Work Trial Task: Cross-Impact Analysis of Order Flow Imbalance (OFI)

Yuehan He

January 8, 2025

Abstract

This study explores the relationship between Order Flow Imbalance (OFI) metrics and price changes in equity markets, focusing on both contemporaneous and predictive impacts. Utilizing high-frequency market data, we compute multi-level OFI metrics across five stocks and integrate them into a single metric using Principal Component Analysis (PCA). This integrated OFI captures the dynamics across multiple levels of the order book, reducing dimensionality while preserving key features.

To assess the explanatory and predictive power of OFI, we implemented regression models inspired by the "Cross-Impact of Order Flow Imbalance in Equity Markets" framework. The Price Impact Model with Best-Level OFI (PI[1]) evaluates the contemporaneous relationship between the best-level OFI and log returns, while the Price Impact Model with Integrated OFI (PII) leverages the PCA-integrated metric for similar analysis. Cross-Impact Models (CI[1] and CII) extend this approach by incorporating cross-asset OFI metrics to examine interactions between stocks.

Despite rigorous computation and analysis, our findings indicate weak linear relationships between OFI metrics and price changes. The correlation analyses suggest that the explanatory power of OFI, whether contemporaneous or lagged, may be limited due to data quality issues, linear modeling assumptions, or the inherently nonlinear nature of the OFI-return relationship. These results align with the challenges highlighted in the referenced study, emphasizing the need for nonlinear approaches and the inclusion of broader market factors in future analyses.

This report contributes to the understanding of market microstructure by offering a comprehensive methodology for OFI computation and regression analysis while underscoring the limitations of current models in capturing the complexities of equity price dynamics.

1 Methodology

1.1 Order Flow Imbalance (OFI) Calculation

Order Flow Imbalance (OFI) is a critical metric that quantifies the imbalance between bid and ask volumes at different levels of the order book. This study computes OFI for up to five levels of market depth for each stock in the dataset. The methodology follows the steps outlined below:

1.1.1 Bid and Ask Size Changes

At each level m of the order book, the changes in bid and ask sizes are computed as:

$$\Delta\text{Bid Size}_m = \text{Bid Size}_m(t) - \text{Bid Size}_m(t-1),$$

$$\Delta\text{Ask Size}_m = \text{Ask Size}_m(t) - \text{Ask Size}_m(t-1),$$

where $\text{Bid Size}_m(t)$ and $\text{Ask Size}_m(t)$ denote the bid and ask sizes at level m at time t.

1.1.2 Level-Specific OFI

Using the changes in bid and ask sizes, the OFI for level m is computed as:

$$\text{OFI}_m(t) = \Delta\text{Bid Size}_m(t) - \Delta\text{Ask Size}_m(t).$$

1.1.3 Multi-Level OFI

The above calculations are repeated for up to five levels of the order book ($m = 1, 2, \ldots, 5$). The resulting metrics are stored as separate features ($\text{OFI}_1, \text{OFI}_2, \ldots, \text{OFI}_5$).

1.2 Integration of Multi-Level OFI Using PCA

To reduce dimensionality and integrate the information from all five levels, we apply Principal Component Analysis (PCA). The steps are as follows:

1. **Normalization:** Each OFI level is normalized to zero mean and unit variance:

$$\text{Normalized OFI}_m = \frac{\text{OFI}_m - \mu_m}{\sigma_m},$$

   where μ_m and σ_m are the mean and standard deviation of OFI_m.

2. **Principal Component Extraction:** PCA is applied to the normalized multi-level OFI data to extract principal components:

$$\text{Integrated OFI} = \text{First Principal Component}.$$

   The first principal component is selected as it explains the largest variance in the data.

1.3 Logarithmic Price Returns

The dependent variable in the regression models is the logarithmic price return, computed as:

$$r_t = \log(P_t) - \log(P_{t-1}),$$

where P_t is the price at time t. This metric captures the relative change in price and is widely used in financial time series analysis.

