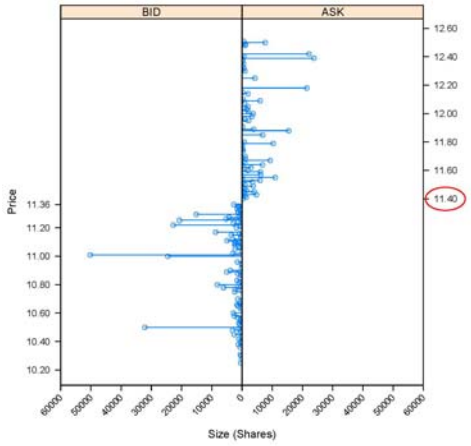
Figure 2.1: An demonstration of different types of limit buy orders.



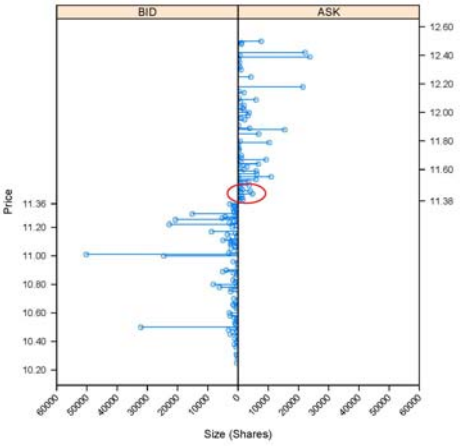
(a) Initial limit order book status.



(b) Updated limit order book status with incoming limit orders.



(c) Updated limit order book status with incoming market orders.



(d) Updated limit order book status with cancellations of limit orders.

Figure 2.2: Different limit order book status with occurrences of market events.

Chapter 3

Parameter Estimation

3.1 Maximum Empirical Likelihood Estimator (MELE)

Likelihood is one of the most important concepts in statistical inference and it has been shown to be useful in nonparametric contexts. Owen (1998, 1990) proposes an empirical likelihood ratio statistic for nonparametric problems and subsequent works include Owen (1991), Chen and Hall (1993), Qin *et al.* (1994), Kitamura (1997), Chan *et al.* (2009). The standard empirical likelihood methods for a mean are outlined in Section 3.1.1, and the more general empirical likelihood methods based on characteristic functions are introduced in Section 3.1.2.

3.1.1 Empirical Likelihood Method for A Mean

Let $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_{N_0}$ be *i.i.d.* observations of $\boldsymbol{Y}$ from a d-variate distribution $F(\boldsymbol{\theta})$. For simplicity we consider the mean $\boldsymbol{\mu}$ of F. The empirical likelihood function is defined by

$$L(F) = \prod_{j=1}^{N_0} dF(\boldsymbol{y}_j) = \prod_{j=1}^{N_0} p_j, \tag{3.1}$$

where

$$p_j = \begin{cases} f(\boldsymbol{y}_j) & \text{if } \boldsymbol{Y} \text{ is continuous and } f \text{ is the density function,} \\ P(\boldsymbol{Y} = \boldsymbol{y}_j) & \text{if } \boldsymbol{Y} \text{ is discrete.} \end{cases}$$

The empirical distribution function is $F_{N_0}(\boldsymbol{y}) = N_0^{-1} \sum_{j=1}^{N_0} I(\boldsymbol{y}_j \leq \boldsymbol{y})$, where $I(\boldsymbol{y}_j \leq \boldsymbol{y}) = \prod_{k=1}^{d} I(y_{jk} \leq y_k)$. The empirical likelihood ratio is then defined as

$$R(F) = L(F)/L(F_{N_0}) = \prod_{j=1}^{N_0} N_0 p_j. \tag{3.2}$$

Note that we do not require the $\boldsymbol{y}_j$'s be distinct.

To estimate the mean $\boldsymbol{\mu}$ of F, the following profile empirical likelihood ratio function is defined:

$$R_E(\boldsymbol{\mu}) = \sup \Big\{ \prod_{j=1}^{N_0} (N_0 p_j) : p_j \geq 0, \sum_{j=1}^{N_0} p_j = 1, \sum_{j=1}^{N_0} p_j \boldsymbol{y}_j = \boldsymbol{\mu} \Big\}. \tag{3.3}$$

Owen (1988, 1990) shows that a unique value for the right-hand side of Equation (3.3) exists, given that $\boldsymbol{\mu}$ is inside the convex hull of $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_{N_0}$. This maximum may be derived by using Lagrange multipliers, which is given in the following theorem.

Theorem 3.1. *The maximum empirical likelihood estimator (MELE) is obtained by maximizing the log-likelihood ratio*

$$l(\boldsymbol{\mu}) = -2 \log R_E(\boldsymbol{\mu}) = 2 \sum_{j=1}^{N_0} \log(1 + \boldsymbol{t}'(\boldsymbol{y}_j - \boldsymbol{\mu})), \tag{3.4}$$

where $\boldsymbol{t}$ is the solution to

$$\sum_{j=1}^{N_0} \frac{\boldsymbol{y}_j - \boldsymbol{\mu}}{(1 + \boldsymbol{t}'(\boldsymbol{y}_j - \boldsymbol{\mu}))} = 0. \tag{3.5}$$

We define

$$\hat{\boldsymbol{\mu}} = \operatorname{argmin} l(\boldsymbol{\mu}). \tag{3.6}$$

Proof. Let

$$H = \sum_{j=1}^{N_0} \log(N_0 p_j) - \lambda(\sum_{j=1}^{N_0} p_j - 1) - N_0 \boldsymbol{t}^{'} \sum_{j=1}^{N_0} p_j(\boldsymbol{y}_j - \boldsymbol{\mu}),$$

where λ and $\boldsymbol{t} = (t_1, t_2, \ldots, t_d)^{'}$ are Lagrange multipliers. By taking derivative with respect to p_j, we have

$$\begin{aligned} & \frac{\partial H}{\partial p_j} = \frac{1}{p_j} - \lambda - N_0 \boldsymbol{t}^{'}(\boldsymbol{y}_j - \boldsymbol{\mu}) = 0, \\ \Rightarrow \quad & p_j = \frac{1}{\lambda + N_0 \boldsymbol{t}^{'}(\boldsymbol{y}_j - \boldsymbol{\mu})}. \end{aligned}$$

Note that

$$\sum_{j=1}^{N_0} p_j \frac{\partial H}{\partial p_j} = N_0 - \lambda - \sum_{j=1}^{N_0} N_0 p_j \boldsymbol{t}^{'}(\boldsymbol{y}_j - \boldsymbol{\mu}) = N_0 - \lambda = 0.$$

Therefore, $\lambda = N_0$ and

$$p_j = \frac{1}{N_0(1 + \boldsymbol{t}^{'}(\boldsymbol{y}_j - \boldsymbol{\mu}))}. \tag{3.7}$$

The last constraint in Equation (3.3) gives that $\boldsymbol{t}$ satisfies

$$\sum_{j=1}^{N_0} \frac{\boldsymbol{y}_j - \boldsymbol{\mu}}{(1 + \boldsymbol{t}^{'}(\boldsymbol{y}_j - \boldsymbol{\mu}))} = 0.$$

Moreover, $\boldsymbol{t}$ is the unique solution of the above system, provided that $\sum_{j=1}^{N_0}(\boldsymbol{y}_j - \boldsymbol{\mu})(\boldsymbol{y}_j - \boldsymbol{\mu})^{'}$ is positive definite. Thus, the explicit expression for $R_E(\boldsymbol{\mu})$ is

$$R_E(\boldsymbol{\mu}) = \prod_{j=1}^{N_0} (1 + \boldsymbol{t}^{'}(\boldsymbol{y}_j - \boldsymbol{\mu}))^{-1}, \tag{3.8}$$

and the log-likelihood ratio is

$$l(\boldsymbol{\mu}) = -2 \log R_E(\boldsymbol{\mu}) = 2 \sum_{j=1}^{N_0} \log(1 + \boldsymbol{t}^{'}(\boldsymbol{y}_j - \boldsymbol{\mu})). \tag{3.9}$$

□

3.1.2 Characteristic Functions

Now assume that $d = 1$ and $F(\boldsymbol{\theta})$ is a univariate distribution. The characteristic function of Y is $\phi(t;\boldsymbol{\theta})$. Since $\phi(t;\boldsymbol{\theta}) = Ee^{itY}$, the aforementioned empirical likelihood method for a mean can be used, which identifies the empirical likelihood ratio as

$$\begin{aligned} R_E(t;\boldsymbol{\theta}) =& \sup\Big\{\prod_{j=1}^{N_0}(N_0p_j) : p_j \geq 0, \sum_{j=1}^{N_0} p_j = 1, \sum_{j=1}^{N_0} p_j e^{ity_j} = \phi(t;\boldsymbol{\theta})\Big\} \\ =& \sup\Big\{\prod_{j=1}^{N_0}(N_0p_j) : p_j \geq 0, \sum_{j=1}^{N_0} p_j = 1, \sum_{j=1}^{N_0} p_j \cos(ty_j) = \phi^R(t;\boldsymbol{\theta}), \\ & \sum_{j=1}^{N_0} p_j \sin(ty_j) = \phi^I(t;\boldsymbol{\theta})\Big\}, \end{aligned} \tag{3.10}$$

where ϕ^R and ϕ^I denote the real and imaginary parts of ϕ, respectively. By using the standard Lagrange multiplier procedure, the log-likelihood ratio becomes

$$\begin{aligned} l(t;\boldsymbol{\theta}) =& -2\log R_E(t;\boldsymbol{\theta}) \\ =& 2\sum_{j=1}^{N_0} \log\Big\{1 + \lambda_1(\cos(ty_j) - \phi^R(t;\boldsymbol{\theta})) \\ & + \lambda_2(\sin(ty_j) - \phi^I(t;\boldsymbol{\theta}))\Big\}, \end{aligned} \tag{3.11}$$

where λ_1 and λ_2 satisfy

$$\begin{cases} \sum_{j=1}^{N_0} \frac{\cos(ty_j) - \phi^R(t;\boldsymbol{\theta})}{1+\lambda_1(\cos(ty_j)-\phi^R(t;\boldsymbol{\theta}))+\lambda_2(\sin(ty_j)-\phi^I(t;\boldsymbol{\theta}))} = 0, \\ \sum_{j=1}^{N_0} \frac{\sin(ty_j) - \phi^I(t;\boldsymbol{\theta})}{1+\lambda_1(\cos(ty_j)-\phi^R(t;\boldsymbol{\theta}))+\lambda_2(\sin(ty_j)-\phi^I(t;\boldsymbol{\theta}))} = 0, \end{cases} \tag{3.12}$$

for all $t \in [-a_1, a_1]$ with $a_1 > 0$. Chan *et al.* (2009) give certain regularity conditions and show that the estimates are not sensitive to the choice of a_1. Then, the MELE is defined as

$$\hat{\boldsymbol{\theta}} = \underset{\theta\in\Omega}{\operatorname{argmin}}\, T_1(\boldsymbol{\theta}), \tag{3.13}$$

where

$$T_1(\boldsymbol{\theta}) = \int_{-a_1}^{a_1} l(t;\boldsymbol{\theta})\,\mathrm{d}G(t),$$

and G is a given smooth distribution function. In the following analysis, we choose $a_1 = \frac{1}{2}$ and $G(t)$ to be the uniform weight function.

3.2 Compound Poisson Processes

Empirical likelihood conducts parameter estimation by profiling a nonparametric likelihood. Compared with parametric methods, this approach provides more flexibility and allows for more general distributions. In this section, we consider parameter estimation of compound Poisson processes by assuming the underlying distribution is approximately a binomial distribution.

To be specific, suppose that $\{L(t)\}_{t\geq 0}$ is a compound Poisson process such that

$$L(t) = \sum_{m=1}^{N(t)} Z_m.$$

$N(t) \sim Poisson(\lambda t)$ and Z_m are *i.i.d.* with

$$P(Z_m = k) = p_k^{(L)} = \frac{Bi(n_0, p_0, k)}{1 - Bi(n_0, p_0, 0)} \quad (k = 1, \ldots, n_0),$$

where $Bi(n_0, p_0, \cdot)$ is the distribution of a binomial(n_0, p_0) random variable. According to the definition of a compound Poisson process, the underlying jump distribution is modified from the binomial distribution such that there is 0 mass at 0.

