Using full limit order book for price jump prediction

Mynbaev, Kairat

New School of Economics, Satbayev University

June 2020

Using full limit order book for price jump prediction

Kairat Mynbaev

New School of Economics
Satbayev University
22a Satpaev str.
Almaty 050013, Kazakhstan
email: kairat_mynbayev@yahoo.com

Abstract. Institutional investors, especially high frequency traders, employ the order information contained in the Limit Order Book (LOB). The main purpose of the paper is to investigate how full information about the LOB can help in predicting the price jump. Normally, a full LOB contains total volumes of orders for hundreds of prices. Using the full information runs into the curse of dimensionality which manifests itself in multicollinearity, insignificant coefficients, inflated estimate variances and high computation time. Due to these problems, order volumes for prices that are distant from ask and bid prices are usually not used in prediction procedures. For this reason we call such information a silent crowd. Here we propose a summary measure of the silent crowd and quantify its influence on trade jump prediction. We use a realistically simulated LOB as a vehicle for experiments and logistic regression as the prediction tool. The full code in Matlab includes 18 blocks.

UDK 519.2, 519.6 Keywords Simulation, trade jump prediction, high frequency trading, logistic regression, limit order book.

AMS-MS Classification. 62F12, 62G07, 62G20.

1 Introduction

The advent of information technologies made possible the transition from quote-driven markets to order-driven trading platforms. On many stock exchanges, including NYSE, NASDAQ, and the London Stock Exchange, trade orders are submitted and executed electronically [1]. Outstanding orders are recorded in what is called a Limit Order Book (LOB). For a fee, clients can have access to either partial or full information contained in the LOB. High speed communications, fast computers and computer algorithms enabled high frequency trading, when orders are submitted every millisecond. Analysing the LOB and making predictions regarding possible market moves in real time is essential for participants of this market.

One direction of research focuses on mathematical models of the LOB [5, 6, 3, 2, 4]. They provide kind of a common denominator for financial phenomena but are too judgmental in the sense that they typically impose restrictions which are hard to validate in practice [7, 8].

On the other hand, machine learning methods do not impose any a priori conditions and attempt to reveal the regularities that are in the data [11, 12, 9, 10]. In particular, statistical methods are used to predict quantities that can be used profitably. The paper [13] presents a non-parametric model for trade sign inference. [14] uses logistic regression to predict occurrence of price jumps. [15, 16] employ support vector machines to capture the dynamics of price movements. [17] suggest a model that describes the evolution of the distribution of limit orders and whose estimates can be used in a regression. [18] analyze the contribution to price discovery of market and limit orders by high-frequency traders (HFTs) and non-HFTs. See the last paper for a valuable review and latest references.

The above references use real-world data. We work with a simulated LOB. The two approaches have different focuses.

The main value of real-world data is that it contains traces of investors' decisions, which are influenced by the shape of the LOB, among other things. The challenge is to infer about investors decisions and use that inference to successfully predict future price movements. This is complicated by many realities: different investors react differently to the market signals contained in the LOB, there are events outside the LOB influencing investors moves and the very invention of successful prediction mechanisms affects investors behaviour.

A simulated LOB should incorporate and exhibit the stylized facts of the real LOB. Different types of orders are posted in accordance with distributional patterns observed in practice, but other than that they are random and independent, at least in our implementation. There are no built-in behavioral assumptions. The simulated LOB is impartial, so to speak. It serves better the purpose of revealing relative importance of quantities contained in the LOB, as opposed to inferring about investors motivations. A real LOB is a snapshot of what has happened, while a simulated LOB can be produced as many times as needed and allows one to fine-tune model parameters to achieve the desired patterns.

In Section 2 we describe the standard features of limit order books. In Section 3 we detail the simulations. Section 4 presents the main results. Section 5 contains conclusions. The Matlab code is available on request.

2 Order types and LOB structure

In order-driven markets investors can submit three order types: limit orders, cancel orders and market orders. The minimum allowed price increment is called a tick. For simulation purposes the tick can be taken to be 1 without loss of generality.