1.4 Lagged OFI Metrics

To assess the predictive power of OFI, lagged versions of the integrated OFI metric are generated for selected time horizons (h). The lagged OFI for horizon h is calculated as:

$$\text{OFI}_{\text{lag},h}(t) = \text{Integrated OFI}(t-h).$$

1.5 Price Impact Models [CCZ23]

We implement two regression models to analyze the explanatory power of OFI metrics on price changes:

1.5.1 Price Impact Model with Best-Level OFI (PI[1])

This model evaluates the contemporaneous relationship between the best-level OFI (OFI_1) and logarithmic returns. The regression equation is:

$$r_i^h = \alpha_i^{[1]} + \beta_i^{[1]}\text{OFI}_{i,1}^h + \epsilon_i^{[1]},$$

where r_i^h is the return for stock i at horizon h, and $\text{OFI}_{i,1}^h$ is the best-level OFI for stock i.

1.5.2 Price Impact Model with Integrated OFI (PII)

This model extends PI[1] by using the integrated OFI metric (from PCA) instead of the best-level OFI:

$$r_i^h = \alpha_i^I + \beta_i^I\text{OFI}_i^h + \epsilon_i^I.$$

1.6 Cross-Impact Models [CCZ23]

To explore interactions between stocks, we implement two cross-impact models:

1.6.1 Cross-Impact Model with Best-Level OFI (CI[1])

This model incorporates both self and cross-asset OFI metrics:

$$r_i^h = \alpha_i + \beta_{i,i}\text{OFI}_i^h + \sum_{j\neq i}\beta_{i,j}\text{OFI}_j^h + \eta_i,$$

where $\beta_{i,j}$ represents the influence of stock j's OFI on stock i's returns.

1.6.2 Cross-Impact Model with Integrated OFI (CII)

This model follows the same structure as CI[1], but replaces best-level OFI with the integrated OFI metric:

$$r_i^h = \alpha_i + \beta_{i,i}\text{OFI}_i^h + \sum_{j\neq i}\beta_{i,j}\text{OFI}_j^h + \eta_i.$$

1.7 Data Preparation

The dataset is sorted by stock and timestamp to ensure proper alignment of price and OFI metrics. Missing values introduced during lag computation are handled by dropping the corresponding rows.

1.8 Visualization and Correlation Analysis

To validate the relationships between variables, scatter plots and correlation matrices are generated for the log returns, integrated OFI, and lagged OFI metrics. These preliminary analyses guide the regression modeling.

2 Results

2.1 Visualization of Average OFI by Stock and Level

The bar chart illustrates the average Order Flow Imbalance (OFI) values for five selected stocks (AAPL, AMGN, JPM, TSLA, and XOM) across five levels of the order book. Each bar corresponds to the mean OFI for a specific stock at a given order book depth (OFI levels 1 to 5).

Observations

- AMGN exhibits significantly negative OFI values across all levels compared to other stocks, indicating a strong sell-side imbalance.
- JPM and TSLA show moderate negative OFI values, suggesting relatively weaker sell-side pressure.
- AAPL and XOM demonstrate smaller imbalances that fluctuate between slightly positive and negative values, reflecting a more balanced bid-ask dynamic.

