Modeling the limit order book

Ph.D. Thesis Proposal

Linqiao Zhao
Department of Statistics
Carnegie Mellon University

March 26, 2008

Introduction

The past two decades have seen the rise of automated continuous double auction (CDA) trading systems in stock exchanges throughout the world. In these markets, orders are matched and executed by computer systems while unmatched orders are stored in the *limit order books* (LOB). The LOB contains rich information about market demand and supply and is of great theoretical and practical value. Theoretically, the LOB will help us understand important aspects of market microstructure, such as price formation and market liquidity. Practically, it will help market participants to form optimal trading strategies, lowering trading costs.

This thesis is about modeling the dynamics of the LOB in some particular markets. First, we propose a model - Event-Type model - for the evolution of the LOB. This model describes an "Ideal" data set, in which all the order arrivals and cancelations are recorded in great detail in real time. However, such data are not accessible to us (and not only not to us). Instead, what we have is "Five-Tick" data, which contains only an aggregated level of information regarding the best five buy and sell order prices and sizes. Estimating the Event-Type model directly on such data is very difficult. Therefore an "indirect" inference approach is proposed to link the Event-Type model with our data set via some auxiliary models, which are predictive models on the Five-Tick data.

This proposal is organized as follows. Section 1 introduces the mechanics of the limit order book. Section 2 briefly reviews models related to the evolution of LOB, and literature on indirect inference. Section 3 introduces data and models. In this section, we propose the Event-Type model for the Ideal data, and explain the difficulty of estimating the Event-Type model based on the Five-Tick data set. Section 4 introduces the indirect inference estimation approach for our Event-Type model. In particular, we propose a new technique - functional indirect inference. The last section describes future work.

1 Mechanics of the Limit Order Book

Here we introduce frequently-used concepts and terminology related to the LOB. There are two basic types of orders - *market orders* and *limit orders*. Traders who submit market orders specify the amount to buy or sell, at the best available price, while traders who submit limit orders specify not only the amount to buy or sell, but the worst acceptable price. At each limit price in the limit order book, there is a total outstanding order size associated with that price, and we use *depth* to refer the total outstanding order size at that price. The highest price of all buy limit orders outstanding is the *bid price*, and the lowest price of all sell limit orders is the *offer price* or the *ask price*. "Best orders" denotes orders with the best prices. The difference between the ask price and the bid price is called the *spread*. Traders can only submit orders with prices along a grid of possible prices. The minimum price difference in the LOB is called a *tick* and the magnitude of the tick varies according to the security and the markets.

There is no need to wait for the incoming orders before executing market orders. When market orders arrive, they are matched immediately against limit orders on the opposite side of the LOB. If the size of a buy market order is greater than the depth at the current ask price, the excess part of this demand is matched at prices higher than the current ask price. Such a buy market order is a *walking-up market order*. Correspondingly, there is also a *walking-down market order* on the sell side. But on some exchanges, such as Paris Bourse, a market order is not allowed to walk. In these exchanges, a market order will only be partially executed if it can not be fully matched at the current best price, and the unmatched part of the market order becomes a limit order with price at the best price on the opposite side before execution. Although limit orders can control transaction prices, they are not usually executed immediately. However, sometimes limit orders are executed at once when a buy limit order arrives whose limit price is greater than the ask, or a sell limit order whose limit price is less than the bid. Such limit orders are called *crossing* or *marketable limit orders*, and are executed like market orders. Therefore, the term *effective limit orders* is sometimes used for non-crossing limit orders including the unexecuted part of crossing limit orders, and *effective market orders* for all orders that bring immediate trades including the executed part of crossing limit orders. Unmatched limit orders will be stored into the LOB according to price priority then time priority. Limit orders can be canceled before expiration and will then be removed at once from the LOB, while market orders can not.

2 Background

Modeling the dynamics of the limit order book is practically attractive. Knowing how the LOB evolves helps traders to take corresponding strategies and lower their trading costs. However, due to its complexity, this problem is not yet systematically solved. A lot of partial research has been done, laying the background for our proposed Event-Type model. Some representative work is summarized in this section. In addition, we briefly review indirect inference, the main estimation approach to our model.

The state of the LOB changes over time in response to traders' order submissions and cancelations, which in turn are affected by the LOB. It is such interactions that make the LOB modeling difficult. Although the effect of traders' behavior on the LOB has been studied, the difficulty of quantifying traders' behavior limits this approach. Instead of attempting to anticipate how traders will behave, Garman (1976) studied the market by assuming that traders' net effect is to generate order flows with some distribution. His work leads the study of the LOB in its steady-state, see Domovitz and Wang (1994) and Luckock (2003). These results have not yet lead to a tractable model of LOB dynamics, but the idea that a market could be characterized by order flows inspires our model. Several aspects of order flows have been explored: the distributions of order prices, the occurrences of order arrivals and cancelations, and order decisions.

Studies of the distributions of order prices have mostly concentrated on the distributions of the relative prices (the difference between buy/sell limit order prices and bid/ask prices). Potters and Bouchaud (2002), Zovko and Farmer (2002), Gabaix et al. (2003) and Lillo (2006) all find that the relative prices follow distributions with fat-tails among different markets. These analyses present static results. Some scholars investigate what information in the LOB contributes additional predictive power for the price changes other than lagged best prices. Maslov and Mills (2001) find that a large imbalance in the number of limit orders placed at the bid and ask sides of the book leads to a predictable short term price change, which is in accord with the law of supply and demand. Cao et al (2004) use the data of 21 stocks from the Australian Stock Exchange and find empirical evidence that the LOB beyond the best prices, specifically the lagged limit order book imbalance, is informative about future price movements. Harris and Panchapagesan(2004) and Pascual and Veredas (2004) find the limit order book is informative about future price movements. Hellstrom and Simonsen (2006) use the integer-valued autoregressive model and obtain similar results. In addition, they argue that the value of the information is short-term. All these papers use parametric linear models only. There is some work on using nonlinear modeling, such as neural networks and genetic algorithms, to predict prices in high frequency time series (Bunis and Zhou,1998).

In the LOB, the occurrences of order arrivals and cancelations are recorded at irregularly spaced time intervals. Standard time series techniques relying on fixed-time intervals do not apply. The clustering of these time intervals led Engle and Russell (1998) to introduce an autoregressive conditional duration model (ACD) to model the intervals between trades. However, it is hard to extend these models to multiple occurrence processes. The irregular occurrences are suitable to be consider as point processes, therefore conditional intensity functions can be used to characterize these point processes. Russell (1999) provides autoregressive conditional intensity models (ACI), which use the framework of ACD and apply it to conditional intensity functions. This ACI model can be used to deal with multiple occurrence processes.

Optimal choices between limit orders and market orders are studied by many scholars. Biais, Hillion and Spatt(1995) present an empirical analysis of the order flow of Paris Bourse, and find that when the spread is larger or the order book is thin, limit orders are submitted more often than market orders. In addition, improvement in the best prices is especially frequent when the depth at the quotes is large. Harris and Hasbrouck (1996), using NYSE data, find that limit orders placed at or better than the prevailing quote perform better than do market orders, even after imputing a penalty for unexecuted orders. Chakravarty and Holden (1995), Handa and Schwartz (1996), Harris (1998), Foucault, Kadan and Kandel (2003), Goettler, Parlour and Rajan (2004, 2005) also address the choice between market orders and limit orders. Few researchers study jointly the price and quantity aspects of the order decision, such as Ellul, Holden, Jain and Jennings (2003) and Ronaldo (2004). However, they focus on specific price-quantity combinations. Lo and Sapp (2005) model quantity simultaneously with the levels of the price aggressiveness, which is defined in terms of how quickly an order is executed, similar to the classification in Bias et al.(1995). In addition, changes in the off-best depth, which may not be observable to the traders, is still significantly related to the price aggressiveness and quantity decision.

The approach we take is to start by supposing that there are some history dependent distributions of these order flows, which need to be calibrated by the data, then we jointly model the price, time, order size and order type. In addition, the model simultaneously considers events' types. However, such a model is difficult to estimate directly due to the need for very detailed data, which are not available to us. Hence we use the "indirect inference" approach, a simulation-based method for estimation. Its main idea is to connect the original model to the data through an auxiliary model. This estimation approach was first introduced by Smith (1990, 1993) and later extended in important ways by Gourieroux, Monfort and Renault (1993) and Gallant and Tauchen (1996). There have been many interesting applications of indirect inference to the estimation of economic and financial models. Because of its flexibility, indirect inference can be a useful way to estimate some complex models. It does not, however, appear to ever have been used to estimate LOB models.