Although the process evolves continuously over time, its observations are only collected at discrete times, say $\delta, 2\delta, \ldots, (N_0 + 1)\delta$, over a time span of $[0, T]$, where $T = (N_0 + 1)\delta$ is the total amount of time the process is observed and δ is the unit sampling interval. Because compound Poisson processes have independent stationary increments, the set of increments $\{Y_j = L_{(j+1)\delta} - L_{j\delta}\}_{j=1}^{N_0}$ is independent with the same distribution. Thus fitting a class

of parameters $\boldsymbol{\theta} = (\lambda, n_0, p_0)$ to the process $\{L(t)\}_{t\geq 0}$ becomes the problem of fitting a class of parameters based on observations $y_1, \ldots, y_{N_0}$.

The real and imaginary parts of the characteristic function of the compound Poisson processes are given in Theorem 3.2. They are required to conduct MELE.

Theorem 3.2. *The characteristic function of the compound Poisson process defined above is*

$$\phi(t;\boldsymbol{\theta}) = \exp\left\{\lambda\delta\big[\sum_{k=1}^{n_0} p_k^{(L)} e^{itk} - 1\big]\right\}.$$

Therefore, the real and imaginary parts of the characteristic function are

$$\begin{cases} \phi^R(t;\boldsymbol{\theta}) = \exp\left\{\lambda\delta[\sum_{k=1}^{n_0} p_k^{(L)}\cos(tk) - 1]\right\}\left\{\cos[\lambda\delta\sum_{k=1}^{n_0} p_k^{(L)}\sin(tk)]\right\}, \\ \phi^I(t;\boldsymbol{\theta}) = \exp\left\{\lambda\delta[\sum_{k=1}^{n_0} p_k^{(L)}\cos(tk) - 1]\right\}\left\{\sin[\lambda\delta\sum_{k=1}^{n_0} p_k^{(L)}\sin(tk)]\right\}. \end{cases}$$

Proof. The characteristic function is

$$\begin{aligned} \phi(t;\boldsymbol{\theta}) &= \mathrm{E}e^{itY} \\ &= \mathrm{E}e^{it(\sum_{m=1}^{N(\delta)} Z_m)} \\ &= \mathrm{E}[\mathrm{E}(e^{it(\sum_{m=1}^{N(\delta)} Z_m)}|N(\delta))] \\ &= \mathrm{E}[(\mathrm{E}e^{itZ_m})^{N(\delta)}] \\ &= \exp\left\{\lambda\delta\big[\sum_{k=1}^{n_0} p_k^{(L)} e^{itk} - 1\big]\right\}. \end{aligned}$$

Then by using the Euler's formula $e^{a+bi} = e^a(\cos b + i\sin b)$, the above equation can be rewritten as

$$\begin{aligned} \phi(t;\boldsymbol{\theta}) &= \exp\left\{\lambda\delta\big[\sum_{k=1}^{n_0} p_k^{(L)} e^{itk} - 1\big]\right\} \\ &= \exp[\lambda\delta(T_1 + iT_2 - 1)] \\ &= \exp[\lambda\delta(T_1 - 1)][\cos(\lambda\delta T_2) + i\sin(\lambda\delta T_2)], \end{aligned}$$

where $T_1 = \sum_{k=1}^{n_0} p_k^{(L)}\cos tk$ and $T_2 = \sum_{k=1}^{n_0} p_k^{(L)}\sin tk$. Thus the real and

imaginary parts of the characteristic function can be obtained. □

Then, the log-likelihood ratio defined in Equation (3.11) is

$$l(t;\boldsymbol{\theta}) = 2\sum_{j=1}^{N_0} \log\Big\{1 + \lambda_1(\cos(ty_j) - \phi^R(t;\boldsymbol{\theta})) + \lambda_2(\sin(ty_j) - \phi^I(t;\boldsymbol{\theta}))\Big\},$$

where λ_1 and λ_2 satisfy Equations (3.12). The MELE is $\hat{\boldsymbol{\theta}} = \operatorname{argmin} T_1(\boldsymbol{\theta})$, and the target function $T_1(\boldsymbol{\theta})$ is obtained by integrating $l(t;\boldsymbol{\theta})$ over a given distribution for t.

The following simulation study is designed to evaluate the performance of this methodology. The true parameters are $\boldsymbol{\theta} = (\lambda, n_0, p_0) = (2, 4, 0.5)$. To find $\boldsymbol{\theta}$ which minimizes the target function $T_1(\boldsymbol{\theta})$, initial guesses about $\boldsymbol{\theta}$ are needed. Four different tests are conducted with different initial parameter values, and for each test 50 data sets are generated with 1,000 observations. The result is reported in Table 3.1. In test 1 the true value of n_0 is used as its initial guess, and in the rest of the tests the start values of n_0 change to 5,6,7 respectively. Although the estimates of λ and n_0p_0 in all tests are satisfactory, the respective estimation results of n_0 and p_0 depend entirely on the prior information of n_0.

The difficulty in estimating n_0 and p_0 respectively for a binomial distribution is a well-known problem. Olkin *et al.* (1981) demonstrate that the method of moments estimator and the MLE of n_0 can be highly unstable. This is illustrated by the following example. Two data sets (16,18,22,25,27) and (16,18,22,25,28) from the binomial distribution are used to estimate n_0 and p_0. The only difference between the two data sets is the fifth observation changes from 27 to 28, but this results in a massive change in the estimate of n_0. The following two methods are proposed to resolve this estimation problem.

	λ	n_0	p_0
True values	2.00	4.00	0.50
Test 1			
Initial values	10.00	4.00	0.10
Estimated values	2.02	4.00	0.50
Test 2			
Initial values	10.00	5.00	0.10
Estimated values	2.09	5.00	0.36
Test 3			
Initial values	10.00	6.00	0.10
Estimated values	2.14	6.00	0.30
Test 4			
Initial values	10.00	7.00	0.10
Estimated values	2.13	7.00	0.25

Table 3.1: Estimation results with original MELE.

3.2.1 Binomial Poisson Approximation

The binomial(n_0, p_0) can be regarded as the distribution of a sum of n_0 *i.i.d.* Bernoulli random variables $X = X_1 + \ldots + X_{n_0}$ with $P(X_i = 1) = p_0$. The normal approximation to the binomial distribution works well when the variance $n_0 p_0(1-p_0)$ is large, which makes each standardized summand $(X_i - p_0)/\sqrt{n_0 p_0(1-p_0)}$ a relatively small contribution to the standardized sum $(X - p_0)/\sqrt{n_0 p_0(1-p_0)}$. When n_0 is large and p_0 is small, in such a way $n_0 p_0$ is a constant not large enough, a different type of approximation to the binomial distribution is better. This is the well known Poisson approximation and it is established in the following procedure.

Suppose $X \sim binomial(n_0, p_0)$ and $Y \sim Possion(\nu)$. The recursion relations of the two distributions are

$$\begin{aligned} P(X=x) &= \binom{n_0}{x} p_0^x (1-p_0)^{n_0-x} \\ &= \frac{n_0-x+1}{x}\binom{n_0}{x-1}\frac{p_0}{1-p_0} p_0^{x-1}(1-p_0)^{n_0-x+1} \\ &= \frac{n_0-x+1}{x}\frac{p_0}{1-p_0}P(X=x-1); \qquad (3.14) \\ P(Y=y) &= \frac{\nu^y}{y!}e^{-\nu} \\ &= \frac{\nu}{y}\frac{\nu^{y-1}}{(y-1)!}e^{-\nu} \\ &= \frac{\nu}{y}P(Y=y-1). \qquad (3.15) \end{aligned}$$

Set $\nu = n_0 p_0$ and, if p_0 is small, we have

$$\frac{n_0-x+1}{x}\frac{p_0}{1-p_0} = \frac{n_0 p_0 - p_0(x-1)}{x - p_0 x} \approx \frac{\nu}{x},$$

since the terms $p_0(x-1)$ and $p_0 x$ can be ignored for small p_0. Thus, to this level of approximation Equation (3.14) becomes

$$P(X=x) = \frac{\nu}{x}P(X=x-1), \qquad (3.16)$$

which is the Poisson recursion relation. To complete the approximation, the only equation that needs to be established is $P(X=0) \approx P(Y=0)$, because all other probabilities will follow from (3.16).

$$P(X=0) = (1-p_0)^{n_0} = (1-\frac{\nu}{n_0})^{n_0} = e^{-\nu},$$

as n_0 goes to ∞. Therefore the Poisson approximation to the binomial distribution is completed. This indicates that the Poisson distribution with parameter $\nu = n_0 p_0$ can be used as an approximation to binomial(n_0, p_0) when n_0 is reasonably large and p_0 is reasonably small, which helps avoid the difficulty of estimating n_0 and p_0 separately.

	λ	ν
True values	2.00	2.00
Initial values	10.00	6.00
Mean	2.02	1.97
Standard deviation	0.10	0.15

Table 3.2: Estimation results with true parameters $\boldsymbol{\theta} = (2, 100, 0.02)$.

In Figure 3.1, the true probabilities of Poisson(2), binomial(100,0.02) and binomial(4,0.5), and their adjusted distributions with 0 mass at 0 are compared. From the two figures, one can tell that the behaviors of Poisson(2) and binomial(100,0.02) are similar, while there is a relatively large difference between these two distributions and binomial(4,0.5).

Two tests are conducted to examine the parameter estimation procedure of combining MELE with binomial Poisson approximation. 50 datasets, each with 1,000 observations, are generated by using two different sets of true parameters $\boldsymbol{\theta} = (\lambda, n_0, p_0) = (2, 100, 0.02)$ and $\boldsymbol{\theta} = (\lambda, n_0, p_0) = (2, 4, 0.5)$, respectively. The estimates and standard errors of λ and $\nu = n_0 p_0$ in each test are reported in Table 3.2 and Table 3.3. For both cases, the problem of estimation results depending heavily on initial guesses has been resolved. Moreover, when the true parameters are $\boldsymbol{\theta} = (2, 100, 0.02)$, the proposed procedure leads to estimates very close to their true values. To better demonstrate the stability of parameter estimates, the estimation results of λ and ν in each sample path are represented as black dots in Figure 3.2, with the straight line indicating the true values.

However, in case 2 when $\boldsymbol{\theta} = (2, 4, 0.5)$, the parameter estimates deviate from their true values to some extent because Poisson may not be a good approximation to binomial distribution in this case. Since in the limit order book problem, n_0 may not be large enough to be well approximated, an alternative estimation procedure is discussed in Section 3.2.2.

	λ	ν
True values	2.00	2.00
Initial values	10.00	6.00
Mean	2.24	1.48
Standard deviation	0.07	0.07

Table 3.3: Estimation results with true parameters $\boldsymbol{\theta} = (2, 4, 0.5)$.

3.2.2 Modified MELE

For a binomial(n_0, p_0) distribution, although estimation of p_0 when n_0 is known is a standard textbook problem, estimating n_0 and p_0 simultaneously when both are unknown remains an interesting problem for half a century. Literatures regarding this estimation problem include Haldane (1941), Draper and Guttman (1971), Olkin *et al.* (1981), Lindsay (1985), Carroll and Lombard (1985), Kahn (1987), Raftery (1988), and Hall (1994). Among these, the Carroll-Lombard (CL) estimate (1985) is one of the best estimates with satisfactory overall performance. Carroll and Lombard (1985) propose a method, where the likelihood function of n_0 and p_0 is integrated over a beta density for p_0, and the resulting beta-binomial distribution gives stable and reasonably efficient estimators of n_0.

The idea behind CL estimate is elimination of nuisance parameters, which is an important but difficult topic in statistical inference and has been studied in various researches (Kalbfleisch and Sprott (1970), Basu (1977), Barndorff-Nielsen (1983), Cox and Reid (1987), Cruddas *et al.* (1989), Berger *et al.* (1999)). There are generally two ways to eliminate nuisance parameters – non-integration and integration methods. In the non-integration method, nuisance parameters are replaced by a point estimate, leading to a profile likelihood which needs to be maximized. The other approach of eliminating nuisance parameters is through integration.