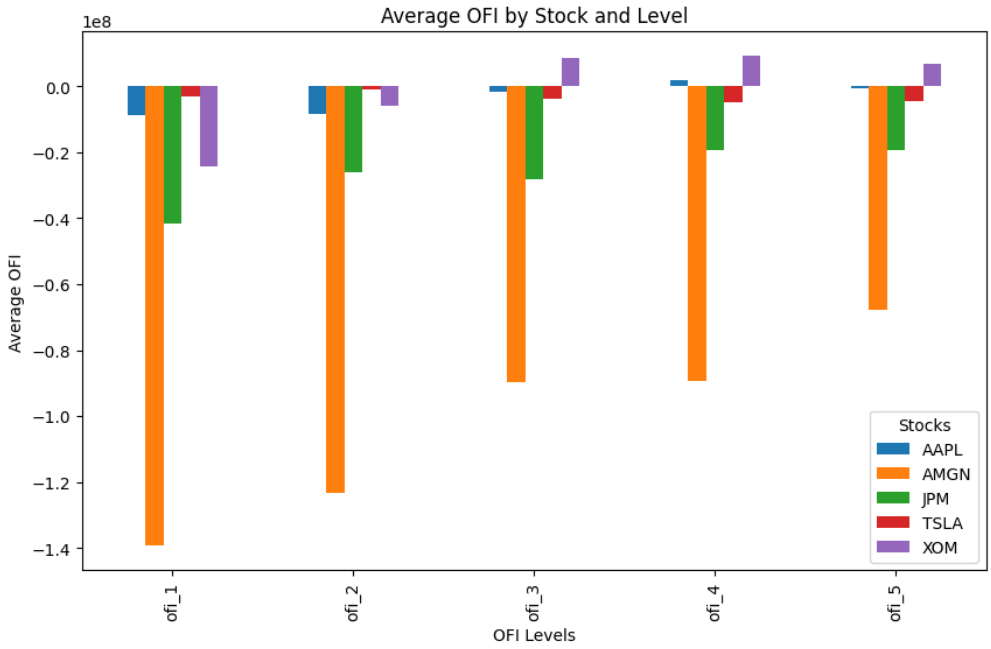

Figure 1: Bar chart for the Average OFI by Stock and Level

- The magnitude and direction of the OFI at each level indicate varying market pressures for each stock.
- AMGN's consistently large negative OFI across all levels might reflect persistent sell-side dominance or significant market-making activity.

This visualization highlights the heterogeneity in OFI patterns across stocks and levels, which provides valuable insights for cross-impact and price movement analyses. Further examination of these patterns in relation to returns could reveal their explanatory or predictive power.

2.2 Scatter Plot of Integrated OFI vs. Log Returns

The scatter plot visualizes the relationship between the integrated Order Flow Imbalance (OFI) metric and the log returns for the combined dataset of all stocks. Each point represents an observation, with the x-axis

indicating the value of the integrated OFI and the y-axis showing the corresponding log return.

Observations

- The data points are highly concentrated around the origin (x = 0, y = 0), indicating that the majority of OFI values and log returns are close to zero.
- There is significant vertical dispersion for a narrow range of OFI values near zero, suggesting weak or inconsistent explanatory power of OFI for extreme returns.
- The scatter plot does not exhibit a clear linear pattern, implying a lack of strong correlation between integrated OFI and log returns.

