MFE Term 3
Fall 2023

(MFE 230ZB - Report) ADIA: Exogenous Information in HFT

Written by:

Felipe Montenegro
Daniel Trivino
Xinyi Tu
Hao Yan
Zhenghui Liu
MFE Students

Under the mentoring of:

Ali Kakhbod
Professor

September 2023

Haas School of Business
UC Berkeley
2220 Piedmont Ave
CA 94720
UNITED STATES

0.1 Introduction

News play a crucial role in determining asset price fluctuations. As investors obtain new information, they adjust their projections of future cash-flows. These adjustments result in a corresponding change in asset prices to reflect updated valuations. According to the Efficient Market Hypothesis (EMH), at any given moment, the trading price of an asset should equal the fundamental value, taking into consideration all available information.

As explained by Foucault, Thierry, Marco Pagano, and Ailsa Roell in "Market Liquidity : Theory, Evidence, and Policy (Chapter 3)", price innovations result from news that are incorporated through market participants believes. However, this benchmark model fails in practice to explain aspects of intraday volatility. Empirical evidence shows that intraday price volatility exceeds what can be explained by news alone, suggesting that the trading process (order-flow) itself contributes to volatility.

In our paper, we adopt Marcaccioli, R., Bouchaud, J.-P., & Benzaquen, M. "Exogenous and Endogenous Price Jumps Belong to Different Dynamical Classes" classification of jumps. Exogenous jumps are triggered by news and lead to immediate volatility spikes that subsequently decline according to a power-law trend. Conversely, endogenous jumps are characterized by a consistent rise in volatility, followed by a symmetrical decay. Thus, we embark on a comprehensive study to determine whether extreme events in the financial markets can be effectively classified as within these two classes.

Our research is of special interest to extend the Queue Reactive Model (QRM) by Huang, W., Lehalle, C.-A., & Rosenbaum, M. "Simulating and analyzing order book data : The queue-reactive model". As highlighted in their work, the QRM model is not robust on exogenous jumps, making it challenging to accurately fit real data. Consequently, the development of our jump classification algorithm was motivated to contribute to the refinement of the QRM model.

0.2 Data

The limit order book (LOB) is a fundamental component in the architecture of modern electronic financial markets, providing a real-time supply and demand for a specific financial instrument. The structure of the LOB includes the bid side (b_t), which lists the prices and quantities at which participants are willing to purchase the asset, and the ask side (a_t), which details the prices and quantities at which they are willing to sell. These orders are arranged in descending order on the bid side and ascending order on the ask side, with the highest bid and lowest ask prices, known as the "best bid" and "best ask," respectively, representing the current market price. The difference between these two prices is termed the "bid-ask spread". Analyzing the dynamics and micro-structure of the LOB has become a focal point for researchers aiming to understand market behavior, especially in high-frequency trading environments.

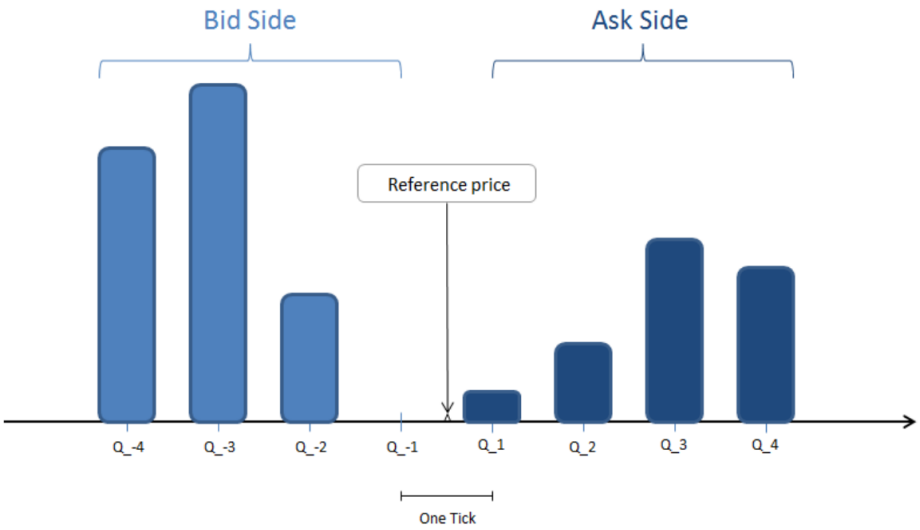

FIGURE 1 – LOB - Huang, W., Lehalle, C.-A., & Rosenbaum, M. "Simulating and analyzing order book data : The queue-reactive model"

To proceed with our empirical study, we selected liquid stocks based on their traded volume during 2019 to achieve standardized returns that are approximately normally distributed. We chose TSLA to showcase our results on the paper, but an exhaustive list can be found in the appendix. Our universe of stocks encompasses continuously traded equities in the Nasdaq exchange from 01/01/2019 to 12/31/2022.