Berger *et al.* (1999) show that there is considerable value in utilizing

integrated likelihood to estimate parameters in a binomial distribution. In the example, they consider k independent observations $s = (s_1, \ldots, s_k)$ from a binomial distribution with unknown parameters (n_0, p_0). The likelihood function is

$$L(n_0, p_0) = \Big[\prod_{j=1}^{k} \binom{n_0}{s_j}\Big] p_0^T (1-p_0)^{n_0 k - T},$$

where $T = \sum_{j=1}^{k} s_j$. Four different approaches of eliminating the nuisance parameter p_0 are discussed, among which the first two are non-integration methods and the last two are integration methods.

(i) The profile likelihood:

$$\hat{L}(n_0) = \Big[\prod_{j=1}^{k} \binom{n_0}{s_j}\Big] \frac{T^T (n_0 k - T)^{n_0 k - T}}{(n_0 k)^{n_0 k}}.$$

(ii) The conditional likelihood:

$$L^C(n_0) = \Big[\prod_{j=1}^{k} \binom{n_0}{s_j}\Big] \Big/ \binom{n_0 k}{T}.$$

(iii) The uniform integrated likelihood:

$$L^U(n_0) = \int_0^1 L(n_0, p_0)\, \mathrm{d}p = \Big[\prod_{j=1}^{k} \binom{n_0}{s_j}\Big] \frac{\Gamma(n_0 k - T + 1)}{\Gamma(n_0 k + 2)}.$$

(iv) The beta integrated likelihood:

$$L^a(n_0) = \frac{\Gamma(2a)}{(\Gamma(a))^2} \Big[\prod_{j=1}^{k} \binom{n_0}{s_j}\Big] \frac{\Gamma(n_0 k - T + a)}{\Gamma(n_0 k + 2a)}.$$

One could consider beta(a, a) as prior densities for p_0, which yields a distribution symmetric around $1/2$ with mean $1/2$ and variance $(4(2a+1))^{-1}$. These densities are appropriate when the prior information of p_0 is lacking. The uniform distribution in (iii) is a special case of beta(a, a) when $a = 1$.

	Datasets		
Likelihood types	s_1	s_2	s_3
$\hat{L}(n_0)$	99	191	69
$L^C(n_0)$	∞	∞	∞
$L^U(n_0)$	51	57	46
$L^2(n_0)$	49	52	45

Table 3.4: Modes of likelihoods for n_0.

For the dataset $s = (16, 18, 22, 25, 27)$, Figure 3.3 gives the graphs of $\hat{L}(n_0)$, $L^C(n_0)$, $L^U(n_0)$ and $L^2(n_0)$. The values of the likelihoods are rescaled in the figure for comparison reasons. $\hat{L}(n_0)$ and $L^C(n_0)$ are nearly constant for a huge range of n_0, which causes a great difficulty in statistical inference. However, $L^U(n_0)$ and $L^2(n_0)$ appear to be much more useful than the non-integrated likelihoods, with $L^2(n_0)$ decreasing even faster than $L^U(n_0)$ in the tail. Meanwhile, a sensitivity analysis is performed to investigate whether small differences in data will result in large changes in the estimates of n_0. The three datasets are $s_1 = (16, 18, 22, 25, 27)$, $s_2 = (16, 18, 22, 25, 28)$ and $s_3 = (16, 18, 22, 25, 26)$. The modes of each likelihood for n_0 are given in Table 3.4. Although modes alone do not provide a sufficient summary of likelihoods, the stability of integrated likelihoods, over minor perturbations in the collection of data, is quite attractive.

In this thesis, we extend the idea of Carroll and Lombard (1985) and Berger *et al.* (1999) to incorporate both the non-integration method to eliminate λ and the integration method to eliminate p_0 so as to improve the original MELE of $\boldsymbol{\theta} = (\lambda, n_0, p_0)$. It is shown to be a fairly stable procedure in the simulation study.

The moment generating function of the process, denoted by $M_Y(t)$, is

$$\begin{aligned} M_Y(t) &= \mathrm{E}e^{tY} = \mathrm{E}e^{t(\sum_{m=1}^{N(\delta)} Z_m)} \\ &= \exp\left\{\lambda\delta\big[\sum_{k=1}^{n_0} p_k^{(L)} e^{tk} - 1\big]\right\}. \end{aligned}$$

By taking derivatives with respect to t, we have

$$M_Y^{(1)}(t) = \exp\left\{\lambda\delta\big[\sum_{k=1}^{n_0} p_k^{(L)} e^{tk} - 1\big]\right\} \cdot \Big[\lambda\delta \sum_{k=1}^{n_0} k p_k^{(L)} e^{tk}\Big].$$

Since the i-th moment is equal to the i-th derivative of $M_Y(t)$ evaluated at $t = 0$, the following equations are obtained.

$$\mathrm{E}Y = \lambda\delta \sum_{k=1}^{n_0} k p_k^{(L)} = \lambda\delta \cdot \frac{n_0 p_0}{1-(1-p_0)^{n_0}}. \tag{3.17}$$

Based on Equation (3.17), we suppose that λ has a point mass at

$$\hat{\lambda}(n_0, p_0) = \bar{y} \cdot \frac{1-(1-p_0)^{n_0}}{\delta n_0 p_0}, \tag{3.18}$$

given n_0 and p_0. In this way the target function which needs to be minimized becomes

$$T_1(\boldsymbol{\theta}) = T_1(\hat{\lambda}(n_0, p_0), n_0, p_0) = T_2(n_0, p_0).$$

Then p_0 is given a density proportional to

$$p_0^a (1-p_0)^b,$$

where a and b are integers. For the moment, we choose $a = b = 1$ (A more detailed discussion about the choice of (a, b) is provided in Chapter 6). The target function $T_2(n_0, p_0)$ is integrated over the given prior density of p_0 and the estimate of n_0 is obtained by minimizing

$$T_3(n_0) = \int_0^1 T_2(n_0, p_0) p_0^a (1-p_0)^b \, \mathrm{d}p_0 \tag{3.19}$$

as a function of n_0 over a certain range of integers. Theoretically we do not

	λ	n_0	p_0
True values	2.00	4.00	0.50
Mean	1.99	3.82	0.54
Standard deviation	0.07	0.39	0.04

Table 3.5: Estimation results with modified MELE.

know if (3.19) always has a unique maximum. In the simulation study, however, we find that (3.19) is either increasing or first decreasing and then increasing in n_0, indicating that the unique maximum always exists for (3.19). Surely, one need not be restricted to assuming the distribution of λ and p_0 as the one above, but this choice does lead to some stable estimators, as is shown in the following simulation study.

After obtaining the estimate for n_0, we put it back to $T_2(n_0, p_0)$ and derive the estimate for p_0 as

$$\hat{p}_0 = \operatorname{argmin} T_2(\hat{n}_0, p_0),$$

subject to the constraint that $p_0 \in (0, 1)$. Finally Equation (3.18) gives the estimate of λ as

$$\hat{\lambda} = \bar{y} \cdot \frac{1 - (1 - \hat{p}_0)^{\hat{n}_0}}{\delta \hat{n}_0 \hat{p}_0}.$$

Again the true values of the parameters are set to be $\boldsymbol{\theta} = (\lambda, n_0, p_0) = (2, 4, 0.5)$ in the simulation study and 50 datasets are randomly generated, each with 1,000 observations. The estimation results are reported in Table 3.5, where $\hat{n}_0 = 4$ for 41 times, and $\hat{n}_0 = 3$ for 9 times. The estimates of λ and p_0 in each dataset are shown as black dots in Figure 3.4, with the horizontal lines representing the true values. These simulation results demonstrate the accuracy and stability of this modified estimation algorithm, and in the following section we will conduct parameter estimation of the limit order book by using the modified MELE.

A by-product of this simulation study is that we find computational efficiency is greatly improved for this newly proposed algorithm. This is because the original MELE requires a simultaneous optimization procedure over three parameters $\boldsymbol{\theta} = (\lambda, n_0, p_0)$. However, in the modified algorithm, the minimization process is over two parameters (n_0, p_0) only and it is conducted in sequence, which largely reduces the computation complexity.

3.3 Limit Order Books

In the limit order book data, our observations include both time-stamped sequences of trades (market orders) and quotes (prices and quantities of outstanding limit orders) for the five best price levels on either side of the limit order book. In Table 3.6, a sample of three trades is constructed, where each row provides the time, size and price of a market order. We also display a sample of the ask-side quotes in Table 3.7, where the five best ask prices (pa1, pa2, pa3, pa4, pa5), together with the outstanding quantities of shares at these respective prices (qa1, qa2, qa3, qa4, qa5) are shown.

From Table 3.6, information about incoming market orders can be read and from Table 3.7, information regarding incoming limit orders and cancellations of limit orders can be collected. After the decomposition of limit order book data, the aforementioned modified MELE can be applied to each compound Poisson process respectively. A simulation study is conducted to examine if the procedure can effectively identify the large amount of unknown parameters in the proposed dynamic model.

We generate simulated limit order book data for 50 times and conduct parameter estimation for each generated dataset. The true values of the model parameters are shown in Table 3.8. The estimation result for $\lambda(1)$, which is the rate for incoming limit orders at a distance of 1 tick away from the opposite best quote, is reported in Figure 3.5. Each black dot is the estimate for a given

Time	Price	Size
10:11:01	7.43	2
10:11:04	7.46	2
10:11:19	7.44	1

Table 3.6: A sample of three consecutive trades.

Time	pa1	pa2	pa3	pa4	pa5	qa1	qa2	qa3	qa4	qa5
10:11:01	7.43	7.44	7.46	7.49	7.50	12	13	1	52	11
10:11:03	7.42	7.43	7.44	7.46	7.49	20	12	13	1	52
10:11:04	7.42	7.43	7.44	7.46	7.49	21	11	13	1	52
10:11:05	7.42	7.43	7.44	7.46	7.49	34	4	13	1	52
10:11:19	7.42	7.43	7.44	7.46	7.49	33	4	13	1	52

Table 3.7: A sample of ask-side quotes.

dataset and the black straight line stands for the true value of $\lambda(1)$. Some of the proposed parameter estimates and their standard errors are reported in Table 3.9.

After parameter estimation, the proposed stochastic model can be used to compute various quantities regarding stock price behaviors, such as the conditional probability of price increase and the conditional distribution of the duration before next price move. This is discussed in Chapter 4.

(a) Parameters related to incoming limit orders.

Arrival rates				
$\lambda(1)$	$\lambda(2)$	$\lambda(3)$	$\lambda(4)$	$\lambda(5)$
0.1	0.08	0.08	0.06	0.05
Jumping distributions $F_j(j = 1, \ldots, 5)$				
n_j^l	p_j^l			
4	0.5			

(b) Parameters related to incoming market orders.

Arrival rate	Jumping distribution Q	
μ	n^m	p^m
0.1	4	0.5

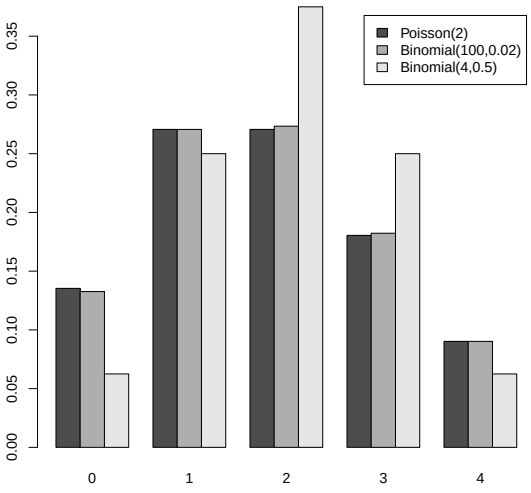
(c) Parameters related to cancellations of limit orders.

Cancellation rates				
$\theta(1)$	$\theta(2)$	$\theta(3)$	$\theta(4)$	$\theta(5)$
0.01	0.008	0.006	0.005	0.004
Jumping distributions $G_j(j = 1, \ldots, 5)$				
n_j^c	p_j^c			
4	0.5			

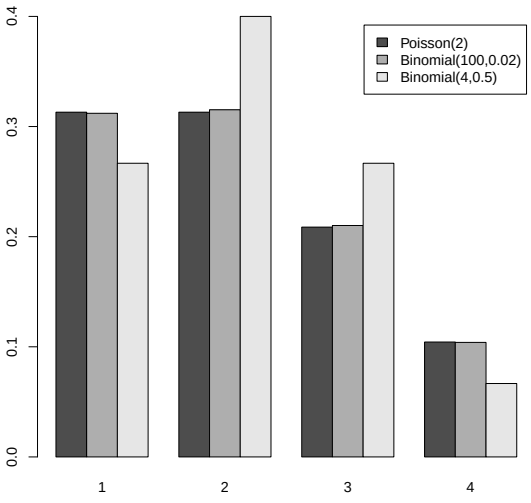
Table 3.8: True values of model parameters.

