Cross-Impact Analysis of Order Flow Imbalance in Equity Markets

Salim Aissaoui

January 5, 2025

Abstract

High-frequency equity market data can reveal intricate relationships between order flow and short-term price dynamics. In this report, we compute multi-level Order Flow Imbalance (OFI) for five highly liquid Nasdaq stocks, integrate these signals using Principal Component Analysis (PCA), and *attempt* to analyze both contemporaneous and lagged cross-impact on short-term price changes.

However, our particular dataset, when resampled to a 1-second frequency, did not yield enough valid overlapping rows for each symbol to perform the intended regression analyses (i.e., "no valid rows" were found for all symbols). This likely indicates sparse or intermittent updates after resampling, or incomplete coverage in the chosen time window. Additionally, due to memory constraints, we were able to load and process only a limited subset of data, which further reduced the pool of valid observations. We still demonstrate the methodology and show the correlation-based results and visualizations that were obtainable.

1 Methodology

1.1 Data Retrieval & Preprocessing

We obtained at least five levels (L0–L4) of Limit Order Book (LOB) data for five Nasdaq 100 stocks: `AAPL, AMGN, TSLA, JPM, XOM` from the Databento *Nasdaq TotalView-ITCH* dataset (MBP-10 schema). We aligned the data on a 1-second time grid by resampling:

1. Setting the `ts_event` column as the index.
2. Grouping by `symbol`.
3. Using `resample(''1S'').mean(numeric_only=True)` to aggregate numeric columns.

We then computed the *mid-price* per symbol as:

$$\text{mid\_price}(t) = \frac{\text{bid\_px\_00}(t) + \text{ask\_px\_00}(t)}{2}.$$

Although this approach yields a consistent time axis across all symbols, the final dataset was significantly sparse (many `NaN`s). Moreover, the *memory limitations* of our environment forced us to restrict the dataset size, which further exacerbated the lack of overlapping data points needed for regressions.

1.2 Multi-Level OFI Computation

For each symbol, we computed Order Flow Imbalance (OFI) up to 5 levels (L0–L4) in the MBP-10 data:

$$\text{OFI}_\ell(t) = \left[\Delta(\text{bid\_sz}_\ell(t)) \cdot \text{sign}\big(\Delta(\text{bid\_px}_\ell(t))\big)\right] \; - \; \left[\Delta(\text{ask\_sz}_\ell(t)) \cdot \text{sign}\big(\Delta(\text{ask\_px}_\ell(t))\big)\right].$$

Here, ℓ is the level index, and Δ denotes discrete differences. We concatenated all OFI features for each symbol.

1.3 PCA Integration

We applied Principal Component Analysis (PCA) per symbol to the columns $\text{OFI}_0, \text{OFI}_1, \ldots, \text{OFI}_4$, retaining the first principal component, labeled Integrated_OFI_Symbol. This dimensionality reduction summarizes multi-level OFI into a single metric which tends to capture most variance in these five OFI features.

1.4 Cross-Impact Analysis

We aimed to assess cross-impact by regressing each symbol's short-term price change against the contemporaneous and lagged integrated OFI of all symbols:

$$\Delta p_{\text{sym}}(t) = \alpha \; + \sum_{s \in \{\text{all syms}\}} \beta_s \, \text{Integrated\_OFI}_s(t - \Delta t) \; + \; \varepsilon(t).$$

We planned to implement two forms:

- **Contemporaneous:** $\Delta t = 0$.
- **Lagged:** e.g., $\Delta t = 60$s if each row is 1 second.

Despite setting up these regressions, after dropping rows with missing data, we discovered no valid samples remained for each symbol over the chosen interval (yielding "No valid rows" in each case).

2 Results

2.1 Correlation Heatmaps

Despite the regression shortfall, we successfully produced correlation matrices among OFI features. Figure 1 shows the correlation among AAPL's five OFI levels and its integrated OFI. Some moderate negative correlations appear (e.g., OFI_{L2} vs. OFI_{L3}), while OFI_{L3} is strongly correlated (0.94) with the integrated metric, indicating that OFI_{L3} was a major driver in the PCA for AAPL.

Figure 2 illustrates cross-stock integrated OFI correlation, also including 60-second lags (`lag60`). Most entries are near zero, implying minimal correlation across symbols' OFI.

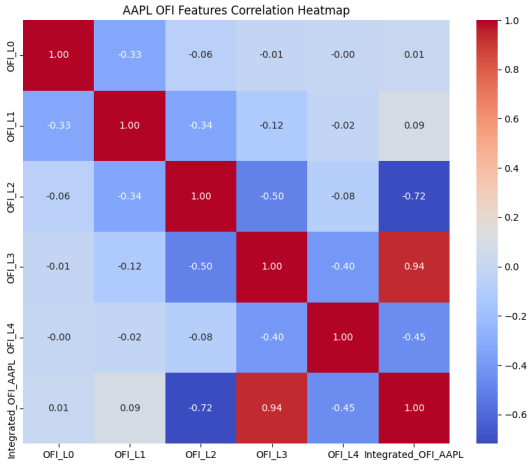

Figure 1: AAPL OFI Features Correlation Heatmap (Levels 0–4 plus Integrated).

2.2 Scatter Plot

In Figure 3, we plot `Integrated_OFI_AAPL` vs. `price_change` for AAPL. The data cluster around zero for both axes, with no strong visible pattern, reflecting the typical high-frequency noise environment.