3 Models and Data

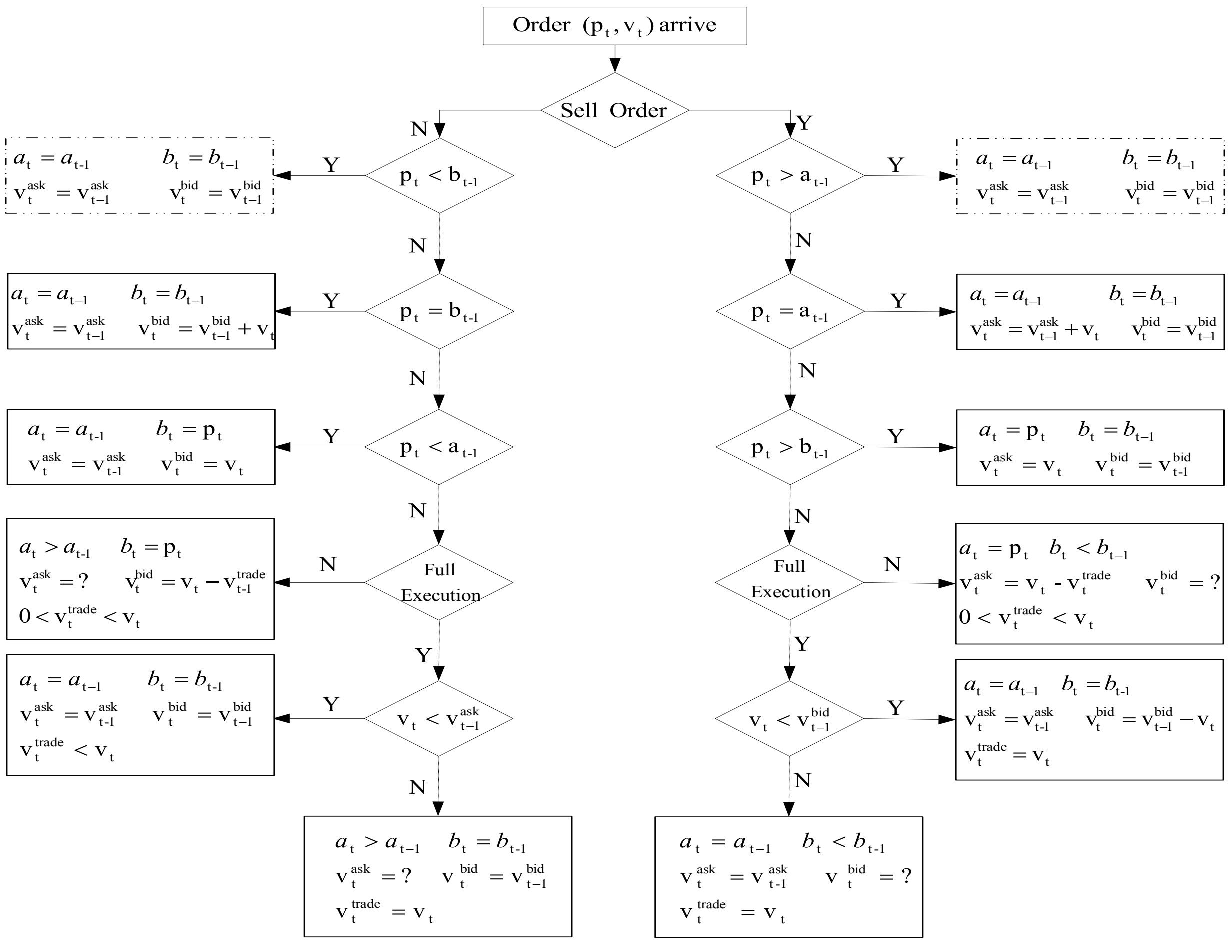

Figure 1: An order arrival

Dashed bracket means that this order will not affect the Level-1 data and therefore will not be recorded in the Level-1 data.

In this section, we first introduce several data sets: the Ideal data, the Level-1 data and the Five-Tick data, where the last two data sets contain only partial information from the first one. Secondly, we propose a model for the Ideal data. The proposed model is based on the idea of distinguishing different types of event, hence it is called the Event-Type model. Then we derive a model for the Level-1 data from the Event-Type model. Finally, we briefly explain the difficulty of directly estimating the Event-Type model based on the Five-Tick data, which calls for some estimation approaches other than maximum likelihood estimation.

3.1 Data

The Ideal data, which we refer to as the complete LOB, consists of depths at each limit order price at any time. Whenever there is a change to any depth, the complete LOB is updated. In addition, whenever there is a trade, information about trade price and size is also recorded.

The Ideal data, a very rich source of information of the trading securities, is not accessible to most traders. Instead, based on their subscription priorities, traders gain access to different levels of data. Most traders have access to the Level-1 data, which only has information about the bid and the ask prices and their depths when any of them changes, and information of trades (trade prices and sizes). Some order arrivals or cancelations may change the complete LOB, while they may not affect the Level-1 data. Figure 1 describes how the Level-1 data will be changed when an order arrives.

The data set we have contains only information of the best five buy and sell prices, which we refer to as the Five-Tick data. It has the following properties:

- Each record in the data contains the depth information at the best five sell and best five buy prices. Trades information is also included.
- The data are aggregated. When more than one event happens within a very short time interval (for example, within less than 100 milliseconds when the market is very active), only the net effect of all the events is recorded in the Five-Tick data. Therefore, there could be more than one depth change in one record.
- The time unit in the Five-Tick data is one second. Several different records may be recorded within one second. Although they are recorded sequentially, the exact time difference between these records is not known.

Table 1 shows a small sample of the 10-year Treasury Note (TNOTE) data , which was traded at CBOT (Chicago Board of Trade) on April 7, 2005. It can be seen that there is more than one record within a one-second interval. In addition, there is more than one event per record, as reflected by the difference between two consecutive records. Another data set we have is International Petroleum Exchange (IPE) data, which is of the same format as TNOTE data.

3.2 Models

3.2.1 The Event-Type Model

In the limit order market, traders' order submission and trading strategies will affect the state of the LOB since traders choose when and at what prices they submit orders. At the same time, the state of the LOB also delivers or disseminates information regarding possible future trades to traders. In other words, traders' trading strategies and the state of limit order book interact with each other. Instead of attempting to anticipate how these interactions work and how traders' orders affect the limit order book, we focus on the evolution of the limit order book by assuming that the aggregated effect of the traders generates order flows with price and order distribution. In the following, we propose a model, which is called the Event-Type model for the complete limit order book within one trading day.

Let $\{t_i\}$ be the point process of times within one trading day when orders either arrive or are

Table 1: 10 year T-note traded on CBOT (4/7/2005)

Bid	Ask	Low	High	Total
08:56:04 1162x109.215	08:56:04 984x109.22	08:56:04 0x109.21	08:56:04 0x109.275	08:56:04 170236x0
08:56:04 1162x109.215	08:56:04 1005x109.22	08:56:04 0x109.21	08:56:04 0x109.275	08:56:04 170236x0
08:56:04 1162x109.215	08:56:04 1158x109.22	08:56:04 0x109.21	08:56:04 0x109.275	08:56:04 170236x0
08:56:04 1162x109.215	08:56:04 1165x109.22	08:56:04 0x109.21	08:56:04 0x109.275	08:56:04 170236x0
08:56:04 1156x109.215	08:56:04 1159x109.22	08:56:04 0x109.21	08:56:04 0x109.275	08:56:04 170242x0
Last[0]	Last[1]	Last[2]	Last[3]	Last[4]
08:56:03 5x109.22	08:56:03 5x109.22	08:56:02 20x109.22	08:56:02 2x109.22	08:56:01 100x109.22
08:56:03 5x109.22	08:56:03 5x109.22	08:56:02 20x109.22	08:56:02 2x109.22	08:56:01 100x109.22
08:56:03 5x109.22	08:56:03 5x109.22	08:56:02 20x109.22	08:56:02 2x109.22	08:56:01 100x109.22
08:56:03 5x109.22	08:56:03 5x109.22	08:56:02 20x109.22	08:56:02 2x109.22	08:56:01 100x109.22
08:56:04 6x109.22	08:56:03 5x109.22	08:56:03 5x109.22	08:56:02 20x109.22	08:56:02 2x109.22
BidDST[0]	BidDST[1]	BidDST[2]	BidDST[3]	BidDST[4]
08:56:04 1162x109.215	08:56:04 4232x109.21	08:56:04 2275x109.205	08:56:04 2015x109.2	08:56:04 2279x109.195
08:56:04 1162x109.215	08:56:04 4232x109.21	08:56:04 2275x109.205	08:56:04 2015x109.2	08:56:04 2279x109.195
08:56:04 1162x109.215	08:56:04 4159x109.21	08:56:04 2275x109.205	08:56:04 2015x109.2	08:56:04 2279x109.195
08:56:04 1162x109.215	08:56:04 4159x109.21	08:56:04 2275x109.205	08:56:04 2015x109.2	08:56:04 2279x109.195
08:56:04 1156x109.215	08:56:04 4159x109.21	08:56:04 2275x109.205	08:56:04 2015x109.2	08:56:04 2279x109.195
AskDST[0]	AskDST[1]	AskDST[2]	AskDST[3]	AskDST[4]
08:56:04 984x109.22	08:56:04 3014x109.225	08:56:04 3642x109.23	08:56:04 2222x109.235	08:56:04 2564x109.24
08:56:04 1005x109.22	08:56:04 3014x109.225	08:56:04 3642x109.23	08:56:04 2222x109.235	08:56:04 2564x109.24
08:56:04 1158x109.22	08:56:04 2993x109.225	08:56:04 3642x109.23	08:56:04 2222x109.235	08:56:04 2564x109.24
08:56:04 1165x109.22	08:56:04 2993x109.225	08:56:04 3642x109.23	08:56:04 2222x109.235	08:56:04 2564x109.24
08:56:04 1159x109.22	08:56:04 2993x109.225	08:56:04 3642x109.23	08:56:04 2222x109.235	08:56:04 2564x109.24

