Feature engineering of a Limit Order Book

XSOR Capital LTD

2021 December

1 Introduction

1.1 What is a Limit Order Book?

A Limit Order Book (LOB for short) is a data structure that is used by market participants to keep track of incoming orders and fill them correctly. Every exchange can use different rules for the implementation of the LOB, but the general structure is always the same.

There are two sides in a LOB, the *ask side* and the *bid side*. In general, the orders in the bid side have a lower price than the orders in the ask side. Orders can be stored based on their *price level*: for each price level, orders are inserted into a queue-like data structure, with their position in the queue depending on different parameters that can vary from exchange to exchange. Inside an order book, it is common to refer to these queues as simply "levels". You can see an example of an order book in Fig. 1.

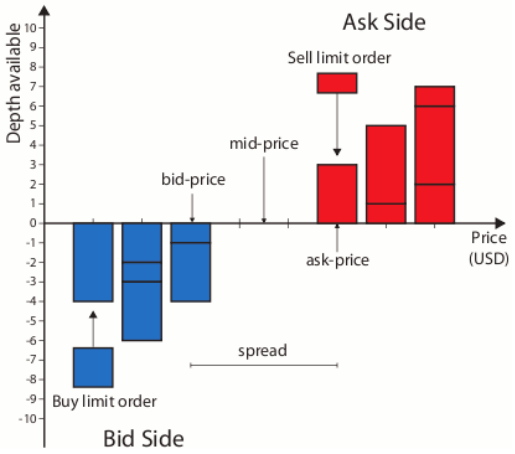

Figure 1: An example of an order book.

The most interesting portion of the LOB is the one around the *mid price*, defined as the average between the highest bid and the lowest ask price. The closer a price level is to the mid price, the earlier its orders are going to be filled if a match happens.

- for the **bid side**, the higher the price, the closer to the mid price. The highest bid price is also referred to as the *first bid level*;
- for the **ask side**, the lower the price, the closer to the mid price. The lowest ask price is also referred to as the *first ask level*.

You can easily convince yourself that this makes sense: bidders want to pay the possible lowest price in order to buy an instrument, and so the higher a bid price is, the more appealing the order is from the other side's point of view. Conversely, for the traders on the ask side, the aim is to sell at the highest price possible, and so an ask order at the lowest price is going to be the one that is most attractive for the bidders on the other side. Thus, the interesting region of the LOB is the one around the mid price.

Trades occur when orders of different sides *match their prices*. This happens because a side takes on the role of the *aggressor*: if, for instance, there is a resting bid order at the best bid price X, an ask order at that price X acts as an aggressor order, and it triggers the trade.

Orders in a LOB can be changed/updated while they are resting (for instance, a bid order of 100 units of security XXX at price 100, could be changed into a bid order of 90 units of that same security at the same price 100), and they can be removed altogether.

1.2 The project: features from a LOB

A LOB stores a great amount of information; depending on the liquidity of the instrument, it can be updated up to hundreds of thousands of times per trading session. It can also produce data about the trades that took place. The aim of the project is to extract some of this information as features that can then be used to better understand how the trading of the instrument under inquiry works. Here are some of the most interesting features you may want to explore.

1.2.1 Distribution of order sizes

An interesting quantity to study is the frequency of trades per trade size. First of all, trade frequency will likely be decreasing in trade size. Furthermore, previous studies have found that trades with round values (5, 10, 20, 50 and so on) are exceptionally common. This is due to the fact that human traders tend to rely on Graphical User Interfaces (GUIs) for their trading, and these platforms provide buttons that have preset round values. Silicon traders (i.e., trading algorithms) instead prefer to use randomized sizes, as to mask themselves better and not leave specific footprints of their activity.

If we divide the time series of trades into chunks, we can compute a frequency table of the trade sizes for that chunk. We can then define a sort of "average trade distribution" that, ideally, should be spiky at the round trade values while being decreasing in trade size. Alternatively, we can simply focus on the frequencies of the round values, and focus on those.

- Is the assumption on the trade distribution confirmed?
- Define a feature that describes the deviation from the "average trade distribution" for a specific time range.

You can refer to the book "Advances in Financial Machine Learning" by Lopez de Prado (2018), chapter 19.6.1.

1.2.2 Order flow imbalance

The LOB can be used to see how the shift in orders volumes and prices can give information about the future movement of prices. If we limit ourselves to the first level of the book (so to the highest bid level and the lowest ask level), we can informally define the order flow imbalance as the imbalance between demand and supply at the best bid and ask prices.

This quantity may be a good price predictor because the relation between the order flow imbalance and the price variation is roughly linear (see "The price impact of order book events", Cont et al., (2011)):

$$\Delta P_k = \beta \ OFI_k + \epsilon_k$$

where ΔP_k is the variation in price at time t_k, β is the price impact coefficient (estimated from the regression), OFI_k is the order flow imbalance at time t_k, and ϵ_k is the error term.

- Implement the order flow imbalance as described in the paper by Cont et al.;
- Use it to go further than level 1; then maybe look for correlations between various levels estimations.

1.2.3 Probability of Informed trading

Some microstructural models of market behavior aim at finding how likely it is that some players engage in informed trading, while the rest simply trade randomly. This information based trading is called Probability of Informed trading (PIN).

Consider a time interval in which some good or bad news can arrive about a security, impacting its price in a favorable or unfavorable way. An informed trader can take advantage of this piece of information. It can be shown that, under some assumptions, the probability of there being such a trader is

$$PIN_t = \frac{\alpha_t \mu}{\alpha_t \mu + 2\varepsilon},$$

where α_t, μ and ε are quantities that can be estimated by fitting the following mixture of three Poisson distributions[1]

$$\begin{aligned} P\left[V^B, V^S\right] = &(1-\alpha)\, P\left[V^B, \varepsilon\right] P\left[V^S, \varepsilon\right] + \\ &\alpha\left(\delta P\left[V^B, \varepsilon\right] P\left[V^S, \mu+\varepsilon\right] + (1-\delta)\, P\left[V^B, \mu+\varepsilon\right] P\left[V^S, \varepsilon\right]\right) \end{aligned}$$

where V^B is the number of buy-initiated trades in the time interval, while V^S is the number of sell-initiated trades in that same time interval.

- Implement the formula for PIN described above. More about the PIN formula and related ones can be found in the book "Advances in Financial Machine Learning" by Lopez de Prado (2018) at chapter 19.5.1.
- If you want, you can also try your hand at implementing the Volume-Synchronized PIN, described in the same book, in chapter 19.5.2. It may be even easier than PIN itself.

1.2.4 Additional research proposals

The following are ideas that you can play around with, but do not come with a specific request. You can unleash your creativity.

- There are definitions of the volatility of a LOB that are based around sampling the time series of the best ask and best bid prices. These methods simply take these time series and compute the standard deviation for the observations in the time series that fall inside a certain time interval. Thus a time series of volatilities can be produced. You can refer to "Limit Order Books", Gould et al. (2013).

  However, this method only takes the first levels into account. Can you think of a method that would use more than one level of the book?

- The LOB can be studied with power laws. Specifically, it has been conjectured that the distribution of the total number of orders inside one price level as a function of the relative price[a] may follow a power law.

  Try to create a feature that is based on fitting a power law to this price-volume profile on both sides of the book. For instance, the scale parameter of the two sides of the book may be combined into one feature.

[a] The bid (respectively, ask) relative price is defined as the absolute value of the difference between the bid (resp., ask) price and the best bid (resp., ask) price.

[1] Please note that the parameter δ, while not in the formula for PIN, has to be estimated nonetheless.