Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 187 (2021) 518–523

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Identification, Information and Knowledge in the Internet of Things, 2020

High-frequency Statistical Arbitrage Strategy Based on Stationarized Order Flow Imbalance

Qingxue Wang[a,c], Bin Teng[a,c], Qi Hao[a,c], Yufeng Shi[a,b,c,*]

[a]*Institute for Financial Studies, Shandong University, Jinan 250100, China*
[b]*School of Mathematics, Shandong University, Jinan 250100, China*
[c]*Shandong Big Data Research Association, Jinan 250100, China*

Abstract

From a high-frequency perspective, the order book data records each market participant's price expectations for the current underlying assets, and therefore it records the fundamental mechanism of pricing. In this paper, we propose a stationarized indicator based on the classical indicator – Order Flow Imbalance (OFI), and empirically find that there is a stronger linear relationship between the new indicator and the mid-price. Furthermore, we develop a statistical arbitrage strategy based on this new indicator. Compared with the classic indicator, the strategy built with the new indicator has better performance.

Keywords: High-frequency trading (HFT); Limit order book (LOB); Order flow imbalance (OFI); Stationary process; Statistical arbitrage

1. Introduction

In recent years, there has been a growing number of high-quality, high-frequency datasets on high-frequency trading (HFT). In particular, the limit order book (LOB), a system of limit order status at different prices, which contains the most microscopic trading information provided by market participants. Smith et al. [13] propose a model that assumes both sequences of order arrivals and cancellations as independent Poisson processes to simulate the effects of order volume changes on prices. Cont et al. [3] extend the model by assuming that the arrival and cancellation rates of limit orders vary with price. Huang et al. [8] propose a model in which order arrival rates depend on order book states, treating the order books as a Markov queuing system, and point out that order cancellations are important factors influencing the price distribution. In Mike and Farmer [11], it is shown that the mid-price of the order flow exhibits evidence of long-term autocorrelation. Eisler et al. [4] decompose the impact of an event into three parts: an

* Corresponding author.
E-mail address: yfshi@sdu.edu.cn

1877-0509
10.1016/j.procs.2021.04.093

instantaneous jump component, the modification of the future rates of the different events, and the modification of the jump sizes of future events.

The indicator order flow imbalance (OFI) is proposed by Cont et al. [1]. When studying the relationship between trade imbalance and the change of the mid-price price over the same period, the authors find that the trade imbalance only describes the completed orders and ignores the impact of the arrival and cancellation of limit orders on the mid-price.

In this paper, we propose a stationarized form of OFI indicator, denoted as log-OFI, based on the classic OFI indicator. In Sect. 2, we first describe the OFI indicator, and observe the frequency distribution of mid-price changes, ask/bid price changes. It can be found that the two variables, mid-price and order size, differ significantly in the magnitude of instantaneous changes, which may weaken the linear correlation. Therefore, we consider the logarithm of the order volume, which is equivalent to using the relative change of the order volume rather than the absolute change, thus proposing a new OFI indicator. Next, we discuss the stationarity of the new indicator, and its linear relationship with price movements. A statistical arbitrage strategy is designed in Sect. 3 based on the stationarity of log-OFI, and compared it with that of the classic OFI indicator.

2. Stationarized order flow imbalance

2.1. Limit order book (LOB) and order flow imbalance (OFI)

The limit order book (LOB) consists of a timestamp, the latest transaction price, and the ask/bid price and volume. LOB records the status of limit orders at different prices. The status of the order book at a given moment contains ask/bid prices and order sizes. The order book is divided into different levels, and the information presented in each level reflects the traders' expectations for the future transaction prices. The order book is an important tool for analyzing the behaviors and expectations of market participants, since the status of order book is constantly updated with new information, and each order execution leaves a trace in the order book.