All the columns are of the format "time size × price". Columns named "Bid" and "Ask" are the information about the bid price, the ask price and their depths. " Low " and "High" are respectively the lowest and highest traded price up to the recorded time. " Total" is the total traded order size up to the recorded time. "Last[0]" is the last traded price and sizes, "Last[1]" is the traded price and sizes before "Last[0]", "Last[n]" is the traded price and sizes before "Last[n-1]" and so on. BidDST[n] indicated the depth at $(n+1)^{th}$ best buy price, which is n ticks less than the bid price in this T-note data, and AskDST[n] is the depth at the $(n+1)^{th}$ best sell price, which is n ticks greater than the ask price in this T-note data. For this T-note data, the tick size is $\frac{5}{320}$. For example, price 109.215 is really $109 + \frac{215}{320}$.

canceled. In order to distinguish these two events, we introduce the random vector E_{t_i} to denote the event type at time t_i:

$$E_{t_i} = \begin{cases} e_1, & \text{if a new order arrives at } t_i \\ e_2, & \text{if an order in the LOB is partially or fully canceled at } t_i. \end{cases}$$

where $e_1 = (1,0)'$ and $e_2 = (0,1)'$. Each event is associated with either the buy side or the sell side of the LOB. The indicator variable Q_{t_i} says which side an event comes from at time t_i, which takes value 1 if the event is related to sell side and 0 otherwise. We make the following assumptions for the model:

A1: Depth at any price is known at time 0[1].

A2: $\{t_i\}$ is an orderly point process, meaning no more than one event happens at each time point. Let $\lambda(t|\mathcal{H}_{t-})$ be its conditional intensity process, where $\mathcal{H}_t$ is the information of the complete LOB up to time t.

A3: $P(E_{t_i} = e_1|\mathcal{H}_{t_i-}) = P_i$ and $P(E_{t_i} = e_2|\mathcal{H}_{t_i-}) = 1 - P_i$ depend on the book history. $P(Q_{t_i} = 1|E_{t_i} = e_j, \mathcal{H}_{t_i-}) = \gamma_i^j$ for $(j = 1,2)$ also depend on the history.

A4: If an order is submitted at time t_i, the price is $p_{t_i} = b_{t_{i-1}} + \Delta h^a_{t_i}$ for a sell order and $p_{t_i} = a_{t_{i-1}} - \Delta h^b_{t_i}$ for a buy order, where a_t and b_t are the ask and bid price at time t, and Δ

[1] In the markets we are modeling, time 0 corresponds to the start of the continuous trading session, not the start of the trading day.

is the tick. Both $h^a_{t_i}$ and $h^b_{t_i}$ are integer-valued random variables whose distributions depend on the history of the LOB. Let $P_a(h^a_{t_i}|\mathcal{H}_{t_i-})$ and $P_b(h^b_{t_i}|\mathcal{H}_{t_i-})$ denote these two conditional probabilities, respectively. The size V_{t_i} of order submitted at time t_i may depend on the order price, the order type and the history of the LOB, its conditional distribution of V_{t_i} follows $P_v(v|p_{t_i}, Q_{t_i}, \mathcal{H}_{t_i-})$.

A5: If a cancelation event happens at time t_i, it happens at the price $p^c_{t_i} = a_{t_{i-1}} + \Delta h^{ac}_{t_i}$ for a sell order cancelation and $p^c_{t_i} = b_{t_{i-1}} - \Delta h^{bc}_{t_i}$ for a buy order cancelation, respectively, where both $h^{ac}_{t_i}$ and $h^{bc}_{t_i}$ are nonnegative integer-valued random variables whose distribution depend on the history of the LOB. Let $P_{ac}(h^{ac}_{t_i}|\mathcal{H}_{t_i-})$ and $P_{bc}(h^{bc}_{t_i}|\mathcal{H}_{t_i-})$ denote these two conditional probabilities, respectively. The proportion of the size canceled relative to the depth at the limit price where the cancelation happens is assumed to be distributed with conditional density $f^a_{cp}(p_c|p^c_{t_i}, \mathcal{H}_{t_i-})$ and $f^b_{cp}(p_c|p^c_{t_i}, \mathcal{H}_{t_i-})$ for the sell order cancelation and the buy order cancelation, respectively, where $0 < p_c \leq 1$.

Under our assumptions, the LOB is a marked point process if we treat the event type, order price and size as "marks" associated with the time that either order submission or cancelation happens. For the proposed Event-Type model, the likelihood function can be derived, which we do in the Appendix.

In the Event-Type model, the order arrival rate, order price and order size submitted, order cancelation rate and cancelation proportion are jointly modeled. This model may be used to describe the evolution of the LOB within one trading day without any significant disturbing, such as the releasing of significant news or reports. Incorporating these exogenous effects into the Event-Type model will not be considered in my thesis, but may be investigated in further research.

3.2.2 Model for the Level-1 Data

Here we derive a model for the Level-1 data from the Event-type model by treating the Level-1 data as the result of a stochastic thinning process from the Ideal data. Let $\{t_i\}$ be the occurrence times of events for the complete LOB, $\{t^*_j\}$ be the occurrence times of events in the Level-1 data, which is a thinned process obtained from the point process $\{t_i\}$, and $\lambda^*(t|\mathcal{F}_{t-})$ be the conditional intensity for the Level-1 data, with filtration $\{\mathcal{F}_t\}$. If t_i is recorded into the Level-1 data, we say t_i is retained; K_{t_i} is the indicator that t_i is retained. The conditional probability of K_{t_i} can be calculated as:

$$
\begin{aligned}
& P(K_{t_i} = 1|\mathcal{H}_{t_i-}) \\
= \; & P_i[P(Q_{t_i} = 1, p_{t_i} \leq a_{t_{i-1}}|E_{t_i} = e_1, \mathcal{H}_{t_i-}) + P(Q_{t_i} = 0, p_{t_i} \geq b_{t_{i-1}}|E_{t_i} = e_1, \mathcal{H}_{t_i-})] \\
& +(1-P_i)[P(Q_{t_i} = 1, p^c_{t_i} = a_{t_{i-1}}|E_{t_i} = e_2, \mathcal{H}_{t_i-}) + P(Q_{t_i} = 0, p^c_{t_i} = b_{t_{i-1}}|E_{t_i} = e_2, \mathcal{H}_{t_i-})] \\
= \; & P_i\gamma^1_i P_a(h^a \leq \frac{S_{t_{i-1}}}{\Delta}|\mathcal{H}_{t_i-}) + P_i(1-\gamma^1_i)P_b(h^b \leq \frac{S_{t_{i-1}}}{\Delta}|\mathcal{H}_{t_i-}) \\
& +(1-P_i)\gamma^2_i P_{ac}(h^{ac}_{t_i} = 0|\mathcal{H}_{t_i-}) + (1-P_i)(1-\gamma^2_i)P_{bc}(h^{bc}_{t_i} = 0|\mathcal{H}_{t_i-}) \qquad (1)
\end{aligned}
$$

where $S_{t_{i-1}}$ is the spread at t_{i-1}. $S_{t_{i-1}}$ can be obtained from the Level-1 data by $S_{t_{i-1}} = S_{t^*_j}$ where t^*_j is the last retained point before time t_i,

In our modeling, we assume that no more than one event happens at a time, hence there can never be a case of infinitely many events within one day, and $\lambda(t|\mathcal{H}_{t-})$ is universally bounded by, say M, so is $\lambda^*(t|\mathcal{H}_{t-})$. Therefore, by the Ogata thinning algorithm (see Appendix), the event occurrence processes for the complete LOB and for the Level-1 data can be thinned from a Poisson process with rate M with retaining probabilities $\frac{\lambda(t|\mathcal{H}_{t-})}{M}$ and $\frac{\lambda^*(t|\mathcal{H}_{t-})}{M}$, respectively. Therefore, we have $\lambda^*(t|\mathcal{H}_{t-}) = \lambda(t|\mathcal{H}_{t-})P(K_t = 1|\mathcal{H}_{t-})$. But to get the likelihood function for the Level-1 data, we need $\lambda^*(t|\mathcal{F}_{t-})$. We imposed some mild assumptions: the Level-data is informative enough such that P_i, P_a, P_b,P_{ac}, P_{bc} and γ_i^j depend only on $\mathcal{F}_{t_i-}$, and $\lambda(t|\mathcal{H}_{t-}) = \lambda(t|\mathcal{F}_{t-})$. Thus, the retaining probability $P(K_{t_i} = 1|\mathcal{H}_{t_i-}) = P(K_{t_i} = 1|\mathcal{F}_{t_i-})$ depends only on the Level-1 data, and we have $\lambda^*(t|\mathcal{F}_{t-}) = \lambda(t|\mathcal{F}_{t-})P(K_t = 1|\mathcal{F}_{t-})$. Under these assumptions, we can derive the likelihood function for the Level-1 data (see Appendix), which has a similar form as the one for the Ideal data.

