Order Book Analysis

Artagan Malsagov

January 18, 2024

1 Data Description

The data consist of 5 levels of both sides of the order book, for 5 different days. Each days spans roughly 10 hours worth of data (36 billion micros, see table below)

date	min timestamp	max timestamp	avg bbo mid	avg 5 level order volume
20190610	0	36000000000	10064	673
20190611	0	36000000000	10127	866
20190612	0	36000000000	9999	955
20190613	0	36000000000	10065	908
20190614	0	35999621354	9894	797

Table 1: Statistics per day of data

1.1 Data resampling

The order book is of the form below:

timestamp	bp0	bq0	ap0	aq0
110	10045	62	10055	98
175	10065	46	10075	42
220	10075	9	10080	25

Table 2: Subset of the data

where the timestamps are in microseconds (and there are 4 additional levels of the orderbook). Plotting a histogram of the frequency of the microsecond timestamps, we see that the updates aren't uniformly distributed:

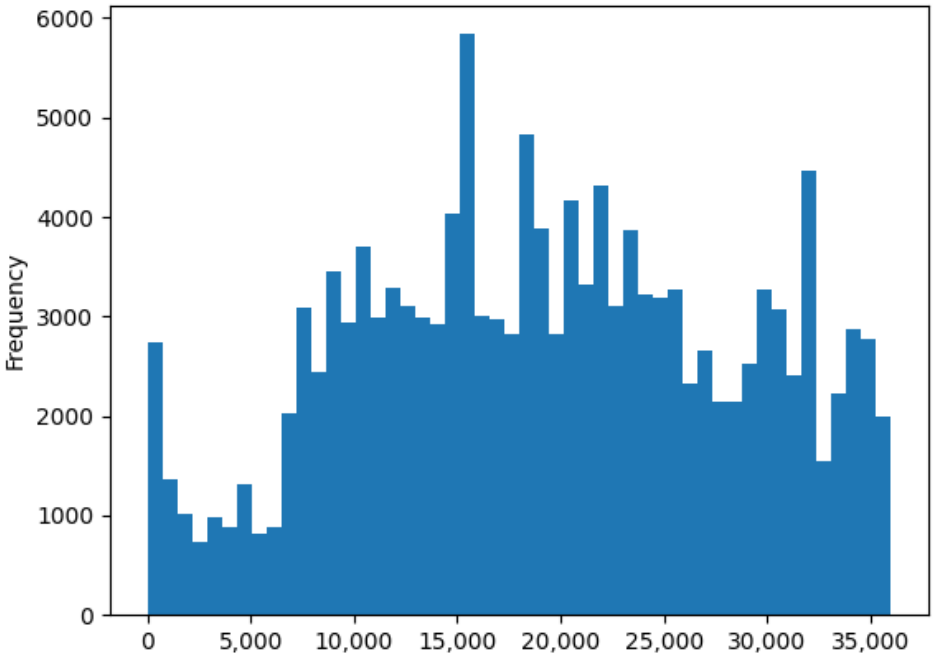

Figure 1: Histogram of order update arrivals for date 20190610

For the analysis the data will be resampled. Specifically, a grid will be used with a time interval of 100,000 micros = 100ms. This is done so as to reduce noise of the raw updates at the microsecond level and thus make signal detection in the data easier. Obviously, this choice of grid size might cause information loss and a more thorough analysis can be done to optimize, but for practical considerations the 100ms will be used as the discretization step, since this will allow for quicker data processing and model estimation. In practice this means the timestamps will be rounded up to the nearest 100,000 micros. The reason to round up is to avoid look-ahead bias when using the data as a trade signal, since rounding down will match an order update with a time-point in the past. In addition, when discretizing to a grid, there might be an issue when there are no updates available. In that case a forward interpolation is done by using the last known value of the previous discretization point.