The order flow imbalance (OFI), proposed in Cont et al. [1], is a quantification of the supply/demand status between buyers and sellers in the market. Let τ_n be the moment when the nth order arrives or cancels, $b(\tau_n)$, $a(\tau_n)$, $r(\tau_n)$, and $q(\tau_n)$ as the bid price, ask price, bid volume, and ask volume at time τ_n, respectively. Between the occurrence of two consecutive order book events, i.e., from τ_{n-1} to τ_n, the OFI indicator is defined as follows:

$$\Delta W(\tau_n) = \begin{cases} r(\tau_n), & \text{if } b(\tau_n) > b(\tau_{n-1}). \\ r(\tau_n) - r(\tau_{n-1}), & \text{if } b(\tau_n) = b(\tau_{n-1}). \\ -r(\tau_{n-1}), & \text{if } b(\tau_n) < b(\tau_{n-1}). \end{cases} \tag{1}$$

$$\Delta V(\tau_n) = \begin{cases} -q(\tau_{n-1}), & \text{if } a(\tau_n) > a(\tau_{n-1}). \\ q(\tau_n) - q(\tau_{n-1}), & \text{if } a(\tau_n) = a(\tau_{n-1}). \\ q(\tau_n), & \text{if } a(\tau_n) < a(\tau_{n-1}). \end{cases} \tag{2}$$

$$\mathrm{OFI}^{\mathrm{Tick}}(\tau_n) = \Delta W(\tau_n) - \Delta V(\tau_n). \tag{3}$$

The OFI for a given time interval, e.g. $(t_{k-1}, t_k]$, is represented by the summation of all the OFIs generated by order book events during that interval:

$$\mathrm{OFI}(t_{k-1}, t_k] = \sum_{\tau_n \in (t_{k-1}, t_k]} \mathrm{OFI}^{\mathrm{Tick}}(\tau_n). \tag{4}$$

OFI measures the direction and size of the order flow between the ask price and the bid price within $(t_{k-1}, t_k]$. When OFI is positive, it means that the total amount of new orders in the buying position is larger, or the total amount of orders cancelled in the selling position is larger. At this time, the pressure of the buyer is greater than the pressure of the seller, resulting in an increase in the mid-price. On the contrary, when the OFI is negative, it means a larger total amount of new orders to sell the position, or a larger total amount of cancellations to buy, and then the pressure of the seller is greater than the pressure of the buyer, which leads to a decline in mid-price. Cont et al. [1] believe that the mechanism of LOB trading shows that the greater the value of OFI, the greater the probability of mid-price increase, and vice versa.

Cont et al. [1] have further investigated the linear correlation between OFIs and mid-price movements, that is, the least square regression (OLS):

$$\Delta P(t_{i,k-1}, t_{i,k}] = \alpha + \beta OFI(t_{i,k-1}, t_{i,k}] + \epsilon, \tag{5}$$

where $\Delta P(t_{i,k-1}, t_{i,k}] := p(t_{i,k}) - p(t_{i,k-1})$, and the mid-price $p(\tau_n) := \frac{r(\tau_n)a(\tau_n)+q(\tau_n)b(\tau_n)}{r(\tau_n)+q(\tau_n)}$. They conclude that the average R^2 of the regression is about 65%. This means that OFI is a valid explanatory indicator of price movements.

2.2. Towards stationarized order flow imbalance

For the empirical analysis in this paper, we use data of the coke futures contract of Dalian Commodity Exchange (DCE), China, dated October 2019. We first calculate all the mid-price changes and order volume changes in the order book, and their frequency distributions are shown in Fig. 1. The volume and price changes all show significant peaks. The range of mid-price changes is very narrow, with no long tails. The order volume changes shows the characteristics of long-tail ranged in $[-400, 400]$.