3.3 Tension between Model and Real Data

In principle, since the Five-Tick data consists of partial information of the complete LOB, one may derive a likelihood function from the one for the complete LOB by incorporating all possible event flows that could generate our data set. However, in practice, this is very difficult. For example, the number of events may be much more than the number of depth changes, and may not be deducible from the Five-Tick data. It is not only very difficult, but almost numerically infeasible to incorporate every possible number of events whose combined effect can generate the same record, as there could be arbitrarily many. Besides, the exact occurrence times of events and the sequential orders of these events sometimes can not be deduced from the Five-Tick data. In addition, events that take place at prices other than the best ten prices can not be identified from the Five-Tick data. Because the Five-Tick data contains only aggregated information, which does not meet high requirements of the Event-type model, estimation of the Event-type model directly using such data set will be intractable.

4 Estimation Approach - Indirect Inference

As explained before, it is infeasible to directly estimate the Event-Type model with the Five-Tick data. Instead, we use a simulation based method - indirect inference - to estimate this model. The fully specified probabilistic structure of the Event-Type model allows us to simulate a complete LOB, from which a Five-Tick data can be extracted and compared with the real Five-Tick data via some auxiliary models. In this section, we first explain indirect inference in detail. Then we state the auxiliary models used in parametric indirect inference. Finally we state our functional indirect inference, a nonparametric approach, and use it to estimate our Event-Type model.

4.1 Introduction to Indirect Inference

Indirect inference is a simulation-based method, which estimates the original model via an auxiliary model. This approach only requires that it be possible to simulate data from the original model for different values of its parameters, and therefore is especially useful when the likelihood function (or other criterion function used to estimate the original model) is intractable. Its principle is the following: for each choice of the parameters in the original model, a simulated data series is generated. The auxiliary model, which can be a misspecified model, is then estimated on both the real data and the simulated data, and compatibility between these two data sets may be gauged by the closeness of the corresponding auxiliary parameter estimates. Indirect inference chooses the parameters that minimizes the distance between the auxiliary parameters estimated from the real data and the simulated data series.

To illustrate how indirect inference works, we cite details from Gourieroux and Monfort (1996). Suppose that the original model(M^O) takes the form: $y_t = F(\underline{y_{t-1}}, \underline{z_t}, u_t; \theta)$, where $\{y_t\}_{t=1}^T$ is a sequence of observed endogenous variables, $\underline{y_{t-1}} = (y_0, y_1, ..., y_{t-1})$, $\{z_t\}_{t=1}^T$ is a sequence of observed exogenous variables, $\underline{z_t} = (z_0, z_1, ..., z_t)$, and $\{u_t\}_{t=1}^T$ is a sequence of unobserved random errors whose distribution G is known. For convenience, we define the auxiliary model (M^A) by a conditional probability density function $f^a(y_t|\underline{y_{t-1}}, \underline{z_t}; \beta)$. The parameters of model (M^A) can be estimated by

$$\hat{\beta}_T = \underset{\beta}{\operatorname{argmax}} \sum_{t=1}^{T} \log f^a(y_t|\underline{y_{t-1}}, \underline{z_t}; \beta).$$

In parallel, we pick a parameter vector θ and generate a simulated sequence of endogenous variables $y_t^s(\theta)$ using model (M^O). We replicate such simulations S times. Then the estimated auxiliary parameters are obtained on the simulated series by

$$\hat{\beta}_{ST}(\theta) = \underset{\beta}{\operatorname{argmax}} \sum_{s=1}^{S} \sum_{t=1}^{T} \log f^a(y_t^s|\underline{y_{t-1}^s}, \underline{z_t^s}; \beta),$$

which is a function of θ. This $\hat{\beta}_{ST}(\theta)$ can be replaced by $\frac{1}{S} \sum\limits_{s=1}^{S} [\underset{\beta}{\operatorname{argmax}} \sum\limits_{t=1}^{T} \log f^a(y_t^s|\underline{y_{t-1}^s}, \underline{z_t^s}; \beta)]$.
The final step in the indirect inference is to choose $\hat{\theta}$ such that $\hat{\beta}_{ST}(\theta)$ and $\hat{\beta}_T$ are as close as possible, where their distance is measured by

$$\hat{\theta}_{ST}(\Omega) = \underset{\theta}{\operatorname{argmin}} [\hat{\beta}_T - \hat{\beta}_{ST}(\theta)]' \Omega [\hat{\beta}_T - \hat{\beta}_{ST}(\theta)], \tag{2}$$

and Ω is a predefined symmetric nonnegative definite matrix.

The above approach can be generalized to estimate auxiliary parameters based on maximizing a criterion function (auxiliary function) ψ_T satisfying some conditions, then the auxiliary parameters for the real data and the simulated data are estimated by

$$\hat{\beta}_T = \underset{\beta}{\operatorname{argmax}} \, \psi_T(\underline{y_T}, \underline{z_T}; \beta),$$

$$\hat{\beta}_{ST}(\theta) = \underset{\beta}{\operatorname{argmax}} \sum_{s=1}^{S} \psi_T(\underline{y_T^s}(\theta), \underline{z_T}; \beta).$$

This $\hat{\beta}_{ST}(\theta)$ can be replaced by $\frac{1}{S} \sum\limits_{s=1}^{S} \underset{\beta}{\operatorname{argmax}} \, \psi_T(\underline{y_T^s}(\theta), \underline{z_T}; \beta)$. Then minimizing the above distance metric, we get the indirect inference estimator.

Intuitively, the quality of indirect inference may depend on how well the auxiliary model approximates the original model. In fact, such intuition is formalized by Gallant and Tauchen (1996). They showed that when the auxiliary model is correctly specified, the indirect inference estimator is as efficent as maximum likelihood. Therefore, from the perspective of efficiency, it is important to choose an auxiliary model that could provide a good statistical description of the observed data.

To do the indirect inference, we need to estimate the auxiliary parameters, either parametrically or nonparametrically. We say "parametric" indirect inference when these auxiliaries are estimated parametrically, such as, maximizing a (quasi-)likelihood function; otherwise we call it nonparametric indirect inference. In the following, we will use both parametric and nonparametric indirect inference to estimate the Event-Type model.

4.2 Parametric Indirect Inference

In our parametric indirect inference, we use some parametric predictive models for the Five-Tick data as auxiliary models. These predictive models for the Five-Tick data focus on two critical price processes: bid and ask prices. Let $\{r_t^b = \ln b_t - \ln b_{t-1}\}$ and $\{r_t^a = \ln a_t - \ln a_{t-1}\}$ be the two logarithmic price changes (we may simply refer to them as price changes sometimes) for the bid price process $\{b_t\}$ and the ask price process $\{a_t\}$, respectively. Our focus is to model the conditional mean of the two price change processes, instead of their conditional distribution.

Firstly, we need to specify what time index is used here. Our Five-Tick data set is recorded at irregularly spaced time intervals. Thus, the standard time series techniques relying on fixed time intervals can not be applied. We solve this problem by sampling the original data at regular intervals with a chosen length (use the nearest record as approximate if there is no record at some regular time intervals). As to the choice of the length of the time interval, we use the average time gaps of bid and ask price changes. The sensitivity of our results to the choice of the interval length will be investigated.

Both $\{r_t^a\}$ and $\{r_t^b\}$ have significantly negative first-lag autocorrelation in the Five-Tick data, but have strong positive cross-correlations. Hence, vector autoregressive models (VAR) can be considered as baseline models. For convenience, we use Y_t to stand for (r_t^a, r_t^b) at time t. The VAR model is

$$Y_t = \sum_{k=1}^{p} A_k Y_{t-k} + E_t \tag{3}$$

where E_t is the vector of error terms, and A_k's are matrices. This only uses the best quotes. For the Five-Tick data, additional information is available to predict future prices, such as the imbalance between the buy and sell order sizes (the difference between the depths at some buy and sell limit prices). We use several variables (see Appendix), similar to Cao et al (2004) and Pascual and Veredas (2004), which summarize some information from the LOB, and use Z_t to denote the vector of these variables at time t. Incorporating these covariates Z_t into the VAR model, we have the following vector autoregressive models with exogenous variables model (VARX):

$$Y_t = \sum_{k=1}^{p} A_k Y_{t-k} + BZ_{t-1} + E_t. \tag{4}$$

The estimation procedures for VAR and VARX models are quite standard and will not be explained in this proposal. These coefficient matrices are the auxiliary parameters in our parametric indirect inference. The estimator for our Event-Type model will be chosen such that the difference between these VAR(X) coefficients estimated from the Five-Tick data and from the simulated data is minimized.

4.3 Functional Indirect Inference

As compared with parametric indirect inference, we use kernel-based nonparametric predictive time series models as auxiliary models in our nonparametric indirect inference. The auxiliary parameters in our nonparametric indirect inference are kernel regression functions. These functional forms estimated from the simulated data and the Five-Tick data are compared to obtain the indirect inference estimator for our Event-Type model, thus we call this new technique - functional indirect inference.

4.3.1 Nonparametric Auxiliary Models

As in parametric indirect inference, we still consider the predictive models for the bid and ask price processes as auxiliary models. The difference is that our focus now is nonparametric predictive models. Since these two processes are the results of interaction among a large number of heterogeneous market participants. It would be a surprise if all these complex interactions were to average out in a linear way, and simple VAR(X) models would provide a good description of the real data. Therefore, it is natural to use nonparametric methods. Instead of predicting the conditional distribution of the bid and ask price changes Y_t, we focus on predicting the conditional mean $E(Y_t|Y_{t-1},..,Y_1)$ by a nonparametric approach.