Limit order arrival rates					
	$\lambda(1)$	$\lambda(2)$	$\lambda(3)$	$\lambda(4)$	$\lambda(5)$
True values	0.1	0.08	0.08	0.06	0.05
Mean	0.098	0.081	0.075	0.057	0.053
S.D.	0.011	0.008	0.009	0.004	0.005
Market order arrival rates					
True values	0.1	Mean	0.095	S.D.	0.010
Cancellation rates					
	$\theta(1)$	$\theta(2)$	$\theta(3)$	$\theta(4)$	$\theta(5)$
True values	0.01	0.008	0.006	0.005	0.004
Mean	0.0111	0.0071	0.0056	0.0047	0.0041
S.D.	0.0016	0.0009	0.0009	0.0005	0.0004

Table 3.9: Estimated parameters and their standard deviations.

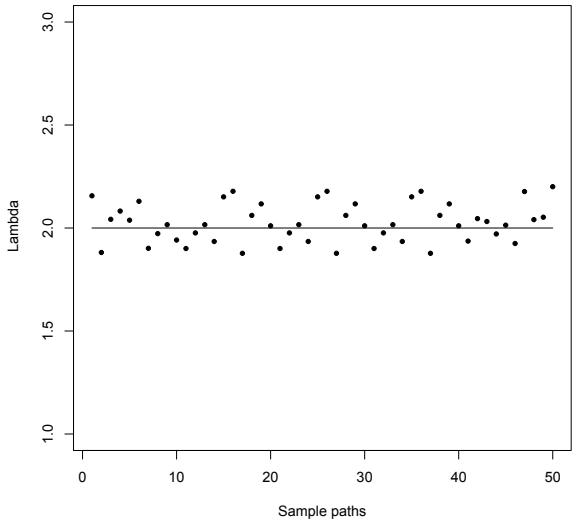


(a) True probability distributions of Poisson(2), binomial(100,0.02) and binomial(4.0.5).

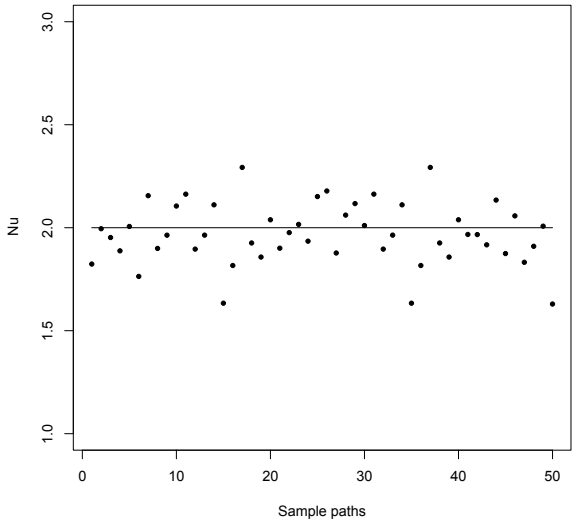


(b) Probability distributions with 0 mass at 0.

Figure 3.1: Binomial Poisson approximation.



(c) Estimation results of λ.



(d) Estimation results of ν.

Figure 3.2: Estimation results with binomial Poisson approximation when the true values are $\boldsymbol{\theta} = (2, 100, 0.02)$.

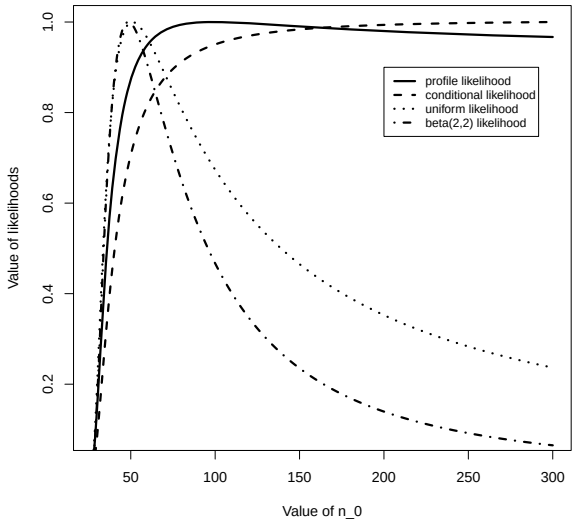
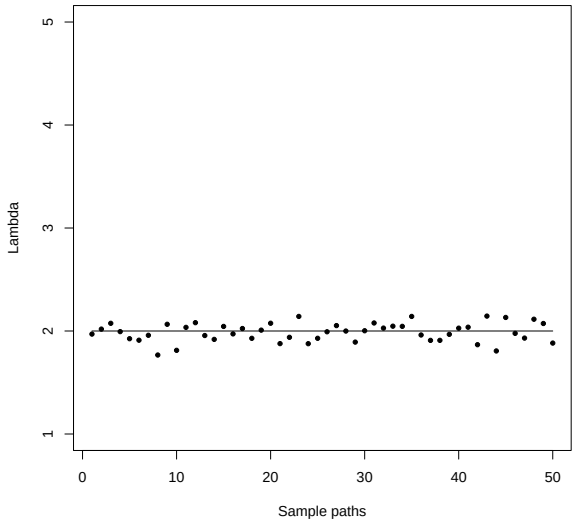
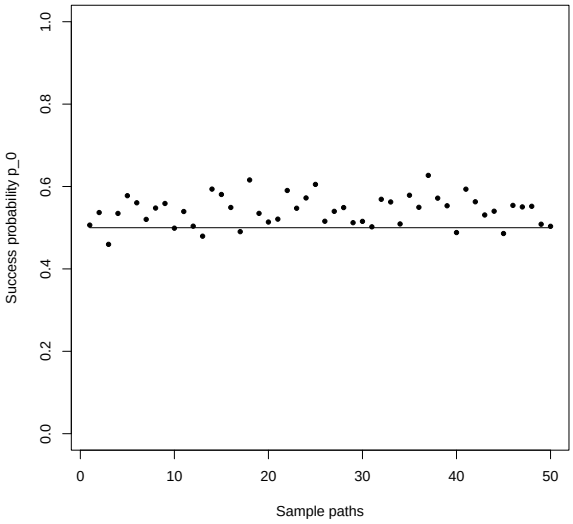


Figure 3.3: The rescaled values of different likelihoods.



(a) Estimation results of λ.



(b) Estimation results of p_0.

Figure 3.4: Estimation results with modified MELE.

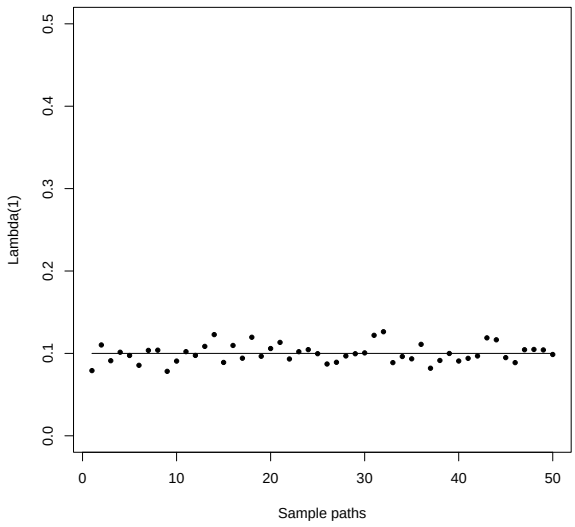


Figure 3.5: The estimation results of $\lambda(1)$ in simulated limited order book data.

Chapter 4

First Passage Time

An important motivation for modeling ultra-high-frequency limit order book dynamics is to apply the information from the limit order book for predicting various short-term behaviors in the market that are useful for trade execution and algorithmic trading. In this chapter, we predict whether the next move in midprice is upward or downward, and when that move occurs. After presenting the distribution function of the first passage time of a birth-death process to 0 in Section 4.1, we calculate the probability of midprice increase in Section 4.2 and give the density distribution of the first price move time in Section 4.3.

4.1 Birth-Death Processes

We denote $X_{p_A(0)}(t)$ as $\tilde{X}_A(t)$ and $X_{p_B(0)}(t)$ as $\tilde{X}_B(t)$, which are stochastic processes monitoring the number of outstanding orders at the current ask and bid prices, respectively. Suppose $\sigma_{A,a}$ and $\sigma_{B,b}$ are the first passage times of $\tilde{X}_A(t)$ and $\tilde{X}_B(t)$ to 0 from their initial levels a and b, and the main purpose in this section is to derive the distributions of these first passage times. We consider $\sigma_{A,a}$ and similar results regarding $\sigma_{B,b}$ can be derived analogously.

Various researches have focused on the calculation of first passage times. Alili *et al.* (2005) study the distribution of σ_x when the process X is an Ornstein-Uhlenbeck process and Borodin and Salminen (1996) derive that of a

Bessel process. The first result for the situation when the process X involves jumps is computed by Zolotarev (1964) and Borokov (1964). Other related studies include Kou and Wang (2003), Roynette *et al.* (2008), and Coutin and Dorobantu (2011).

In our model, $\tilde{X}_A(t)$ is a birth-death process (B-D process) consisting of several independent compound Poisson processes:

(i) the compound Poisson process related to incoming limit sell orders, with rate $\lambda(S)$ and jump size distribution F_S, which increases the value of $\tilde{X}_A(t)$;

(ii) the compound Poisson process related to market buy orders, with rate μ and jump size distribution Q, which decreases the value of $\tilde{X}_A(t)$;

(iii) the compound Poisson process related to cancellation of limit orders, with rate $x\theta(S)$ where x is the number of outstanding orders at the current ask and jump size distribution G_S, which decreases the value of $\tilde{X}_A(t)$ as well.

The distribution function of $\sigma_{A,a}$ is given in the following theorem.

Theorem 4.1. *Suppose the initial bid-ask spread* $1 \leq S \leq 5$. *The distribution function of* $\sigma_{A,a}$ *has a right-derivative at 0 and is differentiable at every point of* $(0, \infty)$. *The right-derivative at* $t = 0$ *is*

$$f^+_{\sigma_{A,a}}(0) = \mu P(Q \geq a) + a\theta(S)P(G_S \geq a) \tag{4.1}$$

and for every $t > 0$, *the derivative* $f_{\sigma_{A,a}}(\cdot)$, *is equal to*

$$f_{\sigma_{A,a}}(t) = E[\mathbf{1}_{\{\sigma_{A,a} \geq t\}}(\mu P(Q \geq \tilde{X}_A(t)) + \tilde{X}_A(t)\theta(S)P(G_S \geq \tilde{X}_A(t)))]. \tag{4.2}$$

Proof. Consider the right-derivative at $t = 0$ first and then extend the analysis to the derivative at $t > 0$.

The probability $P(\sigma_{A,a} \leq h)$ can be split according to N_h, which is the total number of occurrences during time period h, as

$$P(\sigma_{A,a} \leq h) = P(\sigma_{A,a} \leq h, N_h = 1) + P(\sigma_{A,a} \leq h, N_h \geq 2).$$

It suffices to prove the following two properties:

$$\lim_{h\to 0} \frac{P(\sigma_{A,a} \leq h, N_h = 1)}{h} = \mu P(Q \geq a) + a\theta(S)P(G_S \geq a); \quad (4.3)$$

$$\lim_{h\to 0} \frac{P(\sigma_{A,a} \leq h, N_h \geq 2)}{h} = 0. \quad (4.4)$$

By combining Lemma 4.2 and Lemma 4.3, the process $\tilde{X}_A(t)$ has the following features:

(i) the waiting time for an event occurring follows exponential distribution with the rate $\lambda(S) + \mu + a\theta(S)$, whatever type of that market event;

(ii) when there is an event occurring, the probability of an incoming limit sell order equals to $\tilde{p}_1 = \lambda(S)/(\lambda(S)+\mu+a\theta(S))$, the probability of a market buy order equals to $\tilde{p}_2 = \mu/(\lambda(S) + \mu + a\theta(S))$, and the probability of cancellation of limit sell orders is $1 - \tilde{p}_1 - \tilde{p}_2$;

(iii) the size of each event follows the jump distribution of the corresponding compound Poisson process;

(iv) the above distributions are independent.