A sell limit order is an order to sell a certain number of shares at a certain price (called ask) or higher. A buy limit order is an order to buy a certain number of shares at a certain price (called bid) or lower. If there is no offsetting order at the same price, a limit order is recorded in the LOB. Limit orders are executed against offsetting incoming orders in the order they (limit orders) were recorded. Limit orders have an expiration date, unless the investor specifies that the order is good until canceled. Order expiration dates are not seen in the LOB investors have access to. For modeling purposes all limit orders are considered as orders with no expiration date.

An investor can cancel his/her limit order (or its remaining part) at any time. In fact, most limit orders are canceled before their execution.

It is useful to imagine the LOB as consisting of two parts, with a vertical price axis. The upper part contains all sell orders, and the lower one contains all buy orders (more precisely, total volumes against each tick). The lowest sell price is called the best ask and the highest buy price is called the best bid. Because of opposite order matching the best ask is always higher than the best bid. The midprice is defined by $midprice = (best\ ask + best\ bid)/2$. The difference $best\ ask - best\ bid$ is called a spread. The prices and total volumes at the best ask and bid are called first level quotes, the prices and total volumes one tick away from the best ask and bid are called second level quotes and so on.

A market sell order is an order to sell a certain number of shares at the best available price, that is at the best bid. Similarly, a market buy order is an order to buy a certain number of shares at the best available price, that is at the best ask. When a market sell order arrives, the total volume at the best bid may be smaller than the market order size. In this case the market order consumes all of the volume at the best bid, the best bid moves down and the remaining part of the market order is executed against the limit orders at the new best bid. Some exchanges use a different rule: if, say, a sell market order size is larger than the outstanding volume at the best bid, the remaining part of the market order stays in the LOB as a sell limit order. The difference between the first case, when the market order may be executed at several prices, and the second one, when it may be partially executed and the remainder stays as a limit order at the best bid, is that in the first case the best bid moves down (and the spread increases), while in the second case it is the best ask that moves down. In the first case the downward move of the midprice is determined by the relative size of the market order and liquidity at the bid side. In the second case this downward move depends on the spread, and the midprice right after execution of the market order will be lower than the best bid right before the execution. The midprice is more stable under the first arrangement, which we adopt in our simulations. Stability of market prices is one of desirable features.

Market orders are executed immediately, so in case of a real LOB, one can know about their arrival and size only from a change in total volumes of limit orders at the best ask and bid. HFT's often place orders just to cancel them a moment later. There also can be errors in the way the LOB is recorded. This kind of problems do not arise with a simulated LOB. Experiments on a real stock exchange are costly and likely to disrupt its operations; in case of a change in rules governing an exchange, large and technologically advanced players will win at the expense of small investors.

All the information above the ask price characterizes the supply, whereas all the information below the bid price characterizes the demand side.

3 Simulation description

The task of modeling the LOB is complex because the impact of an order on the book depends on the state of the book. Therefore one cannot sum the incoming orders over a period of time and post the sum to the book. The orders have to be generated and posted immediately one by one. This requires a lot of calculation, only a small part of which can be made faster using parallel computing. We have not been able to use the CUDA (parallel computing language from NVIDIA$^{\text{TM}}$) because it can handle only specific types of code.

Application of logit requires measuring depths at equally spaced moments, and their number should be large enough. With short time intervals (on the order of several milliseconds) the LOB is too poor. Increasing the lengths of time intervals increases the complexity of calculations.

Following the empirical pattern [7], the distribution of orders is defined in such a way that the spread of limit orders is very large, ±50% of the midprice or more. On both sides of the midprice the distribution declines as a power law, up to 100 ticks from the midprice, and then falls to zero. Orders arrive independently at exponential rates.

Cancel order sizes are given as a fraction of the order depth.

The Matlab code consists of 18 programs. The first character in the program name indicates its level. The lowest-level programs start with A, next-level programs start with B and so on. The level of a program is determined by the references contained in it. For example, the program C_AllOrdersTimesAndPrices.m may refer to levels A and B but not higher.