We approximate the conditional changes by

$$E(Y_t|Y_{t-1},..,Y_1) \approx f(Y_{t-1},...,Y_{t-k}), (t > k), \tag{5}$$

where function f and integer k in (5) are estimated using kernel-based nonparametric approach. When an output V_t is regressed on a vector input U_t, the general form of the kernel estimate of the regression function is

$$r_n(u, h_n) = \sum_{t=1}^{n} \frac{K(\frac{u-U_t}{h_n})}{\sum\limits_{s=1}^{n} K(\frac{u-U_s}{h_n})} V_t \tag{6}$$

where K is the kernel function and $h_n \in \mathbb{R}$ is the bandwidth. For our time-series prediction, the input $U_t = (Y_t, \ldots, Y_{t+k-1})$, the output $V_t = Y_{t+k}$, and $n = N - k$. However, k is unknown. Cross-validation is applied to estimate the integer k and bandwidth h. Let $CV(k,h)$ be the cross-validation score for k and h,

$$CV(k,h) = \frac{1}{N-k} \sum_{t=k+1}^{N} ||Y_t - r_{(-t)}(U_t, h)||^2. \tag{7}$$

For each fixed k within $[1, k_0]$, we choose bandwidth by $\hat{h}_k = \underset{0<h\leq h_0}{\operatorname{argmin}} CV(k,h)$. We define $CV(k) = CV(k, \hat{h}_k)$, and estimate k by $\hat{k}$: $\hat{k} = \underset{1\leq k\leq k_0}{\operatorname{argmin}} CV(k)$. For such estimator $\hat{k}$, we use $h_{n-\widehat{k}}$ to denote the optimal bandwidth $\hat{h}_{\hat{k}}$. Then the estimator for (5) is given by $\widehat{E}(Y_t|Y_{t-1},..,Y_1) = \widehat{f}(Y_{t-1},...,Y_{t-\widehat{k}}) = r_{N-\widehat{k}}(U_{t-\widehat{k}}, h_{n-\widehat{k}})$

Parallel to the VARX model in parametric indirect inference, we may incorporate the effect of covariates $\{Z_t\}$ and consider the model

$$E(Y_t|Y_{t-1},..,Y_1, Z_{t-1},...,Z_1) = f(Y_{t-1},...,Y_{t-k}, Z_{t-1}) \tag{8}$$

This model assumes that, in addition to the conditional stationarity of $\{Y_t\}$, only lag-1 covariates may have non-negligible effect on the price changes. The above nonparametric prediction methodology can also be used with covariates. Now $\widehat{E}(Y_t|Y_{t-1},..,Y_1, Z_{t-1},...,Z_1) = \widehat{f}(Y_{t-1},...,Y_{t-\widehat{k}}, Z_{t-1})$ $= r_{N-\widehat{k}}(U_{t-\widehat{k}}, h_{N-\widehat{k}})$ and $U_t = (Y_t,...,Y_{t+k-1}, Z_{t+k-1})$.

4.3.2 Estimation Procedure for Functional Indirect Inference

We present the functional indirect inference estimation for our Event-Type model in this subsection. The presented procedures are for the functional indirect inference that uses model (5) or, similarly, model (8), as the auxiliary model. Our functional indirect inference procedure for the Event-Type model is the following:

(1) Based on the observed bid and ask price changes $\{Y_t\}$, we use the kernel-based nonparametric approach to estimate model (5). $\hat{\beta}$ denotes the estimated auxiliary parameters, that is, $\hat{\beta} = (\hat{f}(\cdot), \hat{k}, \hat{h})$, where $\hat{E}(Y_t|\underline{Y_{t-1}}) = \hat{f}(Y_{t-1}, ..., Y_{t-\hat{k}}) = r_{n-\hat{k}}((Y_{t-\hat{k}}, ..., Y_{t-1}), \hat{h})$ is the kernel estimator for model (5), $\hat{k}$ is the estimated lag, $\hat{h}$ is the estimated optimal bandwidth, and $\underline{Y_{t-1}}$ stands for the history of price change processes up to time $t-1$.

(2) For a chosen parameter vector $\theta \in \Theta$, simulate S paths of the bid and ask price changes $\{Y_t^s(\theta)\}$ based on Event-Type model (M^O) $(s = 1, ...S)$. We use the same kernel function as in the first step, and estimate model (5) for the simulated data. The estimated auxiliary parameters $\hat{\beta}^s(\theta) = (\hat{r}^s(\cdot, \theta), \hat{k}^s(\theta), \hat{h}^s(\theta))$, where the estimator of the function of model (5) based on this simulated data is denoted by $\hat{r}^s(\cdot, \theta) = r_{n-\hat{k}^s(\theta)}(\cdot, \hat{h}^s(\theta))$, and $\hat{k}^s(\theta)$ and $\hat{h}^s(\theta)$ are the estimated lag and bandwidth, respectively.

(3) We use $Pd(g, \{a_t\})$ to denote the sequence of predicted values for $\{a_t\}$ using the predictor g. So $Pd(\hat{\beta}, \{Y_t\}_{t=1}^N)$ is the sequence of predicted values based on the model associated with auxiliary parameters $\hat{\beta}$ for data set $\{Y_t\}_{t=1}^N$ (note that $Pd(\hat{\beta}, \{Y_t\}_{t=1}^N)$ is of length $N - \hat{k}$). Correspondingly, we can obtain $Pd(\hat{\beta}^s(\theta), \{Y_t\}_{t=1}^N)$, $Pd(\hat{\beta}, \{Y_t^s(\theta)\}_{t=1}^N)$, and $Pd(\hat{\beta}^s(\theta), \{Y_t^s\}(\theta)_{t=1}^N)$, which are sequences of length $N - \hat{k}$, $N - \hat{k}^s(\theta)$ and $N - \hat{k}^s(\theta)$ respectively. We introduce the following error statistic

$$D(\theta) = \sum_{s=1}^{S} \frac{[||Pd(\hat{\beta}, \{Y_t\}) - Pd(\hat{\beta}^s(\theta), \{Y_t\})||_*^2 + ||Pd(\hat{\beta}, \{Y_t^s(\theta)\}) - Pd(\hat{\beta}^s(\theta), \{Y_t^s\}(\theta))||_*^2]}{N - \max(\hat{k}, \hat{k}^s(\theta))} \tag{9}$$

where $||\cdot||_*$ denotes the Euclidean distance between the common parts of the two sequences.

(4) The indirect inference estimator $\hat{\theta}_{SN}$ is chosen by

$$\hat{\theta}_{SN} = \underset{\theta}{\operatorname{argmin}}\, D(\theta) \tag{10}$$

which relies on the replicate number of simulation S and the number of observations N.

The logic behind our functional indirect inference estimation is the following. Assuming that there exists a unique $\theta_0 \in \Theta$ of Event-Type model (M^O), such that the observed data set $\{Y_t\}$ evolves like a simulated path $\{Y_t^*(\theta_0)\}$ from model (M^O), the estimated auxiliary parameters from these two time series should be very close, that is, $\hat{\beta}(\theta_0) \approx \hat{\beta}$. In particular, the predictions $Pd(\hat{\beta}, \{Y_t\})$ and $Pd(\hat{\beta}(\theta_0), \{Y_t\})$, $Pd(\hat{\beta}, \{Y_t^*(\theta_0)\})$ and $Pd(\hat{\beta}(\theta_0), \{Y_t^*(\theta_0)\})$ should be very similar. Hence if such θ_0 exists, our defined $D(\theta)$ should be able to pick it up and attain its minimum at $\theta = \theta_0$.

5 Future Work

The goal of this thesis is to estimate the Event-Type model for the limit order book. Due to the limitation of our Five-tick data, this model can not be directly estimated. We apply an indirect inference approach to estimate our model instead. The model, the data and the estimation methodology have already been introduced above. Here, we state the detailed work that needs to be done.

5.1 Conditions Checking for Indirect Inference

To implement indirect inference, there are several requirements. First, the original model should be able to be used to simulate data for different parameter values ("Feasibility requirement"). Second, the auxiliary parameter space should be rich enough to distinguish the parameters in the original model. ("Identification requirement"). Finally, an auxiliary model should be chosen so that a consistent estimator of the original model can be obtained ("Consistency requirement").

The complete probabilistic structure of the Event-Type model allows us to simulate a limit order book. Therefore, the feasibility requirement is satisfied. As to the identification requirement, for different parameter vectors in the Event-Type model, it is unlikely that the bid and ask price processes extracted from the simulated Five-Tick data would exhibit the same probabilistic structure. Therefore they will lead to different estimated VAR(X) models and kernel regression functions.

The major and important task is to verify the consistency requirement for both parametric and functional indirect inference. In the case of parametric indirect inference, Gourierous and Monfort (1996) showed that some regularity conditions for auxiliary models are needed to ensure the consistency of parametric indirect inference estimators. Hence we will validate whether our auxiliary VAR(X) models satisfy these regularity conditions. In the case of functional indirect inference, there are no theorems that we can use directly to verify consistency. Therefore, one challenge work in our functional indirect inference estimation is that consistency needs to be proven theoretically, presumably under some specific regularity conditions. Moreover, we will use simulation to provide evidence of the consistency of our functional indirect inference estimator. The main steps are as follows:

(1) Pick a $\theta^* \in \Theta$ and treat it like the true value for the Event-Type model. Fix an integer S.