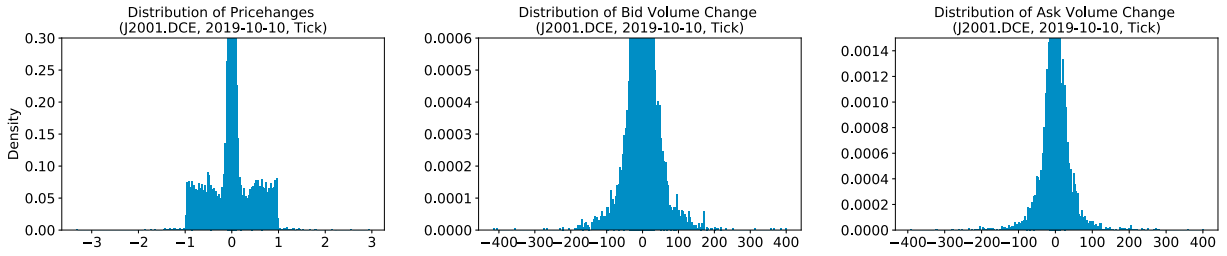


Fig. 1. Distribution of mid-price changes & Order volume changes (tick). (a) Mid-price changes; (b) Bid-volume changes; (c) Ask-volume changes.

We consider logarithmic difference. Logarithmic difference is a simple and effective method for stationarity. We take the logarithm of the ask/bid volume and then make the difference, whose frequency distributions are shown in Fig. 2. The logarithmic difference order volume retains the sharp peak characteristic, but is no longer sensitive to unusually large changes and the long tail almost disappears, with a range around $[-6, 6]$.

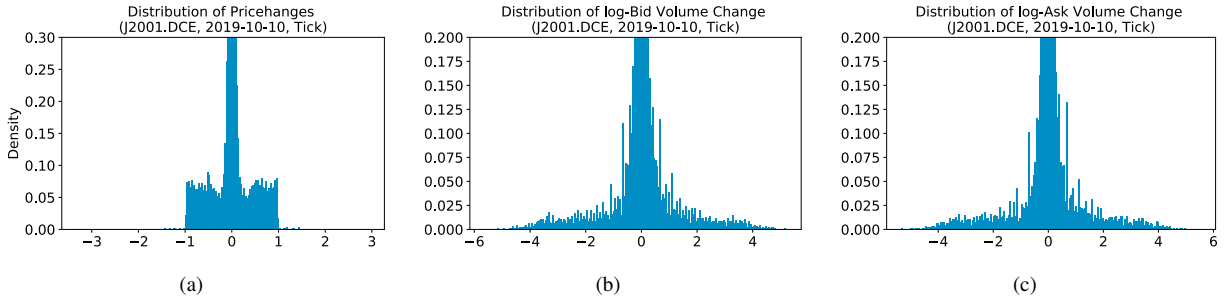


Fig. 2. Distribution of mid-price changes & log-volume changes (tick). (a) Mid-price changes; (b) Bid-volume changes; (c) Ask-volume changes.

Therefore, we construct a new indicator referred to as log-OFI, which is specific defined as follows:

$$\Delta\tilde{W}(\tau_n) = \begin{cases} \log r(\tau_n), & \text{if } b(\tau_n) > b(\tau_{n-1}). \\ \log r(\tau_n) - \log r(\tau_{n-1}), & \text{if } b(\tau_n) = b(\tau_{n-1}). \\ -\log r(\tau_{n-1}), & \text{if } b(\tau_n) < b(\tau_{n-1}). \end{cases} \tag{6}$$