Henceforth, Equation (4.3) can be proved as follows.

$$\begin{aligned} & \lim_{h\to 0} \frac{P(\sigma_{A,a} \leq h, N_h = 1)}{h} \\ = & \lim_{h\to 0} \frac{a_0 h \exp\{-a_0 h\} \times [P(Q \geq a)\mu/a_0 + P(G_S \geq a)a\theta(S)/a_0]}{h} \\ = & \mu P(Q \geq a) + a\theta(S)P(G_S \geq a), \end{aligned}$$

where $a_0 = \lambda(S) + \mu + a\theta(S)$.

Since

$$P(\sigma_{A,a} \leq h, N_h \geq 2) \leq P(N_h \geq 2) = 1 - \exp\{-a_0 h\} - a_0 h \exp\{-a_0 h\},$$

by L'Hopital's rule, we have

$$\lim_{h \to 0} \frac{P(\sigma_{A,a} \leq h, N_h \geq 2)}{h} = 0.$$

Equation (4.4) is therefore proved.

Now consider the derivative at $t > 0$. Similarly, we split the probability $P(t \leq \sigma_{A,a} \leq t + h)$ into two parts, according to the value of $N_{t+h} - N_t$:

$$\begin{aligned} & P(t \leq \sigma_{A,a} \leq t + h) \\ = \; & P(t \leq \sigma_{A,a} \leq t + h, N_{t+h} - N_t = 1) + P(t \leq \sigma_{A,a} \leq t + h, N_{t+h} - N_t \geq 2) \\ = \; & B_1(h) + B_2(h). \end{aligned}$$

Note that

$$B_2(h) \leq P(N_{t+h} - N_t \geq 2),$$

and thus

$$\lim_{h \to 0} \frac{B_2(h)}{h} = 0.$$

With respect to $B_1(h)$, by the Markov property at t,

$$B_1(h) = E[\mathbf{1}_{\{\sigma_{A,a} \geq t\}} P^t(\sigma_{\tilde{X}_A(t)} \leq h, N_h = 1)],$$

where $P^t(\cdot) = P(\cdot | \tilde{X}_A(t))$. By using Equation (4.3), $B_1(h)/h$ converges to

$$E[\mathbf{1}_{\{\sigma_{A,a} \geq t\}} (\mu P(Q \geq \tilde{X}_A(t)) + \tilde{X}_A(t) \theta(S) P(G_S \geq \tilde{X}_A(t)))],$$

as $h \to 0$. Therefore, the theorem is proved. □

Lemma 4.2. *Let X, Y, Z be independent random variables. Suppose they are exponentially distributed with parameters μ, λ and θ. Then $X \wedge Y \wedge Z$ is also exponentially distributed with parameter $\mu + \lambda + \theta$.*

Proof. We consider X and Y first, and compute the cumulative density function of $X \wedge Y$ as follows.

$$\begin{aligned} P(X \wedge Y \leq x) &= 1 - P(X \wedge Y \geq x) \\ &= 1 - P(X \geq x)P((Y \geq x) \\ &= 1 - \exp\{-(\mu + \lambda)x\}. \end{aligned}$$

By taking derivative with respect to x, the density of $X \wedge Y$ is given by

$$f(x) = (\mu + \lambda) \exp\{-(\mu + \lambda)x\}.$$

Therefore, the conclusion follows by induction. □

Lemma 4.3. *Let X, Y, Z be three independent random variables. Suppose they are exponentially distributed with parameters μ, λ and θ. Then the probability that $X \leq Y$ and $X \leq Z$ is*

$$P(X \leq Y, X \leq Z) = \frac{\mu}{\mu + \lambda + \theta}.$$

Proof. From Lemma 4.2, we conclude that $Y \wedge Z$ is exponentially distributed with the rate $\lambda + \theta$. Since X and $Y \wedge Z$ are independent, we have

$$\begin{aligned} P(X \leq Y, X \leq Z) &= P(X \leq Y \wedge Z) \\ &= \int_0^\infty \int_x^\infty \mu \exp\{-\mu x\}(\lambda + \theta) \exp\{-(\lambda + \theta)y\} \, \mathrm{d}y \, \mathrm{d}x \\ &= \int_0^\infty \mu \exp\{-(\mu + \lambda + \theta)x\} \, \mathrm{d}x \\ &= \frac{\mu}{\mu + \lambda + \theta}. \end{aligned}$$

□

In Figure 4.1, we compare the empirical density of the first passage time (dashed line) with the estimated theoretical density given in Theorem 4.1 (solid line). On the left part of this figure, we use a sample size of 1,000 to fit the empirical distribution curve as well as estimate the expectation in Equation

(4.2). On the right side, the sample size increases to 10,000. By checking these curves, we see that with enough number of simulations, both curves are smoothed and merged.

4.2 Direction of First Price Move

Let τ be the time of the first change in midprice, that is,

$$\tau = \inf\{t > 0, p_M(t) \neq p_M(0)\}.$$

Given the initial description of the limit order book, the probability that the next move in midprice in an up can be denoted by

$$P[p_M(\tau) > p_M(0) | \tilde{X}_A(0) = a, \tilde{X}_B(0) = b, p_S(0) = S], \tag{4.5}$$

where $1 \leq S \leq 5$. The following theorem computes this probability.

Theorem 4.4. *When the initial bid-ask spread $1 < S \leq 5$, the probability of an increase in midprice in equation (4.5) is given by*

$$\int_0^\infty \int_0^{t_2} e^{-\Lambda(t_1+t_2)} [f_{\sigma_{A,a}}(t_1) + \Lambda \int_{t_1}^\infty f_{\sigma_{A,a}}(x)\,\mathrm{d}x][f_{\sigma_{B,b}}(t_2) + \Lambda \int_{t_2}^\infty f_{\sigma_{B,b}}(x)\,\mathrm{d}x]\,\mathrm{d}t_1\,\mathrm{d}t_2, \tag{4.6}$$

where $\Lambda = \sum_{i=1}^{S-1} \lambda(i)$. $f_{\sigma_{A,a}}(\cdot)$ and $f_{\sigma_{B,b}}(\cdot)$ are the pdf of the first passage times $\sigma_{A,a}$ and $\sigma_{B,b}$ and are given in Theorem 4.1. When $S = 1$, (4.6) reduces to

$$\int_0^\infty \int_0^{t_2} f_{\sigma_{A,a}}(t_1) f_{\sigma_{B,b}}(t_2)\,\mathrm{d}t_1\,\mathrm{d}t_2. \tag{4.7}$$

Proof. Let $\sigma_{A,a}$ and $\sigma_{B,b}$ denote the first passage times of $\tilde{X}_A(t)$ and $\tilde{X}_B(t)$ to 0 from their initial levels a and b, respectively.

When the initial bid-ask spread $S = 1$, the first change in midprice happens exactly when either the bid or ask queue hits the state 0 for the first time. Therefore, τ is given by the minimum of the two independent first passage times $\sigma_{A,a}$ and $\sigma_{B,b}$. Moreover, the quantity in Equation (4.5) is given by

$P(\sigma_{A,a} < \sigma_{B,b})$. The probability distribution of $\sigma_{A,a}$ is given in Theorem 4.1, and the probability distribution of $\sigma_{B,b}$ can be obtained analogously. Thus, their joint distribution is

$$f_{\{\sigma_{A,a},\sigma_{B,b}\}}(t_1, t_2) = f_{\sigma_{A,a}}(t_1) f_{\sigma_{B,b}}(t_2),$$

and our desired probability when $S = 1$ is shown in Equation (4.7).

Now consider the case when $1 < S \leq 5$. Suppose σ_A^i denote the first time an ask order arrives i ticks away from the initial bid price and σ_B^i denote the first time a buy order arrives i ticks away from the initial ask price, where $i = 1, \ldots, S-1$. The first change in midprice is given by

$$\begin{aligned} \tau &= \sigma_{A,a} \wedge \sigma_{B,b} \wedge \min\{\sigma_A^i, \sigma_B^i, i = 1, \ldots, S-1\}, \\ &= (\sigma_{A,a} \wedge \sigma_{B,\Lambda}) \wedge (\sigma_{B,b} \wedge \sigma_{A,\Lambda}), \end{aligned}$$

where $\sigma_{A,\Lambda} = \min\{\sigma_A^i, i = 1, \ldots, S-1\}$ and $\sigma_{B,\Lambda} = \min\{\sigma_B^i, i = 1, \ldots, S-1\}$. Notice that σ_A^i and σ_B^i are mutually independent and exponentially distributed with parameters $\lambda(i)$, for $i = 1, \ldots, S-1$. Lemma 4.2 gives that $\sigma_{A,\Lambda}$ and $\sigma_{B,\Lambda}$ are exponential as well, both with the rate $\Lambda = \sum_{i=1}^{S-1} \lambda(i)$. Moreover, $\sigma_{A,a}$ and $\sigma_{B,b}$ are also independent of the arrival times $\sigma_A^i,\ \sigma_B^i$, for $i = 1, \ldots, S-1$. Thus the distributions of the independent random variables $\sigma_{A,a} \wedge \sigma_{B,\Lambda}$ and $\sigma_{B,b} \wedge \sigma_{A,\Lambda}$ are given in Lemma 4.5.

Recall that the midprice is defined as

$$p_M(t) = \frac{p_B(t) + p_A(t)}{2}.$$

The first move in midprice occurs when either the bid price $p_B(t)$ or the ask price $p_A(t)$ changes. The bid price changes when (a) the bid sequence $\tilde{X}_B(t)$ reaches zero or (b) there is a limit buy order arriving within the spread. Similarly, the ask price changes when (c) the ask sequence $\tilde{X}_A(t)$ reaches zero or (d) there is a limit sell order arriving within the spread. Moreover, the first

change in midprice is upward if (b) or (c) happens prior to the occurrences of (a) or (d). Therefore the probability quantity in Equation (4.5) can be rewritten as

$$P(\sigma_{A,a} \wedge \sigma_{B,\Lambda} < \sigma_{B,b} \wedge \sigma_{A,\Lambda}).$$

According to the preceding analysis, this probability equals to

$$\int_0^\infty \int_0^{t_2} e^{-\Lambda(t_1+t_2)}[f_{\sigma_{A,a}}(t_1)+\Lambda \int_{t_1}^\infty f_{\sigma_{A,a}}(x)\,\mathrm{d}x][f_{\sigma_{B,b}}(t_2)+\Lambda \int_{t_2}^\infty f_{\sigma_{B,b}}(x)\,\mathrm{d}x]\,\mathrm{d}t_1\,\mathrm{d}t_2,$$

where $f_{\sigma_{A,a}}(\cdot)$ and $f_{\sigma_{B,b}}(\cdot)$ are the pdf of the first passage times $\sigma_{A,a}$ and $\sigma_{B,b}$, respectively. □

Lemma 4.5. *Let X be an exponentially distributed random variable with rate λ. $\sigma_{A,a}$ and X are independent. Then the distribution of $\sigma_{A,a} \wedge X$ is given by*

$$e^{-\lambda t}[f_{\sigma_{A,a}}(t) + \lambda \int_t^\infty f_{\sigma_{A,a}}(x)\,\mathrm{d}x],$$

where $f_{\sigma_{A,a}}(\cdot)$ is the pdf of $\sigma_{A,a}$ and is given in Theorem 4.1.