(2) Simulate a $\{Y_t\}_{t=1}^N$ from the Event-Type data, and treat this sequence as the observed data.

(3) Perform the functional indirect inference method to get an estimator $\hat{\theta}_{SN}$.

(4) Let N increase gradually, then repeat (2) and (3) for these different N.

(5) Check whether the distance between $\hat{\theta}_{SN}$ and θ^* gradually approaches zero as N increases.

5.2 Implementing Indirect Inference

There are three main steps in our indirect inference estimation: simulation of the Event-Type model, estimation of auxiliary models, and optimization (minimization) of parameter distance functions. The estimation procedures for parametric auxiliary models (VAR/ VARX) are standard. As for nonparametric auxiliary models, the estimation methodologies have been explained in detail in the previous section. Hence we will only address the work related to the other two steps.

5.2.1 Simulation from Event-Type Model

To simulate a limit order book, we need to specify the full probabilistic structure. As an initial trial, we will use a simple structure, where most distributions are path independent. Simple constant distributions are used for event types and event directions. Since almost all relative limit prices in order arrivals and cancelation concentrate on integers within some range, we assign constant probabilities for relative limit price at each of these integers. As to the order size, we use the fat-tailed Pareto distribution. The uniform distribution is used for the cancelation proportion distribution.

One major task is to simulate a point process with a given conditional intensity function. In our model, a self-exciting intensity function is used, $\lambda(t|\mathcal{H}_{t-}) = \lambda_0 + \xi \sum_{i=0}^{\infty} g(t - t_i)$, where g is a scaled beta density with shape parameters α and β and effect window δ; δ, α, β, ξ and λ_0 are all constants. Such an intensity function makes durations between events dependent, and captures the delayed effects of events. We use the Ogata thinning algorithm (Y. Ogata (1981)), which only requires that the intensity be bounded sequentially by constants, true for our intensity function, to simulate the event occurrence process for the LOB.

We have already simulated a LOB based on the simple structure. The number of parameters is fairly high, which already makes our optimization step difficult. So only if the initial structure fails to provide a good model of the LOB, will we consider complicated path-dependent probabilistic structures.

5.2.2 Optimization

To get the indirect inference estimator, we need to minimize the distance functions (2) and (9). Since the number of parameters in the Event-Type model is big, it is infeasible to use an exhaustive-search minimization method. Instead, we use the simulated annealing algorithm, which is an effective optimization method to locate a good approximation to the global optimum of a given function in a large search space (for details of the algorithm, see Belisle (1992)). Other global optimization may be considered if the simulated annealing algorithm performs poorly.

Currently, we are using R for both the simulation and optimization. One future work is to port the computations to C/C++, so that the computational efficiency of the indirect inference estimation could be increased.

5.3 Model Specification Checking

The specification checking for our Event-Type model will be conducted in several ways: stylized facts, pattern matching, and quantitative testing.

(1) Stylized facts.
There are some stylized facts that are commonly observed in financial high-frequency data (see Rama Cont (2001)). Limit order book, one special type of high-frequency data, should present some of these stylized facts. Furthermore, we do find some of these stylized facts in our Five-tick data set, such as, heavy tails of distribution of returns, absence of autocorrelations of returns for long time scales, and slow decay of autocorrelation in absolute returns. We will use these stylized facts observed in our Five-tick data to test our Event-Type model. Specifically, if the LOB simulated from our Event-Type model does not exhibit these stylized fact, the model is misspecified.

(2) Pattern matching.
Even if all these testable stylized facts appear in our simulated LOB, we can not conclude that our model is correctly specified. Simulated Five-tick data (extracted from simulated LOB) should also show patterns similar to the real data. The patterns used in our test include: distribution of traded prices and sizes, distributions of spreads, bids, asks, and relative limit prices and their autocorrelations, distributions of depths at each limit price, and distributions of shapes of depth profile at some random chosen time.

(3) Quantitative testing.
Model specification test will also be conducted quantitatively. We use $MSR(\hat{\beta}, \{Y_t\})$ to denote

the mean square residual between $\{Y_t\}$ and $\{Pd(\hat{\beta}, \{Y_t\}\}$, similarly, for $MSR(\hat{\beta}^s(\theta), \{Y_t\})$, $MSR(\hat{\beta}, \{Y_t^s(\theta)\})$, and $MSR(\hat{\beta}^s(\theta), \{Y_t^s(\theta)\})$. In addition, we define the cross mean square residuals,

$$MSR(\hat{\beta}^s(\theta), \{Y_t^{s'}(\theta)\}) = ||\{Y_t^{s'}(\theta)\}_{t=1}^N - Pd(\hat{\beta}^s(\theta), \{Y_t^{s'}(\theta)\})||_*^2/(N - \hat{k}^s(\theta)) \quad (s \neq s').$$

If the Event-Type model is correctly specified, then there exists a θ_0 such that the model $(M^E(\theta_0))$ corresponds to the observed time series $\{Y_t\}$. Therefore, $\{Y_t\}$ and $\{Y_t^s(\theta_0)\}(s = 1, ..., S)$ will lead to similar functional estimators in model (5). Hence, conditioning on θ_0,

$$MSR(\hat{\beta}, \{Y_t\}) \sim \{MSR(\hat{\beta}^s(\theta_0), \{Y_t^s(\theta_0)\}), s = 1, ..., S\} \tag{11}$$

In order to perform the test, we will use $S \gg 20$. If (11) holds, $MSR(\hat{\beta}, \{Y_t\})$ should fall into the 95% confidence interval of the empirical distribution calculated from the RHS of (11). However, since θ_0 is unknown exactly, we can not obtain accurate values of RHS of (11), but only estimated values. Hence, when implementing, we may use a wide confidence interval to compensate for the uncertainty of θ_0.
As to the cross mean square residuals, conditioning on θ_0, for each $s' = 1, ..S$,

$$\{MSR(\hat{\beta}, \{Y_t^s(\theta_0)\}), s = 1, ..., S\} \sim \{MSR(\hat{\beta}^{s'}(\theta_0), \{Y_t^s(\theta_0)\}), s = 1, ..., S, s \neq s'\} \tag{12}$$

$$\{MSR(\hat{\beta}^s(\theta_0), \{Y_t\}), s = 1, ..., S\} \sim \{MSR(\hat{\beta}^s(\theta_0), \{Y_t^{s'}(\theta_0)\}), s = 1, ..., S, s \neq s'\} \tag{13}$$

That is, if such θ_0 exists, then the distribution of the cross MSR between the simulated data and its predicted values using the predictor estimated from the real data, and the one between the real data and its predicted values using the predictor estimated form the simulated data should both be representative of the cross MSR between the simulated data and its predicted values using predictors estimated from another simulated data. We will find a suitable approach to test these implications.

If the Event-Type model behaves reasonably well in all three tests, we would like to believe that it is a suitable model for the LOB market.

In summary, this thesis provides a new way to model the evolution of the limit order book, which considers the joint effects of all the order flows in the market. To overcome the difficulty of the model estimation due to the limitation of Five-tick data, the indirect inference approaches are used. Moreover, we provide a new nonparametric indirect inference method - functional indirect inference, for the estimation. A theory of the consistency of our functional indirect inference estimator will be shown, and also used to guide our estimation procedure. As to practical applications, our model, if proven to be correctly specified, will help market participants to better understand the dynamics of the LOB. In addition, our estimation approach will help them to quantify the processes of the LOB, even using only Five-tick data, so that some predictable results can be used to improve their trading strategies.

A Appendix

A.1 The Likelihood Function for the Event-type Model

For an orderly point process $\{t_i\}_{i=1}^n$ in $[0,T]$ with conditional intensity $\lambda(t|\mathcal{H}_{t-})$ $(0 = t_0 < t_1 < \cdots < t_n \le T\ )$, the log-likelihood function is

$$\begin{aligned}
\ln \mathcal{L}(\lambda) &= \int_0^T (1 - \lambda(s|\mathcal{H}_{s-}))ds + \int_{(0,T]} \ln \lambda(s|\mathcal{H}_{s-}))dN(s) \\
&= \int_0^T (1 - \lambda(s|\mathcal{H}_{s-}))ds + \sum_{i=1}^n \ln \lambda(t_i|\mathcal{H}_{t_i-})
\end{aligned} \tag{14}$$