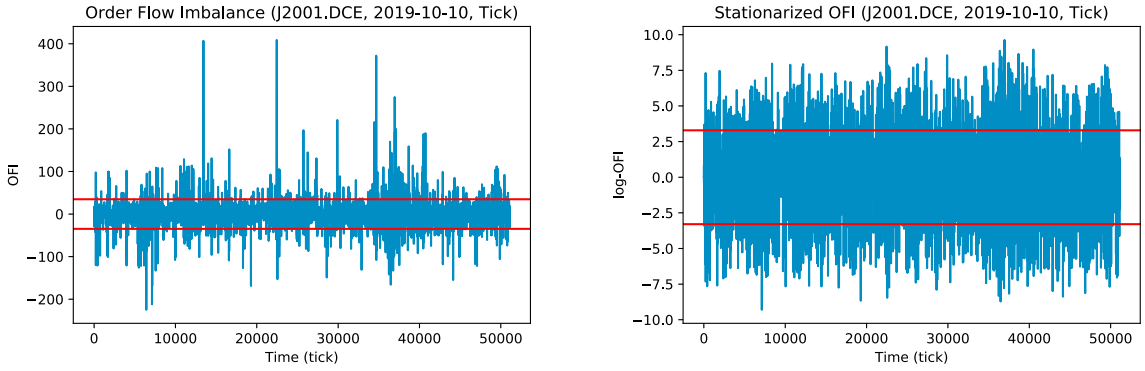
$$\Delta\tilde{V}(\tau_n) = \begin{cases} -\log q(\tau_{n-1}), & \text{if } a(\tau_n) > a(\tau_{n-1}). \\ \log q(\tau_n) - \log q(\tau_{n-1}), & \text{if } a(\tau_n) = a(\tau_{n-1}). \\ \log q(\tau_n), & \text{if } a(\tau_n) < a(\tau_{n-1}). \end{cases} \tag{7}$$

$$\text{log-OFI}^{\text{Tick}}(\tau_n) = \Delta\tilde{W}(\tau_n) - \Delta\tilde{V}(\tau_n), \tag{8}$$

$$\text{log-OFI}(t_{k-1}, t_k] = \sum_{\tau_n \in (t_{k-1}, t_k]} \text{log-OFI}^{\text{Tick}}(\tau_n). \tag{9}$$

Next, we will verify that this new indicator has better stationarity than OFI.

Time series is a sequence of random variables arranged in time order. A stationary process means that its statistical properties do not vary with time. See more details about the theory of stationary process and statistical arbitrage in [14]. we examine the stationarity of OFIs and log-OFIs. From Fig. 3, we can see that log-OFIs have fewer outliers than OFIs. This illustrates that we have constructed a stationarized indicator.



(a) (b)

Fig. 3. Order Flow Imbalance (OFI) sequences (Tick). The horizontal line represents three times the standard deviation. (a) OFIs; (b) Stationaried log-OFIs.

2.3. R^2 of the linear regression

Consider the linear regression (5). We use R^2 to evaluate the effectiveness of the linear fit between OFIs and mid-price changes. We use the same definition as [5]:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} {y_i}^2}, \tag{10}$$

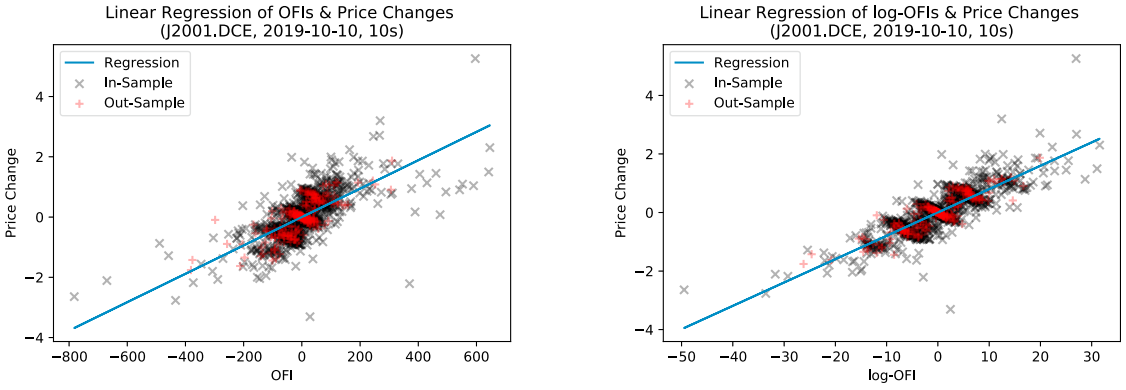
where y_i is true value and $\hat{y}_i$ is predicted value. When the R^2 is closer to 1, the estimated value of the regression is closer to the true value.

We use 1500 data points within each trading day as in-sample data and 300 data points as out-of-sample. Then we compare the R^2 of the in-sample and out-of-sample data, denoted as R^2_{IS} and R^2_{OOS}, respectively.