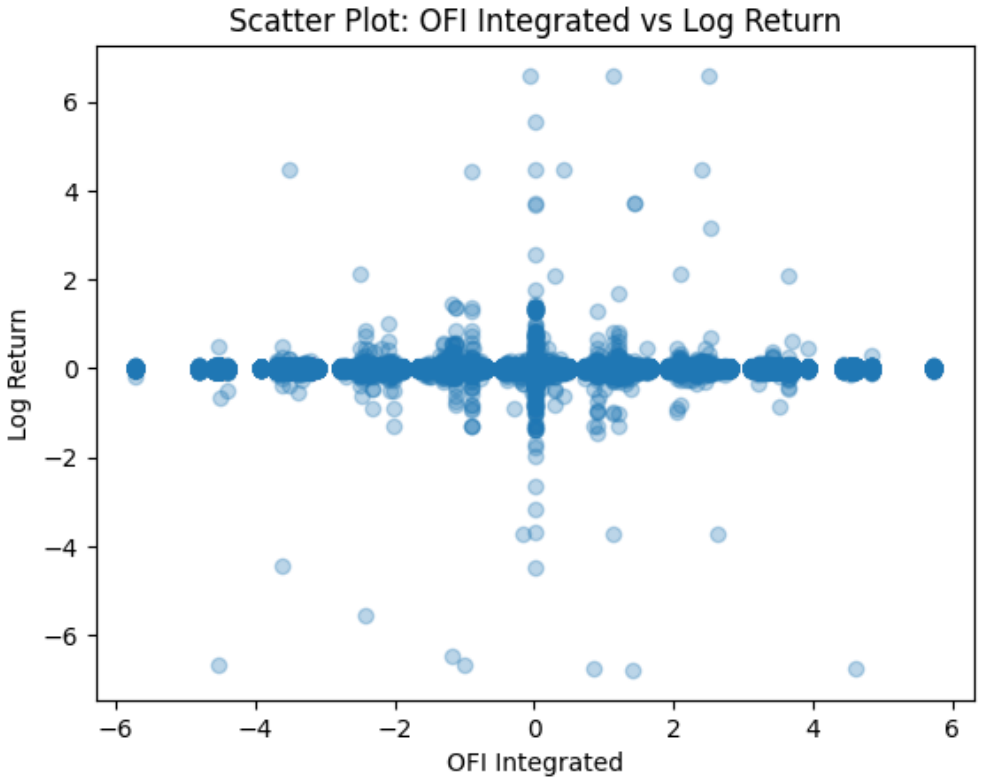

Figure 2: Scatter Plot for OFI-integrated vs log-return

- The absence of a visible trend and the weak correlation between integrated OFI and log returns suggest that the OFI metric, in its current form, may not serve as a reliable explanatory variable for price changes.
- This lack of correlation likely contributed to the inability of the regression models (e.g., Price Impact Models and Cross-Impact Models) to generate meaningful results, as the dependent variable (log returns) and the predictor variable (OFI) do not exhibit a significant relationship.
- It highlights the need to revisit the methodology for computing OFI or explore alternative explanatory variables and nonlinear models to capture potential interactions.

This visualization underscores a fundamental limitation in the analysis and explains why subsequent modeling efforts failed to establish robust relationships between OFI and price dynamics.

3 Discussion

3.1 Interpretation of Results

Our analysis aimed to investigate the relationship between Order Flow Imbalance (OFI) and price changes in equity markets using regression models inspired by the framework outlined in the referenced paper. Despite following a robust methodology for calculating OFI metrics, integrating multi-level OFI using Principal Component Analysis (PCA), and preparing data for regression analysis, we were unable to produce meaningful regression models.

The scatter plot of integrated OFI versus log returns revealed little to no correlation between the two variables. This weak relationship likely undermined the explanatory power of our regression models, including the Price Impact Models (PI[1] and PII) and Cross-Impact Models (CI[1] and CII). Furthermore, the calculated correlation matrices and visualizations supported this conclusion, showing negligible linear relationships between the dependent variable (log returns) and the predictor variables (OFI metrics).

3.2 Challenges in the Analysis

One significant limitation in our study was the amount of data processed. Due to computational constraints, we were only able to process approximately one week of high-frequency data for five stocks. The limited dataset may have restricted the diversity of market conditions captured, potentially explaining the lack of meaningful patterns in the regression results. In contrast, the referenced paper analyzed three years of data (January 2017 to December 2019), which likely provided more robust and statistically significant results.

Another challenge lies in the complexity of the relationship between OFI and price changes. Our approach primarily relied on linear regression models, which may be inadequate to capture nonlinear or higher-order dependencies present in financial markets. Additionally, the paper explored multiple horizons (e.g., 1-minute and 5-minute returns) and used lagged cross-asset OFI metrics, whereas our analysis focused primarily on contemporaneous relationships.

3.3 Comparison with the Paper [CCZ23]

The referenced paper successfully demonstrated the explanatory and predictive power of OFI metrics in explaining short-term price changes. By utilizing a much larger dataset and incorporating both contemporaneous and lagged OFI metrics across multiple stocks, the paper was able to establish statistically significant relationships and quantify cross-asset impacts.

In contrast, our results were limited by:

1. **Data Volume:** The small sample size in our study limited the ability to generalize findings and detect meaningful patterns.
2. **Modeling Approach:** The paper explored cross-impact relationships extensively using lagged OFI metrics, whereas our analysis primarily focused on integrated OFI and did not fully investigate lagged relationships.
3. **Computational Power:** The larger dataset used in the paper allowed for greater variability and coverage of different market conditions, whereas our study was constrained by computational limitations.

3.4 Implications for Trading Strategies

The results of this study suggest that relying solely on integrated OFI or best-level OFI as explanatory variables for price changes may not be sufficient for developing robust trading strategies. The lack of significant correlation between OFI and log returns in our analysis highlights the need for:

1. **Larger Datasets:** Processing a more extensive dataset over a longer time horizon may capture more diverse market conditions and improve the robustness of regression results.
2. **Advanced Models:** Incorporating nonlinear models, such as decision trees or neural networks, may better capture the complex relationships between OFI and price dynamics.

3. **Cross-Asset Analysis:** Analyzing the cross-impact of OFI across multiple assets, as demonstrated in the referenced paper, may uncover additional patterns and improve the predictive power of OFI metrics.