We used Databento as it provides an API with comprehensive order-book data, including bid and ask prices, depth (quantity), and accurate timestamps. Moreover, we focused on the mbp-10 schema which provides up to 10 levels of depth. We filtered on trades and the first two levels of the order book. Then, we resampled the data into different timeframes and achieved similar

results which indicate robustness in the model. However, we chose one-minute snapshots that provide a clear picture of the price dynamics over our observation period to showcase our results.

Our decision to fetch high-frequency data instead of minute time data is the flexibility that it provides when creating different price level aggregations as discussed in the methodology section. This comes at the expense of memory and required machine power as discussed in the challenges section.

From our filtered dataset on trades and the first two levels of depth, we created a subset for market hours and another one for after-market hours. Most papers avoid using after-market hours, but we extend our analysis due to the exuberant number of news that is released after hours.

ts_event	symbol	action	side	depth	price	size	bid_px_00	ask_px_00	bid_sz_00	ask_sz_00	bid_px_01	ask_px_01	bid_sz_01	ask_sz_01
2019-01-02 09:30:00.423428343	TSLA	T	N	0	306.07	308	306.06	306.51	3	3	305.25	306.97	5	250
2019-01-02 09:30:00.548210475	TSLA	T	B	0	306.51	3	306.06	306.97	3	250	305.25	307.00	5	100
2019-01-02 09:30:00.593769772	TSLA	T	N	0	306.95	17	306.06	306.97	3	250	305.25	307.00	5	100
2019-01-02 09:30:00.594026954	TSLA	T	N	0	306.95	33	306.06	306.97	3	250	305.25	307.00	5	100
2019-01-02 09:30:00.594026954	TSLA	T	N	0	306.96	200	306.06	306.97	3	250	305.25	307.00	5	100
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2019-12-31 15:59:59.349468986	TSLA	T	N	0	418.18	5	418.17	418.33	200	1939	418.00	418.34	15	1195
2019-12-31 15:59:59.349517271	TSLA	T	N	0	418.18	15	418.17	418.33	200	1939	418.00	418.34	15	1195
2019-12-31 15:59:59.349517271	TSLA	T	A	0	418.17	85	418.17	418.33	115	1939	418.00	418.34	15	1195
2019-12-31 15:59:59.349565578	TSLA	T	A	0	418.17	15	418.17	418.33	100	1939	418.00	418.34	15	1195
2019-12-31 15:59:59.349565578	TSLA	T	A	0	418.17	100	418.00	418.33	15	1939	417.90	418.34	100	1195

FIGURE 2 – LOB Data

We acquired ticker specific news data through StockNewsAPI for the same timeframe from 01/01/2019 to 12/31/2022. We trust our results are possible due to the curated influx of news articles from credible sources such as CNBC, Zacks, Bloomberg, The Motley Fool, Fox Business, The Street, and many others that are used by many practitioners.

StockNewsAPI offers versatile access options :

— **Ticker-Based Access :** We can retrieve news associated with a specific company ticker for a specified period, including crucial information such as date, ticker symbol, news item, and source, among other parameters.

— **No Filters :** StockNewsAPI enables us to obtain all news articles without applying any filters, ensuring a comprehensive dataset for analysis.

— **Raw Data :** The API provides JSON data that encompasses textual content, images, links, and more, offering a rich dataset for in-depth analysis

We construct a dedicated dataframe using StockNewsAPI data, featuring columns for sentiment analysis, source name, tickers, date, title, and text. This structured data allows us to integrate news-related insights into our research seamlessly.