Fig. 4 shows that the log-OFIs we defined not only reduce the outliers, but also reduce the bias of the regression. The regression line of OFI is skewed, in fact, this is because the majority of data points are concentrated at the origin point, and the definition of OFI at the origin may be inaccurate. Table 1 shows the R^2 of the coke contract (J2001.DCE) in October 2019. We find that the R^2 of log-OFIs is significantly larger than that of OFIs, both in-sample and out-of-sample. It is also noteworthy that R^2_{IS} is not significantly higher than R^2_{OOS}, which indicates that the linear correlation is stable.

3. Statistical arbitrage via stationary log-OFIs

In this paper, we take a statistical arbitrage approach in our strategy. The main idea of the strategy is that log-OFI indicator and price changes maintain a high linear relationship in both in-sample and out-of-sample dataset. We suggest that the sufficiently stationary log-OFIs are filters for price changes. Once the log-OFI deviates too much, it



(a) (b)

Fig. 4. Linear regression of OFIs & Price changes (10s). (a) Unstationary OFIs; (b) Stationarized OFIs.

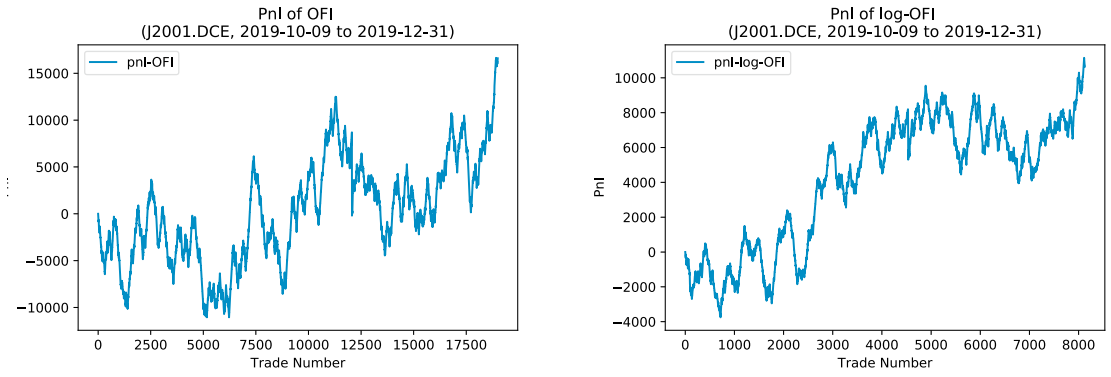
Table 1. R^2 of linear regression between OFIs and price changes.

Date	OFI (R^2_{IS})	OFI (R^2_{OOS})	log-OFI (R^2_{IS})	log-OFI (R^2_{OOS})
2019-10-08	0.482	0.548	0.677	0.684
2019-10-09	0.542	0.548	0.775	0.772
2019-10-10	0.454	0.546	0.699	0.774
2019-10-11	0.451	0.432	0.726	0.711
2019-10-14	0.535	0.766	0.741	0.801
2019-10-15	0.459	0.502	0.714	0.803
2019-10-16	0.513	0.560	0.761	0.794
2019-10-17	0.503	0.538	0.751	0.734
2019-10-18	0.542	0.627	0.776	0.834
2019-10-21	0.509	0.574	0.738	0.804
2019-10-22	0.443	0.456	0.717	0.764
2019-10-23	0.392	0.405	0.693	0.648
2019-10-24	0.388	0.473	0.686	0.737
2019-10-25	0.465	0.246	0.721	0.801
2019-10-28	0.472	0.489	0.762	0.722
2019-10-29	0.400	0.414	0.731	0.666
2019-10-30	0.518	0.440	0.748	0.677
2019-10-31	0.461	0.458	0.745	0.714
Total	**0.474**	**0.501**	**0.731**	**0.747**

indicates a drastic price movement, while it is necessarily mean-reverting due to its stationarity. Take advantage of this idea, we can design high-frequency trading strategies. For the theory of high-frequency trading, please refer to Wang & Zheng [14].

