Limit Order Signal Analysis (Tower Research)

Jing Guo*

October 12, 2016

1 Introduction

The paper is organized as follows: In Section 2 I define *Order Flow Imbalance* and *Trade Imbalance* and study them as price impact features. In Section 3 I analyze the lag factors and apply *ARMA* model to fit a time series model, and further combine the models introduced in Section 2 and 3 in Section 4. I study the cross-stock relationship between mid-price returns of stock A and stock B in Section 5. Arrival times of the next incoming trade is predicted by the *HMM* model in Section 6. Finally, section 7 concludes the findings of the report. Spread effect is studied in Appendix A.

2 Features: Order and Trade Imbalance

Define $(P_n^B, q_n^B, P_n^A, q_n^A)$ as the bid price, bid size, ask price and ask size of the n-th observation. I will define potential features of order and trade imbalance constructed by them, and study their impact on price movements.

2.1 Order Imbalance

I follow Cont et al. (2014) and define e_n which measures the contribution of the n-th event to the size of bid and ask queues:

$$e_n = I_{P_n^B \geq P_{n-1}^B} q_n^B - I_{P_n^B \leq P_{n-1}^B} q_{n-1}^B - I_{P_n^A \geq P_{n-1}^A} q_n^A - I_{P_n^A \leq P_{n-1}^A} q_{n-1}^A,$$

*Department of Industrial Engineering and Operations Research, Columbia University in the City of New York, New York, NY, 10027

and define *Order Flow Imbalance* (OFI) over time intervals $[t_{k-1}, t_k]$ as the sum of individual event contributions e_n over these intervals:

$$OFI_k = \sum_{n=N(t_{k-1})+1}^{N(t_k)} e_n,$$

where $N(t_{k-1})$ and $N(t_k)$ are the index of the first and the last event in the interval $[t_{k-1}, t_k]$.

2.2 Trade Imbalance

I follow Cont et al. (2014) and define the *Trade Imbalance* (TI) during a time interval $[t_{k-1}, t_k]$ as the difference between volumes of buy and sell trades during that interval:

$$TI_k = \sum_{n=N(t_{k-1})+1}^{N(t_k)} a_n - \sum_{n=N(t_{k-1})+1}^{N(t_k)} b_n,$$

where a_n and b_n are trade size on the ask side and bid side at the n-th quote.

2.3 Price Impact on Mid-Price Change

Define P_k as the mid-quote price at time t_k, i.e.,

$$P_k = \frac{P^A_{N(t_k)} + P^B_{N(t_k)}}{2},$$

and the mid-price changes over $[t_{k-1}, t_k]$ as

$$\Delta P_k = P_k - P_{k-1}.$$

I compute mid-price changes in every 10 seconds, and regress them over OFI's and TI's within every 30 minutes. There are total 13 sets of linear regression coefficients calculated, and the percentages of significant coefficients are reported as in Table 1. We can see from Table 1 that except in the first 30 minutes when irregular price jumps exist, OFI and TI features are always significant, i.e., they are strong price impact factors. R-squares of these linear models vary from 70% to 90%.

	Stock A	Stock B
OFI Significant Percentage	92.3%	100%
TI Significant Percentage	92.3%	100%
R-Square 90% Range	75.7%∼85.7%	74.7%∼83.7%

Table 1: Percentages of Significant OFI and TI Features. Mid-price changes are computed in every 10 seconds, and regressed over OFI's and TI's within every 30 minutes. We have total 13 sets of linear regression coefficients calculated, and the percentages of significant coefficients are reported as above. The range of R-squares using OFI and TI as prediction indicator is listed in the last row.

3 Time Dependency: Time Series Analysis

The time series of mid-price changes is sometimes mean-reverting or momentum-driven. I plot the mid-price changes of stock A in every 10 seconds within the first 30 minutes in Figure 1. We can see from the plot that mean-reverting effect may exist within the first 30 minutes for stock A.

Figure 2 shows the PACF and ACF of the mid-price changes of stock A. Based on the PACF and ACF plots, I choose $p = 2$ and $q = 1$ for my time series analysis. I run 13 auto-regressions for every 30 minutes, and list their significant percentages in Table 2. Within the periods of time when lag factors are significant, I compute the percentage of times of them being mean-reverting or momentum-driven as in the last two rows. From Table 2 we can see that lag factors are not always significant, and can vary from being mean-reverting to momentum-driven from time to time.