2 Feature Selection

2.1 Target to predict

For the targets the following metrics were considered:

- The simple average mid price:

$$P_{mid} = \frac{bp0 + ap0}{2} \tag{1}$$

- The inverse volume weighted mid price:

$$P^1_{mid} = \frac{bp0 \times aq0 + ap0 \times bq0}{bq0 + aq0} \tag{2}$$

  This mid has the benefit of taking into account the order imbalance at the top level: if the buy order volume is higher, the price will be skewed higher to the ask, and vice-versa if the sell order volume is higher.

- The inverse volume weighted mid price at the first and second level:

$$P^2_{mid} = \frac{bp0 \times aq0 + ap0 \times bq0 + bp1 \times aq1 + ap1 \times bq1}{bq0 + aq0 + bq1 + aq1} \tag{3}$$

  This mid has the same advantage as the previous one, in addition it also takes into account the second layer of the orderbook and thus incorporates more information.

For the target the simple average mid (equation 1) is used rather than one of the two inverse weighted mids. The reasoning being that the inverse weighted mid prices are already predictors of sorts for the simple average mid.

Another interesting target would be the mid price change from time t to $t + \delta$ where the δ is set to a multiple of the discretization step 100ms. The logic being that price moves are what drives the markets, rather than the absolute price levels. However, to keep things simple, the target variable to be predicted will be the simple average mid price.

An exploratory analysis of the data is first appropriate. Below are the plots of the mid price and the inverse weighted mid price P^1_{mid}. Looking at the plots there are no weird outliers. Also clearly the inverse weighted mid price closely tracks the simple average mid price.

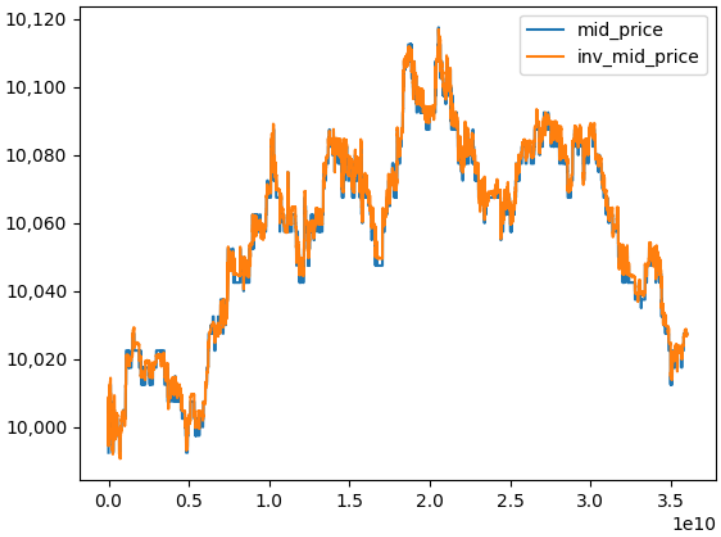

Figure 2: Day 2019-06-10

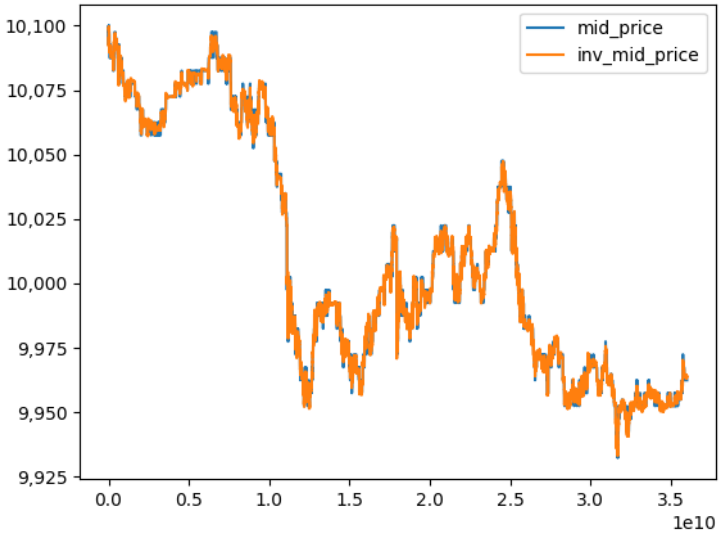

Figure 3: Day 2019-06-12

2.2 Features

The features below are considered for the analysis:

- The bid-offer spread calculated as:

$$BOspread = \frac{ap0 - bp0}{P_{mid}} \tag{4}$$

The intuition of using the spread as a predictor for the change of the mid-price is that if say the spread is relatively wide, then the probability of a non-marketable limit order arriving whose price is inside the spread is also higher. Whereas if the bid-offer spread was very narrow, then the mid-price can only change when one top level side of the order-book is depleted.

- Order imbalance at the top level:

$$OI_0 = \frac{bq0}{bq0 + aq0} \tag{5}$$

The intuition behind this metric is that if the order queue on the bid side is larger than on the ask side, the ask side will be depleted sooner and therefore this is predictive of an upward price move. Vice-versa a downward move if the ask side queue is larger. This ratio is chosen over a simple subtraction $bq0 - aq0$ because the ratio is normalized, which reduces bias in the model fitting. Note that if the top level bid size queue is larger than the one on the ask side, the ratio will be close to 1. And if the ask side has a much larger queue the ratio will be close to 0.

- Order imbalance at the second level: same as the above metric, but at the second level of the order book:

$$OI_1 = \frac{bq1}{bq1 + aq1} \tag{6}$$

- Change in $bq0$ relative to the previous time-point on the discretized grid:

$$\Delta bq0 = bq0_t - bq0_{t-\delta} \tag{7}$$

The intuition behind this is that the arrival of a large order on the bid side indicates more buying pressure and vice versa when a large buy order is removed by a marketable order. Both cases make it likely for the mid price to move.

- Change in $aq0$ relative to the previous time-point:

$$\Delta aq0 = aq0_t - aq0_{t-\delta} \tag{8}$$

The intuition behind this is the same as the one for the bid quantity change above.

- Finally the inverse mid price will also be considered as defined in equation (3). This price is sometimes used as a fair-price in market-making, and the analysis will show whether it has any predictive power. Most likely it will be heavily correlated with the mid-price.

3 Model Selection

The model that will be estimated using a Lasso regression is:

$$P_{mid,t+\delta} = BOspread_t + OI_{0,t} + OI_{1,t} + \Delta bq_{0,t} + \Delta aq_{0,t} + P^2_{mid,t} + \beta_0 + \epsilon \tag{9}$$

where ϵ is the error term and β_0 is the intercept. For ease of notation, the coefficients are omitted in the formula above. The forecasted future value of the mid price will be fixed at $\delta = 100ms$, so one discretization step forward on our grid. Obviously more values can be tested, but due to time constraints only this value will be used.

Note that the inverse weighted mid price of two layers $P^2_{mid,t}$ will be heavily correlated with P_{mid}. Hence, the regression will be run twice: with the inverse weighted mid price which will be model 1 and one time without the inverse weighted mid price, model 2:

$$P_{mid,t+\delta} = BOspread_t + OI_{0,t} + OI_{1,t} + \Delta bq_{0,t} + \Delta aq_{0,t} + \beta_0 + \epsilon \tag{10}$$

4 Results

The model above is estimated using a Lasso regression. Different alphas are experimented with and the tables report the R^2 score and MSE of both the test and train set. The results are given in the table below:

date	test score	test mse	train score	train mse
20190610	0.999716	0.233545	0.999710	0.238649
20190611	0.999766	0.129046	0.999767	0.128842
20190612	0.999935	0.141519	0.999936	0.139918
20190613	0.999639	0.161879	0.999640	0.162173
20190614	0.999900	0.225768	0.999900	0.224700