We set the upper and lower bound of opening positions as 24 and -24, respectively. When the log-OFI crosses over (under) the upper (lower) bound, we sell short (buy) the contracts. We also set the upper and lower bound of closing positions as 12 and -12, respectively. If we hold a long (short) position, and the log-OFI crosses over (under) the upper (lower) bound, we sell (buy to cover) the contracts. Similarly, we apply the same strategy to the indicator OFIs, and the corresponding thresholds are 300 and 150, respectively.

Fig. 5 indicates that the P&L curve of log-OFI strategy has a smaller maximum drawdown (MDD) than that of the classic indicator OFI, and the cumulative expectation is positive. Moreover, under the same parameters, the new strategy has fewer times of opening positions but larger average return of each opening position. The above results show that the new strategy improves the original performance.



(a) (b)

Fig. 5. Profits and Losses (P&L), J2001.DCE, 2019-10-09 to 2019-12-31. (a) OFI ; (b) log-OFI.

4. Conclusions

The order book has always attracted the attention of scholars and investors. In this paper, we propose a stationarized log-OFI indicator based on the OFI indicator, by observing the characteristics of high-frequency data. It is found that log-OFIs are relatively stationary, and the R^2 of their linear regression with price changes is greatly improved. We also use data from the J2001.DCE contract to validate a statistical arbitrage strategy for the log-OFI indicator. The results show that the log-OFI strategy has more stable returns than classical OFI.

Acknowledgements

This work is supported by the National Key R&D Program of China (Grant No. 2018YFA0703900), and the National Natural Science Foundation of China (Grant Nos. 11871309 and 11371226).

References

[1] Cont, R., Kukanov, A., Stoikov, S., 2014. The price impact of order book events. Journal of Financial Econometrics 12, 47–88.
[2] Cont, R., de Larrard, A., 2013. Price dynamics in a markovian limit order market. SIAM Journal on Financial Mathematics 4, 1–25.
[3] Cont, R., Stoikov, S., Talreja, R., 2010. A stochastic model for order book dynamics. Operation Research 58, 1–21.
[4] Eisler, Z., Bouchaud, J.P., Kockelkoren, J., 2012. The price impact of order book events: market orders, limit orders and cancellations. Quantitative Finance 12, 1395–1419.
[5] Gu, S., Kelly, B., Xiu, D., 2020. Empirical asset pricing via machine learning. The Review of Financial Studies 33, 2223–2273.
[6] Hogan, S., Jarrow, R., Teo, M., Warachka, M., 2004. Testing market efficiency using statistical arbitrage with applications to momentum and value strategies. Journal of Financial Economics 73, 525–565.
[7] Huang, H.C., Su, Y.C., Liu, Y.C., 2014. The performance of imbalance-base trading strategy on tender offer announcement day. Investment Management and Financial Innovations 11, 38–46.
[8] Huang, W., Lehalle, C.A., Rosenbaum, M., 2015. Simulating and analyzing order book data: The queue-reactive model. Journal of the American Statistical Association 110, 107–122.
[9] Lillo, F., Farmer, J.D., 2004. The long memory of the efficient market. Studies in Nonlinear Dynamics & Econometrics 8, 1–19.
[10] Mike, S., J. Doyne, F., 2008. An empirical behavioral model of liquidity and volatility. Journal of Economic Dynamics & Control 32, 200–234.
[11] Potters, M., Bouchaud, J.P., 2003. More statistical properties of order books and price impact. Physica A: Statistical Mechanics and its Applications 324, 133–140.
[12] Shen, D., 2015. Order imbalance based strategy in high frequency trading. Ph.D. thesis. Oxford University.
[13] Smith, E., Farmer, J.D., Gillemot, L., Krishnamurthy, S., 2003. Statistical theory of the continuous double auction. Quantitative Finance 3, 1–36.
[14] Wang, Z., Zheng, W., 2015. High-Frequency Trading and Probability Theory. World Scientific.