Proof. We compute the density $f_{\{\sigma_{A,a}\wedge X\}}$ of the random variable $\sigma_{A,a} \wedge X$ in terms of the density $f_{\sigma_{A,a}}$ of the random variable $\sigma_{A,a}$. Because X is exponential with parameter λ, for all $t \geq 0$, we have

$$\begin{aligned} & P[\sigma_{A,a} \wedge X \leq t] \\ = \;& 1 - P[\sigma_{A,a} > t]P[X > t] \\ = \;& 1 - e^{-\lambda t} \int_t^\infty f_{\sigma_{A,a}}(x)\,\mathrm{d}x. \end{aligned}$$

By taking derivatives with respect to t, we have

$$f_{\{\sigma_{A,a}\wedge X\}}(t) = e^{-\lambda t}[f_{\sigma_{A,a}}(t) + \lambda \int_t^\infty f_{\sigma_{A,a}}(x)\,\mathrm{d}x],$$

for $t \geq 0$. Clearly, $f_{\{\sigma_{A,a}\wedge X\}}(t) = 0$ for $t < 0$. □

In the following simulation study, we consider the case when the initial bid-ask spread $S = 1$ and use the true values of the same set of parameters

shown in Table 3.8. The simulation result is reported in Figure 4.2, where the horizontal line stands for ten different cases with different initial limit order book status. In this way, the true probabilities of stock price increase range from 0.1 to 0.9 in different scenarios and the empirical probabilities are compared with these theoretical probabilities in each scenario.

4.3 Time of First Price Move

We continue with our notations in Section 4.2. Denote the time of the first change in midprice by τ:

$$\tau = \inf\{t > 0, p_M(t) \neq p_M(0)\}.$$

The distribution of τ is given in Theorem 4.6.

Theorem 4.6. *When the initial bid-ask spread $1 < S \leq 5$, the density function of τ is*

$$f_\tau(t) = e^{-2\Lambda t}[f_{\{\sigma_{A,a}\wedge\sigma_{B,b}\}}(t) + 2\Lambda(1 - F_{\{\sigma_{A,a}\wedge\sigma_{B,b}\}}(t))], \tag{4.8}$$

where $\Lambda = \sum_{i=1}^{S-1}\lambda(i)$. $f_{\{\sigma_{A,a}\wedge\sigma_{B,b}\}}(\cdot)$ and $F_{\{\sigma_{A,a}\wedge\sigma_{B,b}\}}(\cdot)$ are the pdf and cdf of $\sigma_{A,a}\wedge\sigma_{B,b}$, and

$$f_{\{\sigma_{A,a}\wedge\sigma_{B,b}\}}(t) = f_{\sigma_{A,a}}(t)(1 - F_{\sigma_{B,b}}(t)) + f_{\sigma_{B,b}}(t)(1 - F_{\sigma_{A,a}}(t)).$$

When $S = 1$, (4.8) reduces to $f_\tau(t) = f_{\{\sigma_{A,a}\wedge\sigma_{B,b}\}}(t)$.

Proof. We start with the simple case when the initial bid-ask spread $S = 1$. The midprice changes for the first time when one of the two birth-death processes $\tilde{X}_A(t)$ and $\tilde{X}_B(t)$ reaches 0 for the first time. Therefore, $\tau = \sigma_{A,a}\wedge$

$\sigma_{B,b}$ and for all $t \geq 0$,

$$\begin{aligned} P[\tau < t] &= P[\sigma_{A,a} \wedge \sigma_{B,b} < t] \\ &= 1 - P[\sigma_{A,a} > t]P[\sigma_{B,b} > t] \\ &= 1 - (1 - F_{\sigma_{A,a}}(t))(1 - F_{\sigma_{B,b}}(t)), \end{aligned}$$

where $F_{\sigma_{A,a}}(\cdot)$ and $F_{\sigma_{B,b}}(\cdot)$ are the cdf of $\sigma_{A,a}$ and $\sigma_{B,b}$. Taking derivatives with respect to t gives

$$f_\tau(t) = f_{\sigma_{A,a}}(t)(1 - F_{\sigma_{B,b}}(t)) + f_{\sigma_{B,b}}(t)(1 - F_{\sigma_{A,a}}(t)),$$

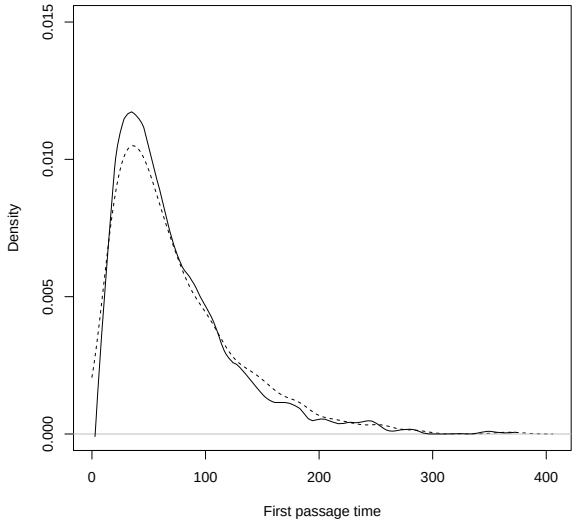
for $t \geq 0$.

Now we move to the case when $1 < S \leq 5$. From the analysis in Section 4.2,

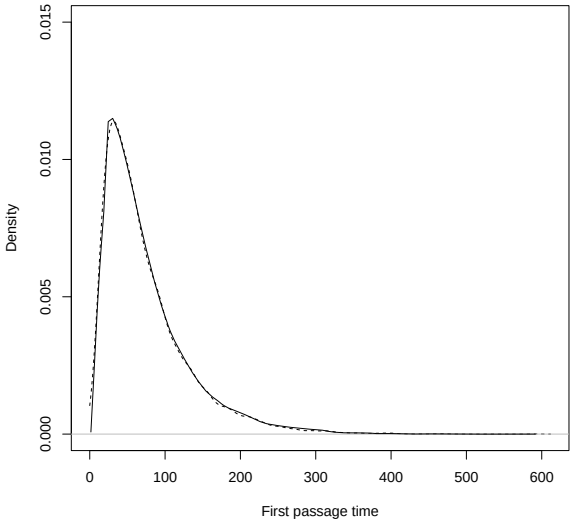
$$\tau = \sigma_{A,a} \wedge \sigma_{B,b} \wedge \min\{\sigma_A^i, \sigma_B^i, i = 1, \ldots, S-1\}.$$

Here $\sigma_{A,a}$ and $\sigma_{B,b}$ are independent of the mutually independent arrival times σ_A^i, σ_B^i for $i = 1, \ldots, S-1$. The distribution of $\sigma_{A,a} \wedge \sigma_{B,b}$ is given in the first case when $S = 1$ and by Lemma 4.2, $\min\{\sigma_A^i, \sigma_B^i, i = 1, \ldots, S-1\}$ is exponential with the parameter $2\sum_{i=1}^{S-1} \lambda(i) = 2\Lambda$. Thus, Lemma 4.5 gives the density function of τ, which is shown as in Equation (4.8). □

Consider the initial bid-ask spread $S = 1$ in the following simulation study, and the result is reported in Figure 4.3. The bar chart depicts the empirical distribution of the time period until the next price change and the black curve stands for its estimated theoretical density as presented in Theorem 4.6.

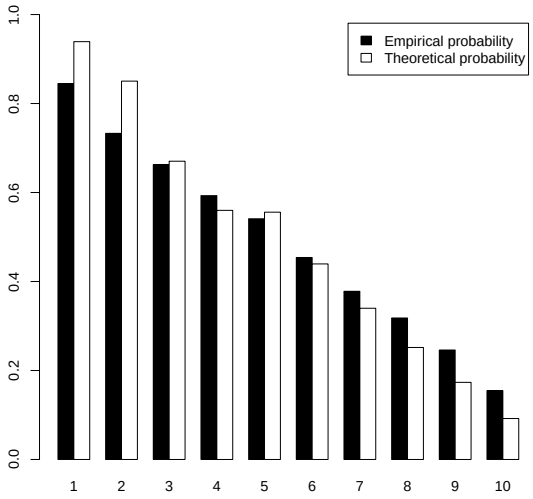


(a) First passage time densities with a sample size of 1,000.

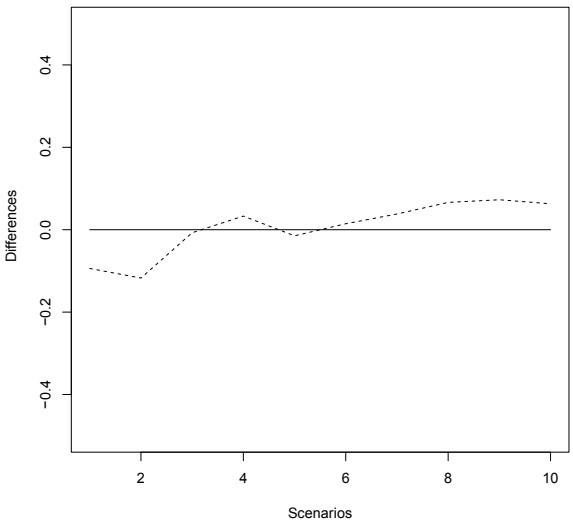


(b) First passage time densities with a sample size of 10,000.

Figure 4.1: First passage time densities with different sample sizes.



(a) Empirical and theoretical probabilities.



(b) Differences between empirical and theoretical probabilities.

Figure 4.2: Probabilities of price increase in 10 different scenarios.

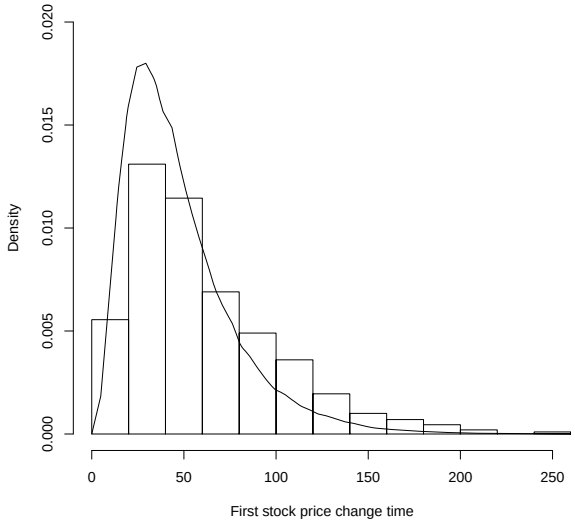


Figure 4.3: The distribution of the duration until the next price move.

Chapter 5

Data Analysis

In this section, the proposed methodology is applied to examine the stock behaviors of different companies. The data are downloaded from LOBSTER (`https://lobster.wiwi.hu-berlin.de/`), which is an on-line limit order book data tool. It gives access to reconstructed limit order book data for the NASDAQ traded stocks. For each stock, there are two related files – a message file and an order book file. Table 5.1 gives an example of the message file and the meaning of each column is explained below.

(i) Time: Seconds after midnight;

(ii) Type:

1: Submission of a new limit order;

2: Cancellation (Partial deletion of a limit order);

3: Deletion (Total deletion of a limit order);

4: Execution of a visible limit order;

5: Execution of a hidden limit order;

(iii) Order ID: Unique order reference number;

(iv) Size: Number of shares;

(v) Price: Dollar price times 10000;

Time	Type	Order ID	Size	Price	Direction
34200.01	4	15835012	34	275200	-1
34200.01	1	16114545	100	275200	-1
34200.01	1	16114695	100	275500	-1
34200.05	3	16063194	100	275000	1
34200.05	3	15833239	100	275100	1

Table 5.1: An example of the message file on LOBSTER.

(vi) Direction:

-1: Sell limit order;

1: Buy limit order.

Note: Execution of a limit sell (buy) order corresponds to a buyer (seller) initiated market trade.

As is shown in Table 5.2, each state update in the order book file corresponds to one trading activity specified in Table 5.1.

We apply the stochastic model to Intel (INTC), which is a U.S. multinational corporation, founded in 1968. It is one of the world's largest semiconductor chip makers and it is also the inventor of the x86 series of microprocessors, which are used in most personal computers. Two different observation periods are selected, one in the morning (09:30:00-09:45:18) and the other in the afternoon (14:00:00-14:52:30), on the same arbitrary trading day. The unit sampling interval is 0.5 seconds. During each observation time period, there are 50,000 market events, including incoming limit orders, incoming market orders and cancellations of limit orders. We choose different observation periods because previous micro-structure studies emphasize the importance of intra-day patterns for volume, price variability and bid-ask spreads (Admati and Pfleiderer (1988), McInish and Wood (1992)) and we want to detect whether the observation duration has any influence on the performance of our model.