Table 3: Model with $\alpha = 0.04$

date	test score	test mse	train score	train mse
20190610	0.996999	2.470786	0.997010	2.464692
20190611	0.995976	2.217756	0.995977	2.224288
20190612	0.998965	2.251413	0.998966	2.255199
20190613	0.993737	2.806957	0.993749	2.817079
20190614	0.998459	3.482738	0.998463	3.468105

Table 4: Model with $\alpha = 1$

date	test score	test mse	train score	train mse
20190610	0.873826	103.867544	0.873907	103.942263
20190611	0.815287	101.806958	0.815320	102.095762
20190612	0.952993	102.279769	0.952984	102.536488
20190613	0.771241	102.520754	0.771078	103.173756
20190614	0.954335	103.184173	0.954352	103.014647

Table 5: Model with $\alpha = 10$

Note that as α is increased, the scores overall go down and the MSEs go up. The scores overall still remain relatively high though. This is due to the inclusion of $P^2_{mid,t}$, which is highly correlated with the simple mid price. Whether $P^2_{mid,t}$ is actually a predictor of the mid price cannot be concluded from this, since correlation does not mean causation.

The regression is done again with $P^2_{mid,t}$ omitted, to zoom in on the other variables. See the tables below. Here it becomes that the other variables don't predict the target variable very well, which can be observed by the higher MSEs and the lower scores. There might be many issues here, the choice of discretization step for the grid, the forecast horizon of $\delta = 100ms$, the form of the target variable: log difference of the mid could have been tested.

date	test score	test mse	train score	train mse
20190610	0.005131	818.981603	0.005719	819.613160
20190611	0.015000	542.894661	0.014752	544.669987
20190612	0.003620	2167.974286	0.003753	2172.710178
20190613	0.074478	414.783460	0.078335	415.389681
20190614	0.186116	1839.052605	0.187572	1833.434636

Table 6: Model with $\alpha = 0.04$

date	test score	test mse	train score	train mse
20190610	0.003133	820.626689	0.003457	821.477544
20190611	0.011403	544.877094	0.011313	546.571361
20190612	0.002394	2170.641543	0.002521	2175.395793
20190613	0.068101	417.641084	0.071400	418.515333
20190614	0.184384	1842.965925	0.185805	1837.423822

Table 7: Model with $\alpha = 1$

date	test score	test mse	train score	train mse
20190610	0.000000	823.205687	-0.000000	824.327841
20190611	0.000000	551.162010	-0.000001	552.825785
20190612	0.000000	2175.850513	-0.000002	2180.897872
20190613	0.000000	448.161418	-0.000012	450.700294
20190614	0.109777	2011.549451	0.110458	2007.460658

Table 8: Model with $\alpha = 10$

Interestingly, date 2019-06-14 has a higher score than the other dates, but also a higher mean square error. This suggests the model fits the data better for that day, but the predictions still have a high amount of error.

The sign of the coefficients is reported in the tables below for $\alpha = 0.04$ for several dates:

feature	coefficient
$BOspread$	0.237939
OI_0	2.080445
OI_1	-0.701317
Δbq_0	-0.092875
Δaq_0	0.117881

Table 9: $\alpha = 0.04$ and 2019-06-10

feature	coefficient
$BOspread$	-1.000401
OI_0	-0.785708
OI_1	-2.308759
Δbq_0	0.000000
Δaq_0	0.002927

Table 10: $\alpha = 0.04$ and 2019-06-12

feature	coefficient
$BOspread$	4.542120
OI_0	13.765068
OI_1	10.569343
Δbq_0	-0.987877
Δaq_0	0.441522

Table 11: $\alpha = 0.04$ and 2019-06-14

The expectation would be for all coefficients to be positive, since increasing values of the variables indicate buying pressure and decreasing value indicate selling pressure. However, several coefficients are negative, however with low absolute value. There was not enough time to include the statistical significance of these coefficients, however, it is likely that they are not statistically significant. Furthermore, on date 2019-06-14, the magnitude of two order imbalance variables (OI_0 and OI_1) is quite substantial. This is likely related with the test scores being very high for that date. Further investigation is needed to improve and test for better models.