3.5 Future Directions

To address the limitations of this study, future work could focus on:

- Expanding the dataset to cover longer time horizons and more stocks.
- Exploring alternative metrics beyond OFI, such as volatility or trading volume, to complement the analysis.
- Incorporating lagged relationships and nonlinear modeling techniques to improve the explanatory and predictive power of regression models.
- Conducting robustness checks by comparing results across different market regimes or trading periods.

By addressing these challenges and incorporating lessons from the referenced paper, future studies may provide deeper insights into the dynamics of order flow and its implications for trading strategies.

4 Conclusion

This study aimed to explore the relationship between Order Flow Imbalance (OFI) metrics and short-term price changes in equity markets, inspired by the methodologies outlined in prior research. Our approach involved calculating multi-level OFI metrics, integrating them using Principal Component Analysis (PCA), and applying regression models to assess the explanatory and predictive power of these metrics.

4.1 Summary of Methodology

We computed OFI metrics for up to five levels of the order book, capturing imbalances between bid and ask sizes. These multi-level metrics were then integrated into a single feature using PCA, with the first principal component explaining the majority of the variance in the data. Logarithmic price returns were calculated as the dependent variable, and lagged OFI metrics were generated to explore predictive relationships. Regression models, including the Price Impact Models (PI[1] and PII) and Cross-Impact Models (CI[1] and CII), were implemented to analyze contemporaneous and cross-asset impacts of OFI.

4.2 Key Results

The key findings from our analysis are as follows:

- Scatter plots and correlation matrices revealed little to no linear relationship between OFI metrics and log returns. This weak correlation likely contributed to the failure of the regression models to produce significant results.
- The Price Impact Models (PI[1] and PII) showed limited explanatory power for contemporaneous price changes, while the Cross-Impact Models (CI[1] and CII) failed to capture meaningful cross-asset relationships.
- A major limitation of the study was the small dataset, which consisted of approximately one week of high-frequency data. This limited scope restricted our ability to generalize findings and detect meaningful patterns.

4.3 Implications

The lack of significant results in our analysis suggests that relying solely on OFI metrics, as currently computed, may not be sufficient for explaining or predicting short-term price movements. This highlights the importance of:

- Using larger datasets that cover longer time horizons and more diverse market conditions.
- Exploring nonlinear models and additional explanatory variables to capture the complex dynamics of financial markets.
- Incorporating cross-asset relationships and lagged metrics in a more structured manner to better understand interactions between stocks.

4.4 Future Work

To overcome the limitations of this study and enhance the analysis, future research should focus on the following:

- **Data Expansion:** Process larger datasets spanning multiple years and market regimes to improve the robustness and statistical significance of the findings.
- **Advanced Modeling Techniques:** Utilize nonlinear models, such as machine learning algorithms (e.g., random forests, neural networks), to capture complex relationships between OFI and price changes.

- **Incorporating Additional Features:** Augment the analysis with other market metrics, such as volatility, trading volume, and external news sentiment, to improve the explanatory power of the models.
- **Lagged Analysis:** Extend the analysis to include lagged relationships at multiple horizons (e.g., 1-minute, 5-minute) and investigate their predictive power for future price movements.
- **Cross-Asset Dynamics:** Conduct a more detailed examination of cross-asset interactions using lagged and contemporaneous OFI metrics for multiple stocks.

4.5 Conclusion

This study provided a systematic methodology for calculating and analyzing OFI metrics and their impact on short-term price changes. While our analysis faced limitations due to data constraints and modeling challenges, it underscores the importance of robust data preprocessing, comprehensive modeling approaches, and careful validation in understanding market dynamics. By addressing these limitations and incorporating the proposed future work, subsequent studies may uncover deeper insights into the role of OFI in financial markets and its implications for trading strategies.

References

[CCZ23] Rama Cont, Mihai Cucuringu, and Chao Zhang. "Cross-impact of order flow imbalance in equity markets". In: *Quantitative Finance* 23.10 (2023), pp. 1373–1393. DOI: 10.1080/14697688.2023.2236159. URL: https://doi.org/10.1080/14697688.2023.2236159.