The function A_InDistr creates the initial distribution of orders.

The function A_OrderTimes generates a sequence of order placement times up to given moment.

The function A_Revert just makes some code more convenient to read.

A_NormConstant realizes an empirically observed pattern in the distribution of orders from [7].

B_OrderTimesFixedPrice generates lists of limit, cancel and market order times (for all price ticks from 1 to $MaxPrice$).

B_AskAndBid finds the best ask (the lowest ask price at which order size is not zero) and the best bid (the highest bid price at which order size is not zero).

The function B_FindCum creates cumulative sums starting from the lower end of B_T. This is the most important part of the method. The silent crowd should be summarized in such a way that the prices close to the midprice should have larger weights. The weights should not be so heavy as to dampen the tail of the silent crowd.

B_LODensity generates sizes of limit orders in the range $(midprice - dist, midprice + dist)$, currently under condition $MaxPrice = 4 * dist$.

C_AllOrdersTimesAndPrices puts into one $MaxPrice \times 3$ matrix M

- all order times from lists of limit, cancel and market orders, unsorted (first column of M),
- order types (1 for Limit, 2 for Cancel, 3 for Market) (second column of M),
- and corresponding prices, numbered 1 through $MaxPrice$ (third column of M).

This is necessary to create a line of orders that later will be posted to the LOB.

Next there are three functions that post three types of orders: C_PostCancelOrder, C_PostLimitOrder and C_PostMarketOrder.

E_Inference_A_B collects statistical characteristics of the LOB. It is important that after about 50 orders the simulated LOB stabilizes and its two-humped shape corresponds to what is observed in practice.

Next we need to see how informative are the prices close to the midprice, compared to the informativeness of the silent crowd.

F_band_A_B finds bands of order sizes of width *band* (*band* up from ask in A_T and *band* down from bid in B_T).

F_weight_A_B prepares weights for averaging order sizes.

Finally, comparison is made between contribution of the prices that are close to the midprice (in the band) and contribution of the silent crowd.

4 Simulation results

The density of incoming limit orders is generated according to what is observed in practice. 200 ticks up and down from the initial midprice the density tapers off. After that, we set it to zero, see Figure 1.

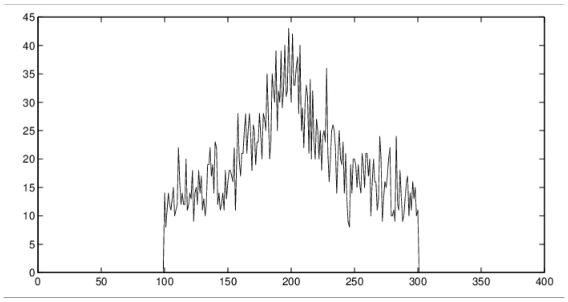

Figure 1: Density of limit orders

As it was mentioned above, after about 50 orders the simulated LOB stabilizes. The midprice falls from the one defined in the initial distribution and afterwards is pretty stable (Figure 2).

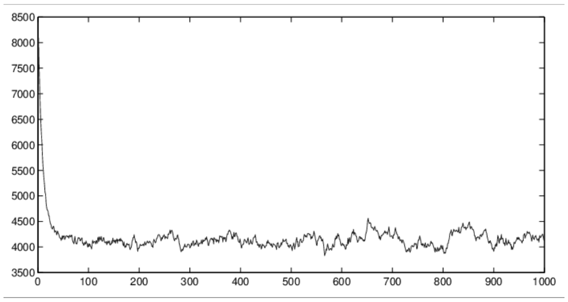

Figure 2: Stabilization of the midprice

The standard deviation of the midprice also stabilizes (Figure 3).

Its two-humped shape corresponds to what is observed in practice, see Figure 4. This is a sign that relative order sizes have been chosen correctly (orders do not accumulate to infinity and are not consumed entirely by incoming buy orders).