	Stock A	Stock B
ar1 Significant Percentage	69.2%	61.5%
ar2 Significant Percentage	7.7%	46.2%
ma1 Significant Percentage	76.9%	61.5%
Significant Mean-Reverting	33.7%	23.0%
Significant Momentum-Driven	38.5%	38.5%

Table 2: Percentages of Significant ar1, ar2 and ma1. Mid-price changes are computed in every 10 seconds, and regressed over lag-1 and lag-2 observations within every 30 minutes. We have 13 sets of linear regression coefficients calculated, and the percentages of significant coefficients are reported above. Within the periods of time when lag factors are significant, I compute the percentage of it being mean-reverting or momentum-driven as in the last two rows.

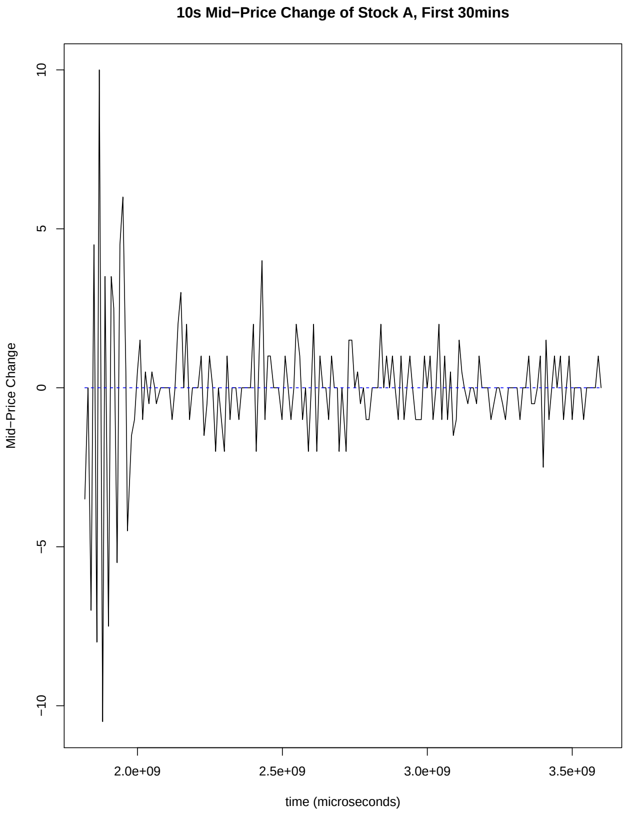

Figure 1: Time Series of mid-price changes of Stock A. Mid-price changes are computed in every 10 seconds, and plotted within the first 30 minutes.

4 ARMAX: Features + Auto-Regression

In this section I estimate ARMA model with external factors OFI and TI (ARMAX model). Here I use $p = 2$ and $q = 1$, estimate ARMAX model in every 30 minutes, and list the percentage of times when ar1, ar2, ma1, OFI and TI are significant as in Table 3. Here obviously *Order Flow Imbalance* and *Trade Imbalance* are still significant and thus strong factors, while lag factors are not always significant.

5 Cross-Asset Relationship

I plot the mid-price returns of stock A and stock B in the first 30 minutes as in Figure 3. Mid-price returns are computed in every 10 seconds. While there are irregular jumps of price in the first 3 minutes, potential positive relationship can be observed in the first

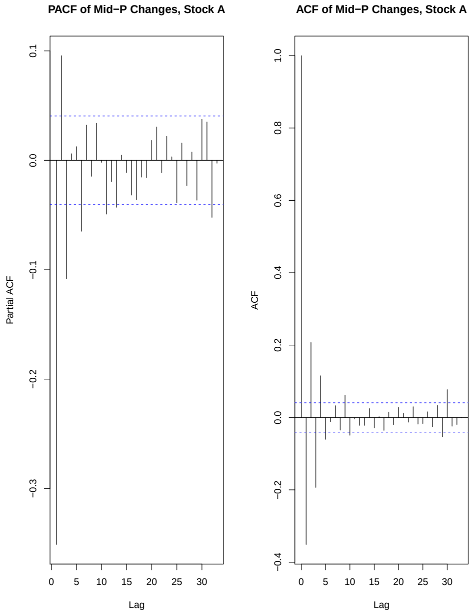

Figure 2: PACF and ACF Plots of mid-price changes of Stock A. Mid-price changes are computed in every 10 seconds. Based on the PACF and ACF plot, I choose $p = 2$ and $q = 1$ for my time series analysis.