If we treat the event type, order price, order size submitted and cancelation size as "marks" associated with the time that an event happens, the LOB is just a marked point process. Based on our model assumption, the likelihood function of our marked point process can be specified (see [11]),

$$\mathcal{L} = \mathcal{L}(\lambda) \prod_{i=1}^n \langle (F_{i1}, F_{i2})', e_{t_i} \rangle \tag{15}$$

where $\langle \cdot, \cdot \rangle$ is an inner product of vectors, $\mathcal{L}(\lambda)$ is the likelihood function regarding λ's, that are of the form as equation (14), and

$$\begin{aligned}
F_{i1} &= P_i[\gamma_i^1 P_a(h_{t_i}^a = \frac{p_{t_i} - b_{t_{i-1}}}{\Delta}|\mathcal{H}_{t_i-})]^{q_{t_i}}[(1-\gamma_i^1)P_b(h_{t_i}^b = -\frac{p_{t_i} - a_{t_{i-1}}}{\Delta}|\mathcal{H}_{t_i-})]^{1-q_{t_i}} \\
&\quad \cdot P_v(v_{t_i}|p_{t_i}, Q_{t_i} = q_{t_i}, \mathcal{H}_{t_i-}) \qquad (16) \\
F_{i2} &= (1-P_i)[\gamma_i^2 P_{ac}(h_{t_i}^{ac} = \frac{p_{t_i}^c - a_{t_{i-1}}}{\Delta}|\mathcal{H}_{t_i-})]^{q_{t_i}}[(1-\gamma_i^2)P_{bc}(h_{t_i}^{bc} = -\frac{p_{t_i}^c - b_{t_{i-1}}}{\Delta}|\mathcal{H}_{t_i-})]^{1-q_{t_i}} \\
&\quad \cdot f_{cp}(p_{ct_i}|p_{t_i}, Q_{t_i} = q_{t_i}, \mathcal{H}_{t_i-}) \qquad (17)
\end{aligned}$$

A.2 The Likelihood Function for the Level-1 data

The Level-1 data consists of time series observations which contain: (1) the ask price a_t and the bid price b_t, (2) the depths V_t^{ask} and V_t^{bid} at the ask and the bid, respectively, (3) the trade size V_t^{trade} and the traded price p_t^{trade}[2]. The time series are recorded whenever one of these quantities changes.

In the main text, we argued that the conditional intensity for the Level-1 data $\lambda^*(t|\mathcal{F}_{t-}) = \lambda(t|\mathcal{F}_{t-})P(K_t = 1|\mathcal{F}_{t-})$ under some assumptions. But to derive the likelihood function of the Level-1 data, we need to deduce the price and size information of order submission and cancelation from the Level-1 data. For the order cancelation in the Level-1 data, the prices and sizes are easily deduced. Hence we only address the order submission case. Let $p_{t_i^*}$ and $V_{t_i^*}$ denote the order price and size submitted into the complete LOB, which causes t_i^* to be recorded into the Level-1 data. Suppose that it is a sell order submission, then the price and the size can be deduced as follows:

1. If $a_{t_i^*} = a_{t_{i-1}^*}$ and $V_{t_i^*}^{ask} > V_{t_{i-1}^*}^{ask}$, then this sell order is submitted at ask price $a_{t_{i-1}^*}$, that is, $p_{t_i^*} = a_{t_{i-1}^*}$, with the size $V_{t_i^*} = V_{t_i^*}^{ask} - V_{t_{i-1}^*}^{ask}$.

2. If $a_{t_i^*} = a_{t_{i-1}^*}$ and $V_{t_i^*}^{ask} = V_{t_{i-1}^*}^{ask}$, then this sell order is submitted at price equal to or lower than $b_{t_{i-1}^*}$ and is fully executed, hence $p_{t_i^*} = p_{t_i^*}^{trade}$ and $V_{t_i^*} = V_{t_i^*}^{trade}$.

[2]This price usually is the traded price of the last traded share

3. If $b_{t_{i-1}^*} < a_{t_i^*} < a_{t_{i-1}^*}$, then this sell order brings a new ask price, hence $p_{t_i^*} = a_{t_{i-1}^*}$ and $V_{t_i^*} = V_{t_i^*}^{ask}$.

4. If $a_{t_i^*} = b_{t_{i-1}^*}$, then this sell limit order is submitted at price $b_{t_{i-1}^*}$ with size more than $V_{t_{i-1}^*}^{bid}$, hence $p_{t_i^*} = b_{t_{i-1}^*}$ and $V_{t_i^*} = V_{t_i^*}^{trade} + V_{t_i^*}^{ask} = V_{t_{i-1}^*}^{bid} + V_{t_i^*}^{ask}$.

5. If $a_{t_i^*} < b_{t_{i-1}^*}$, then this sell limit order is submitted with price lower than $b_{t_{i-1}^*}$ and is partially executed, hence $p_{t_i^*} = a_{t_{i-1}^*}$ and $V_{t_i^*} = V_{t_i^*}^{ask} + V_{t_i^*}^{trade}$.

Similarly, we can deduce the price and size of a buy order submission. However, this reasoning is based on the assumption that we know the order type submitted. Therefore the order type submitted needs be inferred before we deduce the order price and size information. One common way to identify the order type is to judge by the change direction of the middle price (the average of the bid and the ask price). If the middle price increases as compared with the previous middle price, then the order submitted is considered as a buy order. If the middle price decreases, then the order is considered as a sell order. It may happen that the middle price remains the same as the previous middle price. When this happens, the bid and ask price do not change according to the data recording properties of the Level-1 data, therefore one of the depths at the bid and ask price must change. If it is the depth at the bid that changes, then we conclude that the order submitted is a buy order, otherwise it is a sell order. Hence, the order types, prices and sizes can be deduced from our Level-1 data.
The likelihood function for the Level-1 data can be similarly derived as the following,

$$\mathcal{L} \quad = \quad \mathcal{L}(\lambda^*) \prod_{i=1}^{m} \langle (F_{i1}^*, F_{i2}^*)', e_{t_i^*} \rangle \tag{18}$$

where F_{i1}^* and F_{i2}^* have the similar forms as (16) and (17), respectively, but with different filtration.

A.3 Ogata thinning algorithm

When simulating a point process with a given intensity function, an effective way is to use a thinning algorithm. Here we introduce the Ogata thinning algorithm, which generalize the Shedler-Lewis thinning algorithm. The Ogata thinning algorithm states that for a point process $(N, \mathcal{F}, P) = \{N_t, \mathcal{F}_t, (0 < t \leq T), P\}$ with conditional intensity $\lambda(t|\mathcal{F}_{t-})$, there exist a positive $\mathcal{F}$-predictable piecewise constant process $M = \{M_t\}$ such that $\lambda_t \leq M_t$ a.s., then the point process can be thinned from a locally homogenous Poisson process $(N^M, \mathcal{F}, P)$ with intensity process $\{M_t\}$. When M_t is a fixed constant for all t, we get the Shedler-Lewis thinning algorithm. The main step of the Ogata thinning algorithm is as follows (cited from Ogata (1981)):
(1) Suppose that the last point t_i before time t has just been obtained. Then construct M_t which is $\mathcal{F}$-predictable, piecewise constant, and $\lambda_t \leq M_t$ for $t \geq t_i$.
(2) Simulate homogeneous Poisson points $t_j^M (> t_i)$ according to the intensity M_t.
(3) For each of the points $\{t_j^M\}$, retain t_j^M with the probability $\lambda_{t_j^M}/M_{t_j^M}$, which is given conditionally independent of t_j^M under the past history $\mathcal{F}_{t_j^M}$.
(4) t_{i+1} is the first accepted point among $t_j^M (> t_i)$.

A.4 Variables related to Information Summarized from the LOB

The following variables summarize some of the information from the limit order book, and are used in Cao et al (2004) and Pascual and Veredas (2004).

V1: The imbalance in the limit order book:
Volume imbalance: $DQ_t^i = \frac{Q_t^{a,i}-Q_t^{b,i}}{Q_t^{a,i}+Q_t^{b,i}}$, where i stands for i-th price level $(i = 1, ..5)$.
Price imbalance: $DP_t^i = \frac{p_t^{a,i+1}-p_t^{a,i}}{p_t^{b,i}-p_t^{b,i+1}}$ $(i = 1, ...4)$.

V2: The shape of the limit order book:
spread: $S_t = a_t - b_t$.
weighted spread: $WS_t = \frac{\sum_{i=1}^{5} Q_t^{a,i}p_t^{a,i}}{\sum_{i=1}^{5} Q_t^{a,i}} - \frac{\sum_{i=1}^{5} Q_t^{b,i}p_t^{b,i}}{\sum_{i=1}^{5} Q_t^{b,i}}$.
weighted prices: $wp_t^i = \frac{Q_t^{a,i}p_t^{a,i}+Q_t^{b,i}p_t^{b,i}}{Q_t^{a,i}+Q_t^{b,i}} - \frac{p_t^{a,1}+p_t^{b,1}}{2}$ $(i = 1, ..5)$.

V3: The changes of variables in V1 and V2.