Ask price 1	Ask size 1	Bid price 1	Bid size 1	Ask price 2	Ask size 2
275200	66	275100	400	275300	1000
275200	166	275100	400	275300	1000
275200	166	275100	400	275300	1000
275200	166	275100	400	275300	1000
275200	166	275100	300	275300	1000
Bid price 2	Bid size 2	Ask price 3	Ask size 3	Bid price 3	Bid size 3
275000	100	275400	373	274900	200
275000	100	275400	373	274900	200
275000	100	275400	373	274900	200
274900	200	275400	373	274800	661
274900	200	275400	373	274800	661
Ask price 4	Ask size 4	Bid price 4	Bid size 4	Ask price 5	Ask size 5
275600	100	274800	661	275700	100
275600	100	274800	661	275700	100
275500	100	274800	661	275600	100
275500	100	274700	300	275600	100
275500	100	274700	300	275600	100
Bid price 5	Bid size 5				
274700	300				
274700	300				
274700	300				
274600	700				
274600	700				

Table 5.2: An example of the order book file on LOBSTER.

By assuming that the stock price is equal to the average of the best bid and ask prices, we predict short-term stock price trends directly from the order book dynamics and the result is depicted in Figure 5.1. The dashed lines represent the predicted stock prices and the solid lines stand for the real stock prices. The forecasting result is in good agreement with the true price trend within the first 6 minutes in the morning period, and within the first 4 minutes in the afternoon period.

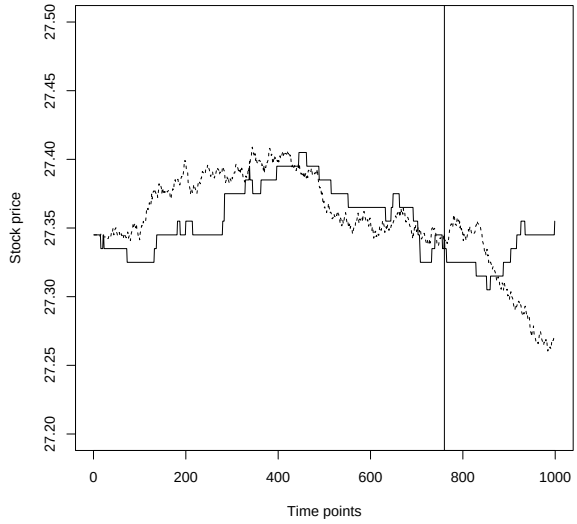
Moreover, we also calculate the theoretical probability of price increase for Intel (INTC) based on observations in the early morning. This probability is repeatedly calculated for 50 times and the calculation results are shown as black dots in Figure 5.2. The average of those probabilities is 0.734, labeled as the black solid line, indicating that there is a large chance that the next move in stock price will be an up. This theoretical result is in accord with the true trend, where stock price increases at the beginning, as is shown in Figure 5.1(a).

Meanwhile, we also investigate the stock behaviors of Amazon (AMZN) and realize our model has its limitations. Amazon.com, Inc., founded in 1994, is an American electronic commerce company headquartered in Seattle, Washington. It is the largest Internet-based retailer in the United States. With different length of history and different business concentration, we will expect that the stock behaviors of Amazon (AMZN) may differ from Intel (INTC) to some extent. Table 5.3 shows the best available bid and ask prices and their respective bid-ask spread of the two companies in the first 3 seconds of one trading day. Although Amazon (AMZN) is considered to be a large and liquid company, its average bid-ask spread is significantly larger than that of Intel (INTC). The lack of liquidity results in a difficulty in applying our model, because there are not enough buying (or selling) activities within 5 ticks of the best ask (or bid) prices, leading to insufficient observations for each proposed dynamic process. To deal with this problem, we may extend our model by

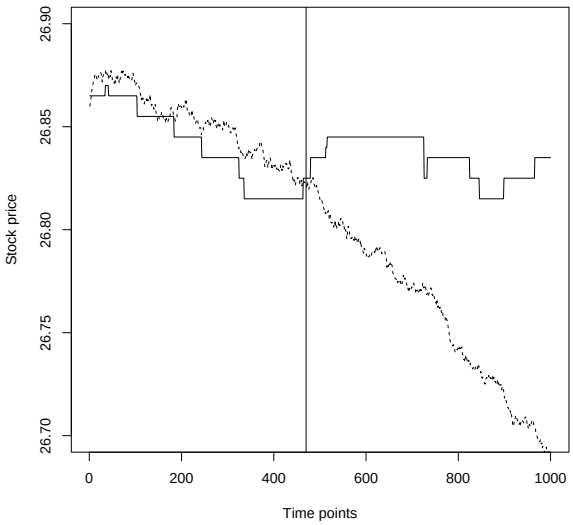
Company	Time	Bid prices	Ask prices	Bid-ask spread
INTC	9:30:00	27.51	27.52	0.01
	9:30:01	27.52	27.53	0.01
	9:30:02	27.52	27.53	0.01
	9:30:03	27.44	27.47	0.03
AMZN	9:30:00	223.81	223.95	0.14
	9:30:01	223.84	223.95	0.11
	9:30:02	223.84	224.24	0.4
	9:30:03	223.89	224.20	0.31

Table 5.3: Best available bid and ask prices and their respective quantities.

incorporating more processes that are far away from the current bid and ask prices in the future.



(a) Prediction results for Intel in the morning period.



(b) Prediction results for Intel in the afternoon period.

Figure 5.1: Short-term stock price prediction.

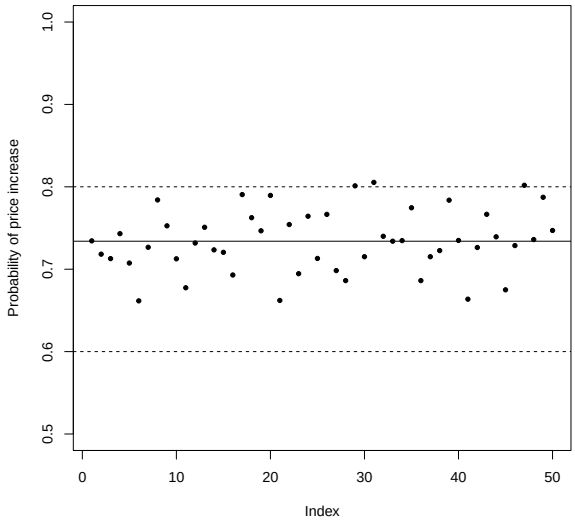


Figure 5.2: The probability of price increase for Intel.

Chapter 6

Discussion

In this thesis, a stochastic model is proposed to describe the dynamics of limit order books, where the order flows are assumed to follow independent compound Poisson processes. The proposed model not only considers the frequency of occurrences of market events, but also different order sizes, which is an important factor in order book dynamics (Cont *et al.* (2010)). Several assumptions are made in Chapter 2 for the sake of building a statistically realistic and quantitatively tractable model. Although the proposed stochastic model seems reasonable, there are several ways to relax these assumptions to account for a richer set of empirically observed features. For example, Biais *et al.* (1995) and Griffiths *et al.* (2000) highlight the autocorrelation in order types, and it remains to be seen if the proposed method can be extended to the case where correlations of order flows are introduced.

The model parameters are estimated in Chapter 3 by means of modified MELE from high-frequency observations in the limit order book. In this modified methodology, λ is replaced by a point estimate

$$\hat{\lambda}(n_0, p_0) = \bar{y} \cdot \frac{1 - (1 - p_0)^{n_0}}{\delta n_0 p_0},$$

given n_0 and p_0, and p_0 is given a prior density proportional to $p_0^a(1 - p_0)^b$. In

$(a, b) = (0, 0)$	
n_0:	$\hat{n}_0 = 3$ for 45 times, and $\hat{n}_0 = 4$ for 5 times
p_0:	$\hat{p}_0 = 0.71$
λ:	$\hat{\lambda} = 1.91$
$(a, b) = (1, 1)$	
n_0:	$\hat{n}_0 = 3$ for 9 times, and $\hat{n}_0 = 4$ for 41 times
p_0:	$\hat{p}_0 = 0.54$
λ:	$\hat{\lambda} = 1.99$
$(a, b) = (2, 2)$	
n_0:	$\hat{n}_0 = 3$ for 10 times, and $\hat{n}_0 = 4$ for 40 times
p_0:	$\hat{p}_0 = 0.53$
λ:	$\hat{\lambda} = 2.02$
$(a, b) = (2, 1)$	
n_0:	$\hat{n}_0 = 3$ for 40 times, and $\hat{n}_0 = 4$ for 10 times
p_0:	$\hat{p}_0 = 0.73$
λ:	$\hat{\lambda} = 1.88$
$(a, b) = (1, 2)$	
n_0:	$\hat{n}_0 = 3$ once, $\hat{n}_0 = 5$ for four times and $\hat{n}_0 = 4$ for 45 times
p_0:	$\hat{p}_0 = 0.49$
λ:	$\hat{\lambda} = 2.04$

Table 6.1: Estimation results with different combinations of (a, b).

this way, the target function in the original MELE $T_1(\boldsymbol{\theta})$ is transformed into

$$T_3(n_0) = \int_0^1 T_2(n_0, p_0) p_0^a (1 - p_0)^b \, \mathrm{d}p_0,$$

where $T_2(n_0, p_0) = T_1(\hat{\lambda}(n_0, p_0), n_0, p_0)$. In Section 3.2.2, we choose $a = b = 1$. To further investigate the influence of the choice of (a, b), different combinations are generated, where $(a, b) = (0, 0), (1, 1), (2, 2), (2, 1), (1, 2)$. The true values of the model parameters are $\boldsymbol{\theta} = (\lambda, n_0, p_0) = (2, 4, 0.5)$ and 50 datasets are randomly generated for each set of (a, b). The estimation results are reported in Table 6.1.

From the table, the following conclusions may be reached.

(i) The beta distribution serves as a better prior distribution of p_0, compared

with the uniform distribution in our problem.

(ii) When the prior information of p_0 is sparse, letting $a = b$ is a safe choice, although this may not give the best estimates.

(iii) The estimate of n_0 tends to be smaller than its true value. The reason of this may be as follows. Figure 6.1 shows the value of $T_3(n_0)$ in one of the generated datasets when $(a, b) = (1, 1)$. $T_3(n_0)$ reaches its minimum when $n_0 = 4$. However, the decreasing rate between $n_0 = 3$ and $n_0 = 4$ is much smaller than the increasing rate between $n_0 = 4$ and $n_0 = 5$, which results in a tendency for the estimate of n_0 to be drifted left.

(iv) In the five combinations above, $(a, b) = (1, 2)$ results in the best estimates. This may be explained in the following way. Figure 6.2 demonstrates the probability distributions when $(a, b) = (1, 1), (2, 2), (2, 1), (1, 2)$, respectively. Since $(a, b) = (1, 2)$ gives a prior distribution of p_0 whose modal is smaller than $1/2$, the estimate of n_0 is inclined to increase, which, to some extent, offsets the shrinkage effects of n_0 explained in (iii) and improves the estimation accuracy.

However, no theoretical results have been achieved as the guidance about choosing (a, b), and we look forward to exploring more about this in the future work.

Furthermore, to apply this estimation approach, either the largest size an order can take or the underlying distribution of a compound Poisson process needs to be given. Although the results seem to be satisfactory by using an adjusted binomial distribution, it will be interesting to consider more general assumptions about the jump distribution.

In Chapter 4, the proposed stochastic model allows for the computation of various quantities, such as the conditional probability of price increase and the conditional distribution of first price change time. These quantities are

wide enough to capture the short-term stock price behaviors and are highly relevant in conducting short-term prediction and designing automated trading strategies. Meanwhile, the theoretical quantities are in good agreement with the corresponding empirical quantities in the simulation studies.

In Chapter 5, a new approach for predicting short-term stock price trends is developed, where data analysis is conducted for Intel (INTC) and Amazon (AMZN). The forecasting result is in good agreement with the theory as well as the short-term trend for very liquid companies with small bid-ask spread, such as Intel (INTC). However, for companies with larger bid-ask spread, the lack of observations in the vicinity of the current bid/ask prices results in great difficulty in the current model, and it remains to be seen if the model can be extended to predict stock prices for such companies in the future.