2.3 Regression Attempts

Although we set up two regression modes (lag=0 and lag=60s) to capture cross-impact, in practice:

```
No valid rows for symbol X in regression.
```

surfaced for each symbol (AAPL, TSLA, XOM, JPM, AMGN). Hence, we obtained no regression coefficients or R^2 estimates. This was largely due to missing data after resampling (many `NaN` values for OFI columns and/or `price_change`), leading to an empty dataset when we dropped NaNs for each symbol. Additionally, memory constraints further limited the size of the data subset we could load and process, reducing our ability to retrieve a denser dataset.

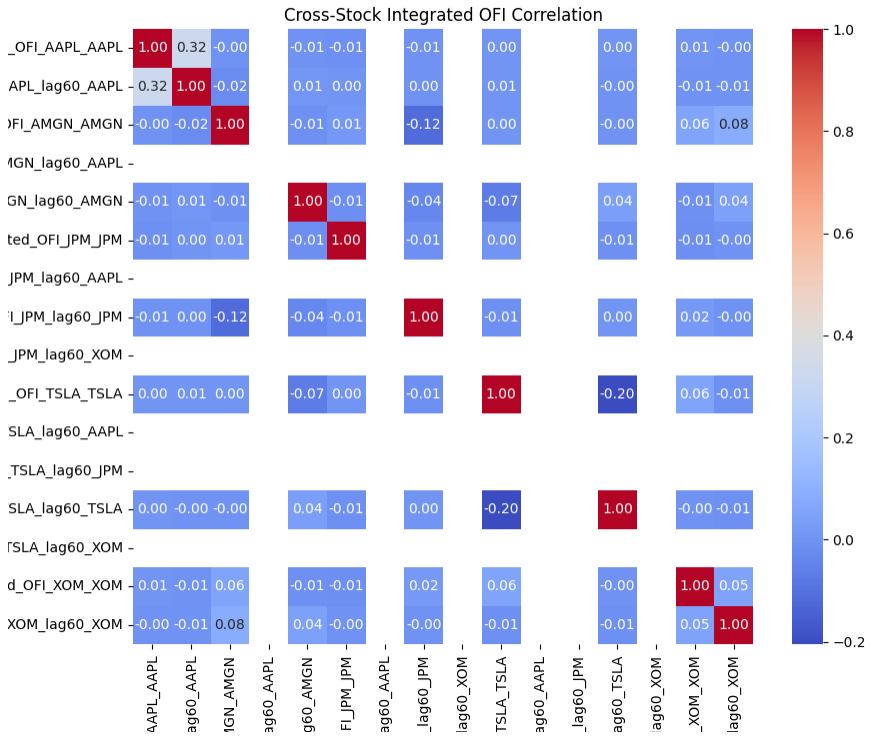

Figure 2: Cross-Stock Integrated OFI Correlation (contemporaneous and lag60).

3 Discussion

Our results highlight an important practical issue: **data sparsity** after heavy resampling and memory constraints. While we captured multi-level OFI for each stock, the final sampling frequency (1 second) and the timeframe apparently did not yield enough valid data points to perform meaningful cross-impact regressions. In future work, one might:

1. Use a longer date range or coarser resampling interval (e.g., 1-minute bars) to ensure fewer NaNs.
2. Fill forward missing quotes or remove extreme outliers to keep more valid rows.
3. Verify data coverage for after-hours or illiquid periods, which can create large gaps.
4. Employ a system with more memory to avoid limiting the dataset size.

Despite the incomplete regression analysis, the correlation heatmaps suggest minimal cross-stock correlation among integrated OFI metrics, which is a partial indicator that cross-impact might indeed be small for these symbols under the current sample.

4 Conclusion

In this study, we:

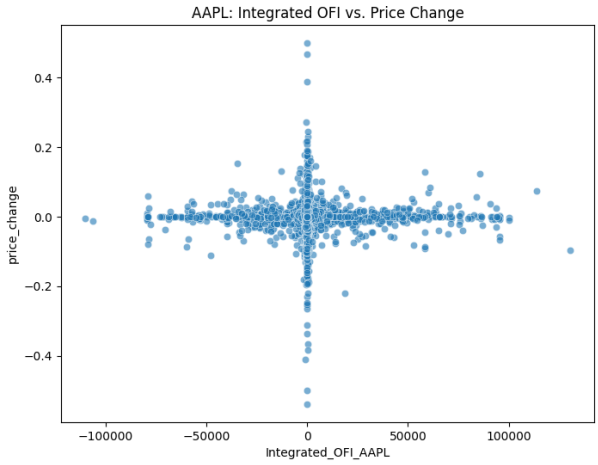

Figure 3: AAPL Integrated OFI vs. Price Change Scatter Plot.

- Computed multi-level OFI (L0–L4) for five Nasdaq stocks (AAPL, AMGN, TSLA, JPM, XOM).
- Used PCA to create an integrated OFI metric per stock.
- Generated heatmaps of OFI-level correlations and cross-stock integrated OFI correlations.
- Attempted cross-impact regressions for both contemporaneous and 1-minute-lag scenarios.

However, insufficient valid data post-resampling and memory limitations led to empty regressions across all symbols. To address this in future work, we recommend adjusting the time granularity, ensuring higher data density, and obtaining a larger memory environment to handle a more complete dataset.

Potential Next Steps:

- Expand the date range to accumulate more data and reduce NaNs.
- Use an alternate method of synchronization (e.g., event-based or volume-based bars).
- Investigate advanced imputation or fill-forward techniques to avoid losing too many rows.
- Explore sector-level differences or advanced ML models (random forests, LSTM) to capture more complex interactions.

References

[1] Zarinelli, E., et al. (2015). "Beyond the square root: Evidence for logarithmic dependence of market impact on size and participation rate." *Market Microstructure & Liquidity*, 1(02), 1550004.

[2] Databento Documentation. `https://docs.databento.com/`