Another sign that the LOB is being simulated correctly is that the order lists in the LOB behave pretty irregularly. See in Figure 5 the behavior of the first five ask sizes.

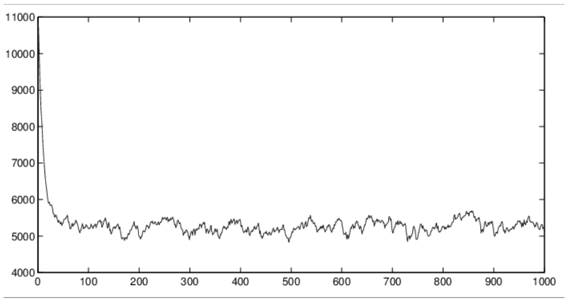

Figure 3: Stabilization of the standard deviation of the midprice

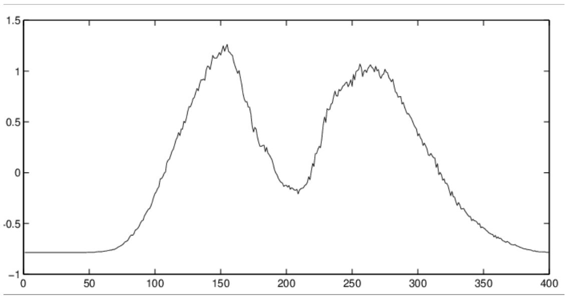

Figure 4: Two-humped distribution of order sizes

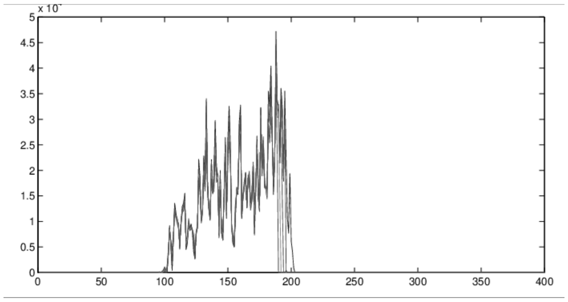

Figure 5: Ask sizes at the first 5 prices

We use the logit model to predict the price jump. This is done with two sets of predictors: one includes only prices close to the midprice and the other additionally includes the index of the silent crowd. Specifically, let a_{it}, b_{it} denote the ask and bid sizes at time t, where $i = 1, 2, ...$ is the quote level. The price jump $j_t = \text{sgn}(midprice_{t+1} - midprice_t)$ is regressed on $a_{it}, b_{it},\ i = 1, ..., I$, in the first regression and on $a_{it}, b_{it},\ i = 1, ..., I, index_{It}$ in the second regression. Here $index_{It} = \sum_{i=I+1}^{MaxPrice} w_i(a_{it} + b_{it})$ is a weighted sum of the representatives of the "silent crowd". We change $I = 1, 2, ..., 24$ to see how the two regressions compare. From Figure 6 it is clear that the silent crowd significantly improves prediction if the number of prices included is low (less than or equal to five). Then its contribution falls and becomes negligible after the number of prices included exceeds eight.

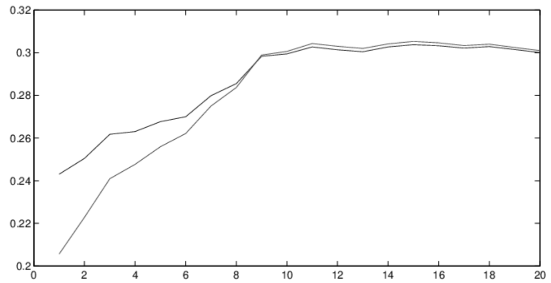

Figure 6: R squared for two sets f predictors

5 Conclusions

We have been able to reproduce the stylized facts of the LOB. Those include the hump-shaped density distribution of order sizes. Using a simulated LOB allows one to achieve desirable distributional properties while preserving unpredictability and to test various forecasting techniques in different scenarios. In our simulations, the midprice stabilizes, which is not a feature observed in practice. It can be easily avoided by introducing a random-walk-like disturbances. However, to obtain distributions of order sizes one would have to detrend the resulting midprice series using moving averages. Because of the lagging nature of moving averages, this would introduce an additional error in estimation. However, we believe that reversion to the mean would at least partially mitigate this problem and the final result would not be very different from ours.