	Stock A	Stock B
ar1 Significant Percentage	30.8%	38.5%
ar2 Significant Percentage	7.7%	23.1%
ma1 Significant Percentage	38.5%	23.1%
OFI Significant Percentage	92.3%	100%
TI Significant Percentage	92.3%	100%

Table 3: Percentages of Significant ar1, ar2, ma1, OFI and TI Features. Mid-price changes are computed in every 10 seconds, and regressed over lag-1, lag-2 observations, OFI's and TI's within every 30 minutes. We have total 13 sets of linear regression coefficients calculated, and the percentages of significant coefficients are reported above.

30 minutes between stock A and stock B. Here I apply *Vector Auto-Regression* (VAR) and study the relationship between returns of stock A and stock B. Define return vector $Y_t = (r_A(t), r_B(t))$, where $r_A(t)$ and $r_B(t)$ are mid-price returns of stock A and stock B at time t. Here I choose lag order $p = 1$. Then VAR model is of the following form:

$$Y_t = A_1 Y_{t-1} + \varepsilon_t, \tag{1}$$

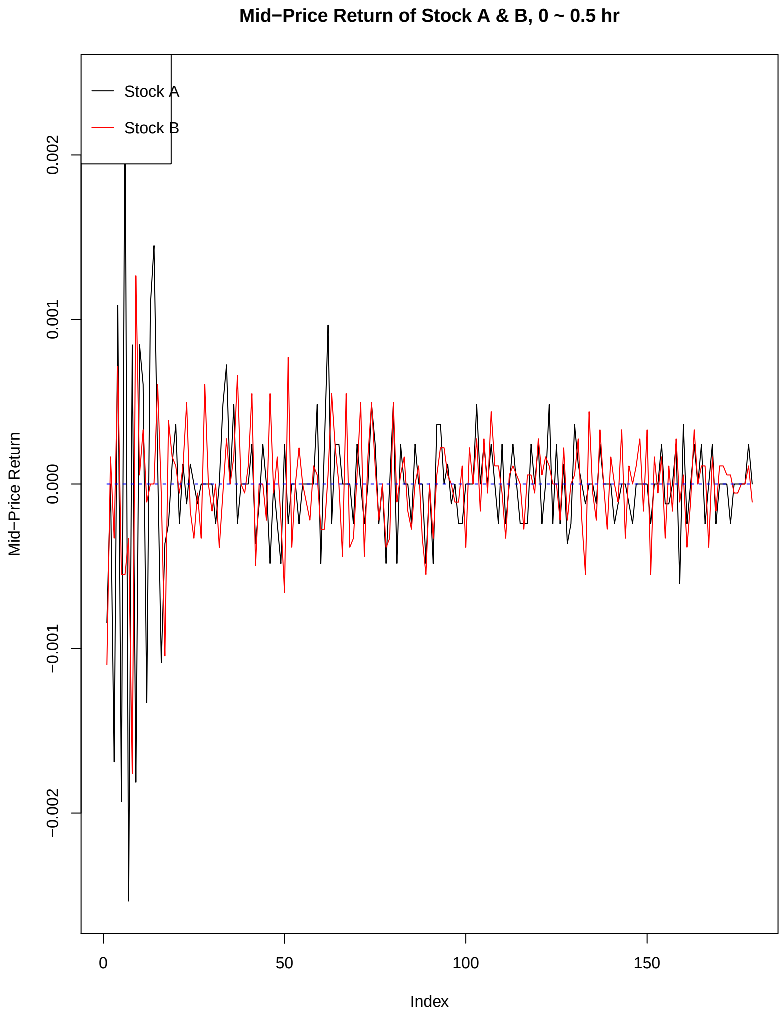

Figure 3: Mid-Price Returns of Stock A and Stock B in the First 30 Minutes. Mid-price returns are computed in every 10 seconds. While there are irregular jumps of price in the first 3 minutes, potential positive relationship can be observed in the first 30 minutes between stock A and stock B.

where $A_1 \in \mathbb{R}^2$ is the coefficient matrix to fit. I estimate model (1) in every 30 minutes with returns calculated in every 10 seconds, and find that the returns of stock A and stock B only display significant linear relationship in the first 30 minutes. I also compute the correlation between stock A and stock B returns in every 30 minutes. The correlations vary from .100 to .412. The correlation is largest around the time when price changes are small.