V4: Trade information:

$$d_t = \begin{cases} 1 & \text{if there is trade at time t with price above the midprice} \\ -1 & \text{if there is trade at time t with price below the midprice} \\ 0 & \text{if there is no trade at time t} \end{cases}$$

A.5 Kernel-based nonparametric prediction method

Let $\{(U_t, V_t)\}_{t=1}^n$ be a $\mathbb{R}^d \times \mathbb{R}$ -valued process. We use $r_n(\cdot)$ to denote the function from $\mathbb{R}^d$ to $\mathbb{R}$ with kernel function K and bandwidth $h_n \in \mathcal{R}$,

$$r_n(u, h_n) = \sum_{t=1}^{n} \frac{K(\frac{u-U_t}{h_n})}{\sum_{s=1}^{n} K(\frac{u-U_s}{h_n})} V_t \tag{19}$$

Let $\{\xi_t\}_{t=1}^N$ be a general $\mathbb{R}^{d_0}$-valued Markov process of order k. To predict $m(\xi_{n+1})$ based on the observed values $\{\xi_1, ..., \xi_n\}$, where m is real-valued measurable function and bounded on compact sets, one can construct the associated process $W_t = (U_t, V_t)$ with $U_t = (\xi_t, ..., \xi_{t+k-1})$ and $V_t = m(\xi_{t+k})$, and consider the kernel regression estimator based on observed $W_t (1 \leq t \leq n-k)$ by the following formula:

$$\widehat{E}[m(\xi_{n+1})|\xi_n, ..., \xi_{n-k}] = r_{n-k}(U_{n-k+1}, h_{n-k}) \tag{20}$$

where $K(\cdot)$ is a chosen kernel function and h_{n-k} is the corresponding bandwidth satisfying some conditions. However, for a general time series, the Markovian assumption may not be satisfied. Carbon and Delecroix (1993) provide a nonparametric predictor to non-Markovian stationary time series. Assuming that $\{\xi_t\}$ is stationary, we know under appropriate theoretical assumptions, we have $E[m(\xi_{t+1})|\xi_t, .., \xi_1]$ and $E[m(\xi_{t+1})|\xi_t, .., \xi_{t-k+1}]$ are close for t large enough and suitable k. Once k is chosen, the above nonparametric prediction method of a Markov process can be used. Bosq(1998) suggests that a suitable k can be chosen to be the one that minimizes

$$\Delta(k) = \sum_{t=k+1}^{N} ||m(\xi_t) - \widehat{m}_k(\xi_t)|| \tag{21}$$

for $1 \leq k \leq k_0$, where k_0 is a predefined bound and $\widehat{m}_k(\xi_t)$ is evaluated using (20).

A.6 Consistency of Indirect Inference Estimator

To get the consistency of the indirect inference estimator, we need some regularity conditions (see Gourieroux and Monfort (1996)),

(A1) $\psi_T(\underline{y_T^s}(\theta), \underline{z_T}; \beta)$ tends almost surely to a deterministic limit function $\psi_\infty(\theta, \beta)$ uniformly in (θ, β) when $T \to \infty$.

(A2) $\psi_\infty(\theta, \beta)$ has a unique maximum with respect to β: $b(\theta) = arg \max_\beta \psi_\infty(\theta, \beta)$. This $b(\theta)$ is referred as **binding function**.

(A3) The equation $\beta = b(\theta)$ admits a unique solution in θ.

References

[1] A. Anand, S. Chakravarty, and T. Martell. Empirical evidence on the evolution of liquidity: Choice of market versus limit orders by informed and uninformed traders. *Journal of Financial Markets*, 8(3):288–308, August 2005.

[2] L. Bauwens and P. Giot. *Econometric modelling of stock market intraday activity*. Kluwer Acadenic Publishers Boston, 2003.

[3] C. Belisle. Convergence theorems for a class of simulated annealing algorithms on $\mathbb{R}^d$. *Journal of Applied Probability*, 29:885–895, 1992.

[4] B. Biais, P. Hillion, and C.Spatt. An empirical analysis of limit order book and the order flow in the paris bourse. *Journal of Finance*, 50:1655–1689, 1995.

[5] R. Bloomfield, M. O'Hara, and G. Saar. The "make or take" decision in an electronic market: Evidence on the evolution of liquidity. *Journal of Financial Economics*, 75(1):165–199, January 2005.

[6] C. Cao, O. Hansch, and X. Wang. The informational content of an open limit order book. EFA 2004 Maastricht Meetings Paper No. 4311 470, Mar. 2004.

[7] S. Chakravarty and C. Holden. An integrated model of market and limit orders. *Journal of Financial Intermediation*, 4:213–241, 1995.

[8] C. Chiarella and G. Iori. A simple microstructure model of double auction markets. *Quantitative Finace*, 2:346–353, 2002.

[9] C. Chiarella and G. Iori. The impact of heterogeneous trading rules on the limit order book and order flows. 2004.

[10] R. Cont. Empirical properties of asset returns: stylized facts and statistical issues. 2001.

[11] D. Daley and D. Vere-Jones. *An Introduction to the Theory of Point Processes*. Springer, 2003.

[12] M. Daniels, J. Farmer, G. Iori, and E. Smith. How storing supply and demand affects price diffusion. 2002.

[13] M. Daniels, J.D.Farmer, L. Gillemot, G. Iori, and E. Smith. A quantitative model of trading and pric efomation in financial markets. 2001.

[14] I. Domowitz and J. Wang. Auctions as algorithms: computerized trade execution and price discovery. *Journal of Economic Dynamics and Control*, 18:29–60, 1994.

[15] D. Easley and M. O'Hara. Price, trade size, and information in securities markets. *Journal of Financial Economics*, 19:69–90, 1987.

[16] D. Easley and M. O'Hara. Time and the process of security price adjustment. *Journal of Finance*, 47:577–606, 1992.

[17] J. Farmer, P. Patelli, and I. Zovko. The predictive power of zero intelligence in financial markets. 2004. AFA 2004 San Diego Meetings.

[18] T. Foucault. Order flow composition and trading costs in a dynamic limit order market. *Journal of Financial Markets*, 2:99C134, 1999.

[19] T. Foucault, O. Kadan, and E. Kandel. Limit Order Book as a Market for Liquidity. *Rev. Financ. Stud.*, 18(4):1171–1217, 2005.

[20] X. Gabaix, P. Gopikrishnan, V. Plerou, and H. Stanley. A theory of power-law distributions in financial market fluctuations. *Nature*, 423:267–270, May 2003.

[21] A. Gallant and G. Tauchen. Which moments to match? *Econometric Theory*, 12(4):657–81, October 1996.

[22] M. Garman. Market microstructure. *Journal of Financial Economics*, 3:257–275, 1976.

[23] L. Glosten. Is the electronic open limit order book inevitable? *Journal of Finance*, 49(4):1127–61, 1994.

[24] R. Goettler, C. Parlour, and U. Rajan. Information acquisition in a limit order market. NYSE working paper, 2004.

[25] R. Goettler, C. Parlour, and U. Rajan. Equilibrium in a dynamic limit order market. *Journal of Finance*, 60:2149–2192, 2005.

[26] C. Gourieroux and A. Monfort. *Simulation-based econometric methods*. Oxford, 1996.

[27] C. Gourieroux, A. Monfort, and E. Renault. Indirect inference. *Journal of Applied Econometrics*, 8:S85–S118, 1993.

[28] P. Handa and R. Schwartz. Limit order trading. *Journal of Finance*, 51:1835–1861, 1996.

[29] L. Harris. Optimal dynamic order submission strategies in some stylized trading problems. *Financial Markets, Institutions & Instruments*, 7:1–76, 1998.

[30] L. Harris and J. Hasbrouck. Market vs. limit orders: The superdot evidence on order submission strategy. *Journal of Financial and Quantitative Analysis*, 31:213–231, 1996.

[31] J. Hasbrouck, G. Sofianos, and D. Sosebee. New york stock exchange system and trading procedures. NYSE working paper, 1993.

[32] J. Hellstrom and O. Simonsen. Does the open limit order book reveal information about short-run stock price movements? Umea Economic Studies 687, Umea University, Department of Economics, 2006.

[33] F. Lillo. Limit order placement as an utility maximization problem and the origin of power law distribution of limit order prices. 2006.

[34] A. Lo, MacKinlay, and J. Zhang. Econometric models of limit order executions. *Journal of Financial Economics*, 65:31–71, 2001.

[35] H. Luckock. A statistical model of a limit order market. 2001.

[36] S. Maslov. Simple model of a limit order-driven market. *PHYSICA A*, 278:571, 2000.

[37] Y. Ogata. On lewis' simulation method for point processes. *Transactions on information theory*, 27(1):23–31, 1981.

[38] C. Parlour. Price dynamics in limit order markets. *Review of Financial Studies*, 11:789–816, 1998.

[39] R. Pascual and D. Veredas. What pieces of limit order book information are informative? an empirical analysis of a pure order-driven market. Mimeo. Dept. of Business. Universidad de las Islas Baleares, 2004.

[40] M. Potters and J. Bouchaud. More statistical properties of order books and price impact. Science & Finance (CFM) working paper archive 0210710, Science & Finance, Capital Fund Management, Oct. 2002. available at http://ideas.repec.org/p/sfi/sfiwpa/0210710.html.

[41] A. Ranaldo. Order aggressiveness in limit order book markets. *Journal of Financial Markets*, 7(1):53–74, January 2004.

[42] I. Rosu. A dynamic model of the limit order book. NYSE working paper, 2005.

[43] J. R. Russell and R. F. Engle. Econometric analysis of discrete-valued irregularly-spaced financial transactions data using a new autoregressive conditional multinomial model. CRSP working papers 470, Center for Research in Security Prices, Graduate School of Business, University of Chicago, Apr. 1998. available at http://ideas.repec.org/p/wop/chispw/470.html.

[44] P. Sandas. Adverse selection and competitive market making: Empirical evidence from a limit order market. *Review of Financial Studies*, 14(3):705–34, 2001.

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

[46] J. Wald and H. T. Horrigan. Optimal limit order choice. *Journal of Business*, 78(2):597–620, 2005.

[47] I. Zovko and J.D.Farmer. The power of patience: a behavioural regularity in limit-order placement. *Quantitative Finace*, 2:387–392, 2002.