References

[1] Parlour C., Seppi D.J. Limit Order Market: A Survey, Elsevier: North-Holland, 2008.

[2] Cont R. Statistical modeling of high-frequency financial data, Signal Processing Magazine, IEEE, 28 (2011), 16–25. https://doi.org/10.1016/b978-044451558-2.50007-6

[3] Cont R., Stoikov S., Talreja R. A stochastic model for order book dynamics, Operations Research, 58(2010), 549–563. https://doi.org/10.2139/ssrn.1273160

[4] He H., Kercheval A.N. A generalized birth-death stochastic model for high frequency order book dynamics, Quantitative Finance, 12 (2012), 547–557. https://doi.org/10.1080/14697688.2012.664926

[5] Rosu I. A dynamic model of the limit order book, Review of Financial Studies, 22 (2009), 4601–4641. https://doi.org/10.1093/rfs/hhp011

[6] Shek H.H.S. Modeling High Frequency Market Order Dynamics Using Self-Excited Point Process, http://dx.doi.org/10.2139/ssrn.1668160

[7] Bouchaud J.-P., Mezard M., Potters M. Statistical properties of stock order books: Empirical results and models, Quantitative Finance, 2 (2002), 251–256. https://doi.org/10.2139/ssrn.507362

[8] Foucault T., Kadan O., Kandel E. Limit order book as a market for liquidity, Review of Financial Studies, 18 (2005), 1171–1217. https://doi.org/10.1093/rfs/hhi029

[9] Jondeau E., Perilla A., Rockinger G. Optimal Liquidation Strategies in Illiquid Markets, 553 (2005), Springer: Berlin Heidelberg. https://doi.org/10.2139/ssrn.1431869

[10] Linnainmaa J.T., Rosu, I. Weather and Time Series Determinants of Liquidity in a Limit Order Market, AFA 2009 San Francisco Meetings Paper. http://dx.doi.org/10.2139/ssrn.1108862.

[11] Crammer K., Singer Y. On the algorithmic implementation of multiclass kernel-based vector machines, Journal of Machine Learning Research, 2 (2001), 265–292.

[12] Tino P., Nikolaev N., Yao X. Volatility forecasting with sparse bayesian kernel models, In 4th International Conference on Computational Intelligence in Economics and Finance 2005, 1052–1058.

[13] Blazejewski A., Coggins R. A Local Non-Parametric Model for Trade Sign Inference, Physica A: Statistical Mechanics and Its Applications, 348 (2005), 481–495. https://doi.org/10.1016/j.physa.2004.09.033

[14] Zheng B., Moulines E., Abergel, F. Price Jump Prediction in a Limit Order Book, Journal of Mathematical Finance, 3 (2013), 242–255. https://doi.org/10.4236/jmf.2013.32024

[15] Fletcher, T., Shawe-Taylor, J. Multiple Kernel Learning with Fisher Kernels for High Frequency Currency Prediction, Comput. Econ. 42(2013), 217–240. https://doi.org/10.1007/s10614-012-9317-z.

[16] Kercheval A.N., Zhang Y. Modelling high-frequency limit order book dynamics with support vector machines, Quantitative Finance, 15 (2015), 1315-1329, DOI: 10.1080/14697688.2015.1032546.

[17] Platania F., Serrano P., Tapia M. Modelling the shape of the limit order book, Quantitative Finance, 18 (2018), 1575–1597. https://doi.org/10.1080/14697688.2018.1433312

[18] Brogaard, J., Hendershott, T., Riordan, R. Price Discovery without Trading: Evidence from Limit Orders, The Journal of Finance, 74 (2019), 1621-1658. https://doi.org/10.1111/jofi.12769