6 Trade Arrival Intensity

In this section I use Hidden Markov Model (HMM) to predict how long it takes for the next trade to come. Define the arrival time of $n-$th incoming trade as

$$y_n = t_n - t_{n-1},$$

where t_n and t_{n-1} are the arrival time of the n-th and $(n+1)$-th trades. Assume that y_n follows distribution:

$$y_n \sim \exp(\lambda(x_n)),$$

where $\lambda(x_n) \in \mathbb{S} = \{\lambda_1, ..., \lambda_K\}$ is a hidden Markov Chain, and time-dependent state process x_n transitions with transition matrix $\mathbb{P}$. I use the first 6 hours as training data set, and apply Forward-Backward algorithm to fit the HMM model with $K = 5$. The arrival times in the last 30 minutes, given all the information up to that them, are predicted using the HMM model.

Predicted arrival times within last 30 minutes are shown in Figure 4. We can see that the real arrival data points almost always fall between the 90% confidence bands.

7 Conclusions

- Order and trade prices are not stable in the first 3 minutes.
- *Order Flow Imbalance* and *Trade Imbalance* are two strong factors.
- Lag (time series) factors are not strong factors.
- Across-stock correlation is generally weak, while it displays positive relationship in the first 30 minutes.
- HMM model can predict trade arrival times well.
- Spread effect is not significant.