Moreover, with the rise of high-frequency/algorithmic trading, it will be interesting to consider if the predatory trading strategies, such as quote stuffing, layering, and spoofing, have any impacts on the implications of our model. We also look forward to exploring if market structure changes, such as the introduction of co-location services and stop logic functionality, will potentially influence our findings.

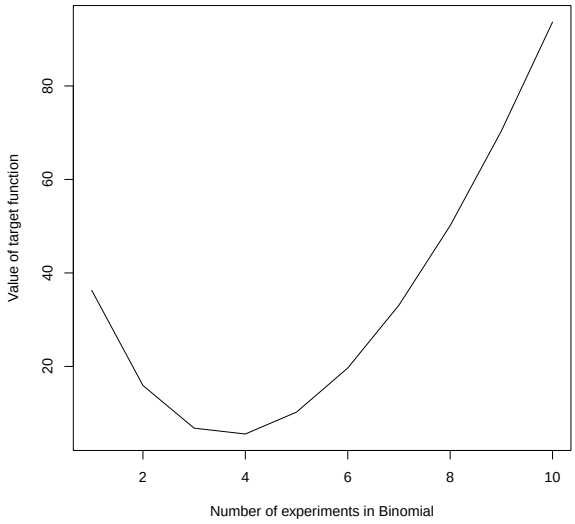


Figure 6.1: Value of $T_3(n_0)$ with different n_0 when $(a, b) = (1, 1)$.

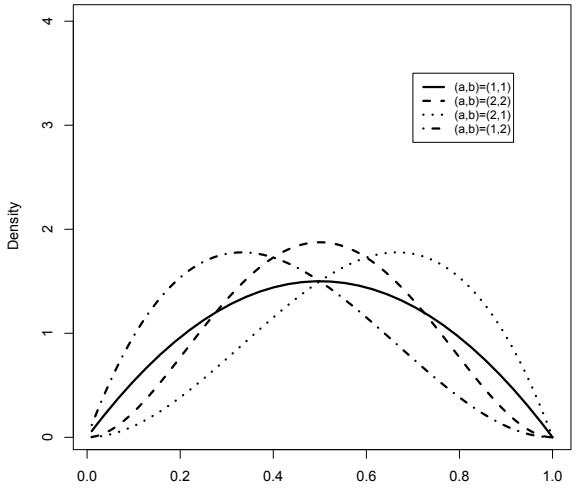


Figure 6.2: Beta distribution with different parameters.

Bibliography

[1] Admati, A. R. and Pfleiderer, P. (1988). A theory of intraday patterns: Volume and price variability. *Review of Financial Studies* **1**, 3–40.

[2] Alfonsi, A., Fruth, A. and Schied, A. (2010). Optimal execution strategies in limit order books with general shape functions. *Quantitative Finance* **10**, 143–157.

[3] Alili, L., Patie, P. and Pedersen, J. L. (2005). Representations of the first hitting time density of an Ornstein-Uhlenbeck process. *Stochastic Models* **21**, 967–980.

[4] Barndorff-Nielsen, O. (1983). On a formula for the distribution of the maximum likelihood estimator. *Biometrika* **70**, 343–365.

[5] Basu, D. (1977). On the elimination of nuisance parameters. *Journal of American Statistical Association* **72**, 355–366.

[6] Berger, J. O., Liseo, B. and Wolpert R. L. (1999). Integrated likelihood methods for eliminating nuisance parameters. *Statistical Science* **14**, 1–28.

[7] Biais, B., Hillion, P. and Spatt, C. (1995). An empirical analysis of the limit order book and the order flow in the Paris Bourse. *Journal of Finance* **50**, 1655–1689.

[8] Borodin, A. and Salminen, P. (1996). *Handbook of Brownian Motion: Facts and Formulae*. Birkhäuser Basel.

[9] Borokov, A. A. (1964). On the first passage time for one class of processes with independent increments. *Theory of Probability and Its Applications* **10**, 331–334.

[10] Bouchaud, J. P., Farmer, D. and Lillo, F. (2008). How markets slowly digest changes in supply and demand. *In: Handbook of Financial Markets: Dynamics and Evolution*. Elsevier: Academic Press, Ch. 2, pp. 57–160.

[11] Bouchaud, J. P., Mezard, M. and Potters, M. (2002). Statistical properties of stock order books: Empirical results and models. *Quantitative Finance* **2**, 251–256.

[12] Bovier, A., Černý, J. and Hryniv, O. (2006). The opinion game: Stock price evolution from microscopic market modeling. *International Journal of Theoretical and Applied Finance* **9**, 91–111.

[13] Brock, W., Lakonishok, J. and LeBaron, B. (1992). Simple technical trading rules and the stochastic properties of stock returns. *The Journal of Finance* **47**, 1731–1764.

[14] Caginalp, G. and Laurent, H. (1998). The predictive power of price patterns. *Applied Mathematical Finance* **5**, 181–205.

[15] Carroll, R. J. and Lombard, F. (1985). A note on N estimators for the binomial distribution. *Journal of the American Statistical Association* **80**, 423–426.

[16] Chan, N. H., Chen, S. X., Peng, L. and Yu, C. L. (2009). Empirical likelihood methods based on characteristic functions with applications to Lévy processes. *Journal of American Statistical Association* **104**, 1621–1630.

[17] Chen, S. X. and Hall, P. (1993). Smoothed empirical likelihood confidence intervals for quantiles. *The Annals of Statistics* **21**, 1166–1181.

[18] Cont, R. and de Larrard, A. (2011). Price dynamics in a Markovian limit order market. *SSRN eLibrary*.

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

[20] Coutin, L. and Dorobantu, D. (2011). First passage time law for some Lévy processes with compound Poisson: Existence of a density. *Bernoulli* **17**, 1127–1135.

[21] Cox, D. R. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society Series B* **49**, 1–39.

[22] Cruddas, A. M., Reid, N. and Cox, D. R. (1989). A time series illustration of approximate conditional likelihood. *Biometrika* **76**, 231.

[23] Draper, N. and Guttman, I. (1971). Bayesian estimation of the binomial parameter. *Technometrics* **13**, 667–673.

[24] Farmer, J. D., Gillemot, L., Lillo, F., Mike, S. and Sen, A. (2004). What really causes large price changes? *Quantitative Finance* **4**, 383–397.

[25] Foucault, T., Kadan, O. and Kandel, E. (2005). Limit order book as a market for liquidity. *Review of Financial Studies* **18**, 1171–1217.

[26] Goettler, R. L., Parlour, C. A. and Rajan, U. (2005). Equilibrium in a dynamic limit order market. *Journal of Finance* **60**, 2149-2192.

[27] Goettler, R. L., Parlour, C. A. and Rajan, U. (2009). Informed traders and limit order markets. *Journal of Financial Economics* **93**, 67-87.

[28] Gourieroux, C., Jasiak, J. and Fol, G. L. (1999). Intra-day market activity. *Journal of Financial Markets* **2**, 193–226.

[29] Griffiths, M. D., Smith, B., Turnbull, D. A. S. and White, R.W. (2000). The costs and determinants of order aggressiveness. *Journal of Financial Economics* **56**, 65–88.

[30] Guilbaud, F. and Pham, H. (2013). Optimal high-frequency trading with limit and market orders. *Quantitative Finance* **13**, 79–94.

[31] Haldane, J.B.S. (1941). The fitting of binomial distributions. *Annals of Eugenics* **11**, 179–181.

[32] Hall, P. (1994). On the erratic behavior of estimators of n in the binomial(n,p) distribution. *Journal of the American Statistical Association* **89**, 344–352.

[33] Harris, L. E. and Panchapagesan, V. (2005). The information content of the limit order book: Evidence from NYSE specialist trading decisions. *Journal of Financial Markets* **8**, 25–67.

[34] Hollifield, B., Miller, R. A. and Sandas, P. (2004). Empirical analysis of limit order markets. *Review of Economic Studies* **71**, 1027–1063.

[35] Kahn, W. D. (1987). A cautionary note for Bayesian estimation of the binomial parameter n. *The American Statistician* **41** 38–39.

[36] Kalbfleisch, J. D. and Sprott, D. A. (1970). Application of likelihood methods to models involving large numbers of parameters. *Journal of the Royal Statistical Society Series B* **32**, 175–208.

[37] Kane, D., Liu, A. and Nguyen, K. (2011). Analyzing an electronic limit order book. *The R Journal* **3**, 64–68.

[38] Kitamura, Y. Empirical likelihood methods with weakly dependent processes. *The Annals of Statistics* **25**, 2084–2102.

[39] Kou, S. G. and Wang, H. (2003). First passage times of a jump diffusion process. *Advances in applied probability* **35**, 504–531.

[40] Larsen, J. I. (2010). *Predicting Stock Prices Using Technical Analysis and Machine Learning*. Master Thesis, Norwegian University of Science and Technology, Trondheim, Norway.

[41] Leigh, W., Purvis, R. and Ragusa, J. (2002). Forecasting the NYSE composite index with technical analysis, pattern recognizer, neural network, and genetic algorithm: a case study in romantic decision support. *Decision Support Systems* **32**, 361–377.

[42] Lindsay, B. G. (1985). Errors in inspection: Integer parameter maximum likelihood in a finite population. *Journal of the American Statistical Association* **80**, 879–885.

[43] Luckock, H. (2003). A steady-state model of the continuous double auction. *Quantitative Finance* **3**, 385–404.

[44] Maslov, S. and Mills, M. (2001). Price fluctuations from the order book perspective – Empirical facts and a simple model. *Physica A* **299**, 234–246.

[45] McInish, T. H. and Wood, R. A. (1992). An analysis of intraday patterns in bid/ask spreads for NYSE stocks. *Journal of Finance* **67**, 753–764.

[46] Obizhaeva, A. and Wang, J. (2013). Optimal trading strategy and supply/demand dynamics. *Journal of Financial Markets* **16**, 1–32.

[47] Olkin, I., Petkau, A. J., and Zidek, J. V. (1981). A comparison of n estimators for the binomial distribution. *Journal of American Statistical Association* **76**, 637–642.

[48] Owen, A.B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* **75**, 237–249.

[49] Owen, A.B. (1990). Empirical likelihood confidence regions. *The Annals of Statistics* **18**, 90–120.

[50] Owen, A.B. (1991). Empirical likelihood for linear models. *The Annals of Statistics* **19**, 1725–1747.

[51] Parlour, C. A. (1998). Price dynamics in limit order markets. *Review of Financial Studies* **11**, 789–816.

[52] Predoiu, S., Shaikhet, G. and Shreve, S. (2011). Optimal execution of a general one-sided limit-order book. *Journal on Financial Mathematics* **2**, 183–212.

[53] Qin, J. and Lawless, J. (1994). Empirical likelihood and general estimating equations. *The Annals of Statistics* **22**, 300–325.

[54] Raftery, A. E. (1988). Inference for the binomial n parameter: A hierarchical Bayes approach. *Biometrika* **75**, 223–228.

[55] Rosu, I. (2009). A dynamic model of the limit order book. *Review of Financial Studies* **22**, 4601–4641.

[56] Roynette, B., Vallois, P. and Volpi, A. (2008). Asymptotic behavior of the hitting time, overshoot and undershoot for some Lévy processes. *ESAIM: Probability and Statistics* **12**, 58–93.

[57] Smith, E., Farmer, J. D., Gillemot, L. and Krishnamurthy, S. (2003). Statistical theory of the continuous double auction. *Quantitative Finance* **3**, 481–514.

[58] Toke, L. M. (2011). "Market making" in an order book model and its impact on the spread. *In: Econophysics of Order-driven Markets: Proceedings of Econophys-Kolkata V*. Springer, pp. 49–64.

[59] Williams, J. B. (1938). *The Theory of Investment Value*. Harvard University Press.

[60] Wyart, M., Bouchaud, J. P., Kockelkoren, J., Potters, M. and Vettorazzo, M. (2008). Relation between bid-ask spread, impact and volatility in order-driven markets. *Quantitative Finance* **8**, 41–57.

[61] Zhao, L. Q. (2010). *A model of limit-order book dynamics and a consistent estimation procedure*. Ph.D. Thesis, Carnegie Mellon University, Pittsburgh, U.S.A.

[62] Zolotarev, V. M. (1964). The first passage time of a level and the behavior at infinity for a class of processes with independent increments. *Theory of Probability and Its Applications* **9**, 653–664.