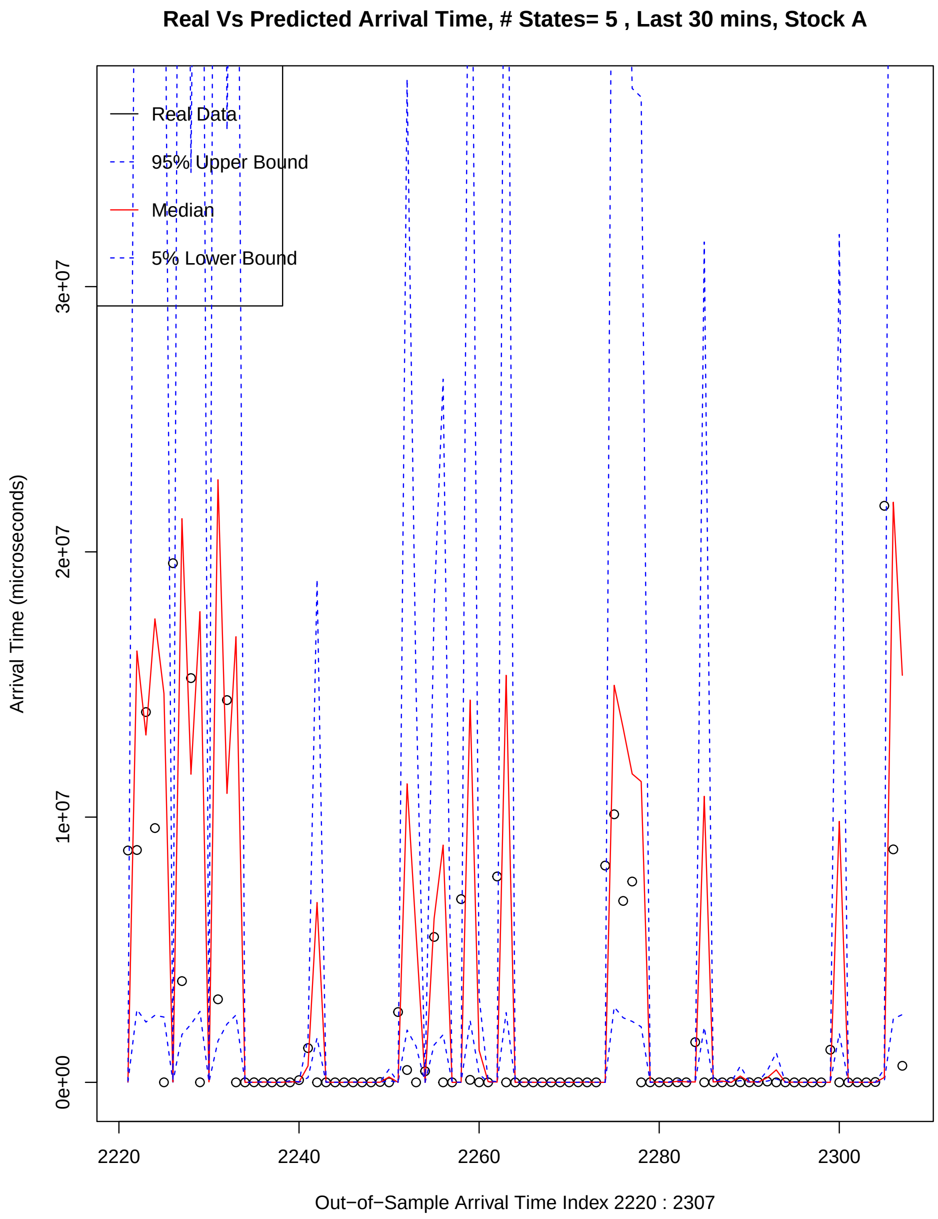

Figure 4: Real versus Predicted Arrival Times. Training data set is the arrival times in the first 6 hours, and Forward-Backward algorithm is applied to fit the HMM model. The arrival times in the last 30 minutes are predicted using HMM model. X axis is the number of arriving trades in the last 30 minutes, while Y axis is the arrival times measured in microseconds. The black dots are real arrival times, the red line is the predicted medians by the fitted HMM model, and the blue lines are the upper and lower 5% quantile boundary. Almost all real arrival times fall between the 90% confidence bands.

A Spread Effect

Shen (2015) argued that mid-price changes are affected by spreads, and one should scale mid-price change by its spread in order to achieve homogeneous variance. I do the relevant research but do not see this effect. See Figure 5 for reference.

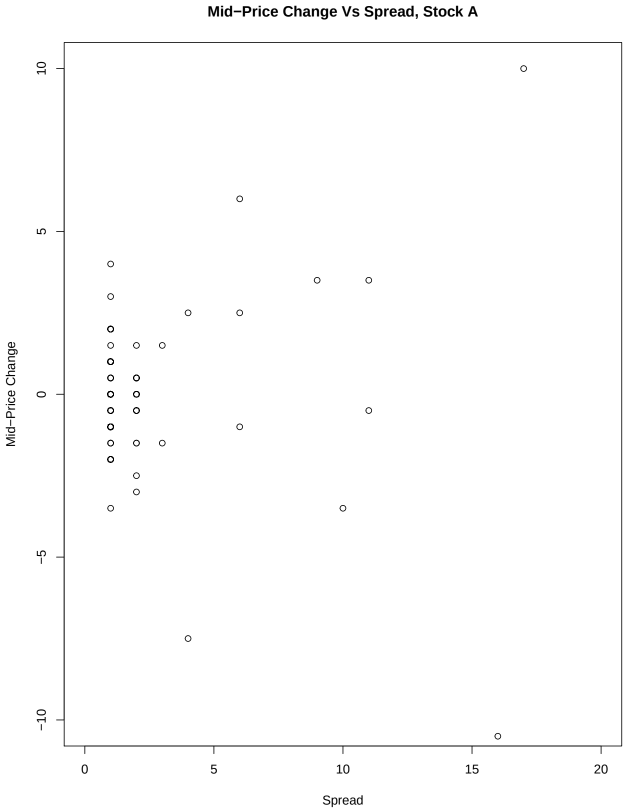

Figure 5: Spread Effect

References

Cont, R., Kukanov, A. and Stoikov, S. (2014). The price impact of order book events, *Journal of Financial Econometrics* **12**: 47–88.

Shen, D. (2015). Order imbalance based strategy high frequency trading, *Unpublished manuscript* .