date	title	text	source_name	sentiment	type	tickers
Tue, 31 Dec 2019 12:58:27 -0500	Tesla must face lawsuit claiming racism at Cal...	A federal judge rejected Tesla Inc's effort t...	Reuters	Negative	Article	['TSLA']
Tue, 31 Dec 2019 09:44:13 -0500	Elon Musk will spend his New Year's Eve workin...	New Year's Eve is just another day at the offi...	New York Post	Neutral	Article	['TSLA']
Tue, 31 Dec 2019 08:55:00 -0500	Why the Stock Market Soured on Tesla and Centu...	The carmaker's CEO had some Boring news to rep...	The Motley Fool	Negative	Article	['CTL', 'TSLA']
Mon, 30 Dec 2019 21:11:35 -0500	Tesla Shanghai reportedly making 1,000 cars pe...	Tesla's Shanghai plant started delivering Mode...	CNBC	Positive	Article	['TSLA']
Mon, 30 Dec 2019 16:51:25 -0500	Bullish on Tesla because of Europe, China: Wed...	Wedbush analyst Dan Ives joins CNBC's "Closing...	CNBC Television	Positive	Video	['TSLA']
...	...	...	...	...	...	...
Mon, 23 Dec 2019 14:03:03 -0500	Tesla Truck Rival Nabs Big Investment From The...	Ford Motor and Amazon are raising bets on elec...	Investors Business Daily	Positive	Article	['TSLA']
Mon, 23 Dec 2019 13:59:08 -0500	China And Tesla Both To Benefit From Shanghai ...	China And Tesla Both To Benefit From Shanghai ...	Seeking Alpha	Positive	Article	['TSLA']
Mon, 23 Dec 2019 13:44:28 -0500	'So high': Tesla shares cross $420 mark over a...	Tesla Inc shares traded above $420 on Monday, ...	Reuters	Positive	Article	['TSLA']
Mon, 23 Dec 2019 11:48:22 -0500	Tesla's stock finally hits $420 a share	Tokes on you! Tesla shares on Monday finally s...	New York Post	Positive	Article	['TSLA']
Mon, 23 Dec 2019 11:15:13 -0500	Tesla's stock just hit a record $420	Tesla CEO Elon Musk once said he had a buyer t...	CNN Business	Positive	Article	['TSLA']

FIGURE 3 – News Data

Having both data sets, we use the timestamp to synchronize them in order to classify the jumps in a further step.

0.3 Jump Modeling

Having high-frequency data, we aggregate it in two different ways which allows us to analyze jumps from various angles. On one hand, we use the mid-price based on the best bid and best ask $(b_t + a_t)/2$. On the other hand, we use OHLC bars to showcase the open, high, low, and close timestamp.

Our research revolves around the "Jump Score" J_t which we define as a metric of standardized returns based on the foundational definition of returns by Andersen and Bollerslev (1997b) and Andersen and Bollerslev (1998b). The authors make the underlying assumption that the returns, denoted as r_t, follow a normal distribution with mean zero. They propose that the standard deviation can be expressed as the multiplication of a deterministic element, f_t (influenced by periodic elements like the time of day, day of the week, etc), and a consistent average volatility factor within a local window K. This leads to the high-frequency return formulation : $r_t = f_t s_t u_t$ where u_t is independently and identically distributed as $N(0, 1)$.

First, we cover the mid-price pipeline. In this case, we use three different summary metrics for our one-minute snapshots. We used the last observation, the mean, and the volume-weighted of the mid-price of each period. We decided to keep the VWMP (volume-weighted mid-price) described as follows as it considers greatly the liquidity component.

We consider the one-minute log-return series on the VWMP. Following a well-known technique on missing values by L. M. Calcagnile, G. Bormetti, M. Treccani, S. Marmi, and F. Lillo, Quantitative Finance 18, 237 (2018), we re-scale the returns during missing periods with the square root of the period length. For example, given the price series $p_0, p_1, \_, \_, \_, p_5$, we construct the following log-returns series $log(p_1/p_0), NA, NA, NA, \frac{1}{\sqrt{4}} log(p_5/p_1)$.

Mid-price returns are known to have approximately zero mean but a strongly fluctuating variance with intra-day seasonality. As such, any standardization procedure must consider both the instantaneous evolution of the variance as well as any seasonality. To standardize the returns, we estimate the consistent average volatility factor s_t using the bipower variation estimator. The bipower variation (BV) estimator is a tool used in the analysis of high-frequency financial data to estimate integrated volatility, which is a measure of how much a financial asset's price moves over a given time period. This estimator is particularly useful because it is robust to the presence of jumps (sudden large price changes) in the data, which can lead to overestimation of volatility.

We develop our formula having on mind that the bipower variation is constant at the daily level.

Thus, K equals the number of returns on a given day with a maximum of 389 observations.

$$s_t^2 = \frac{\pi}{2K} \sum_{i=1}^{K} |r_{t-i}| \, |r_{t-i+1}|$$

Hour	2019-01-02	2019-01-03	2019-01-04	2019-01-05	2019-01-06	2019-01-07	2019-01-08	2019-01-09	2019-01-10	2019-01-11	...
09:31:00	0.001639	0.001195	0.001103	0.0	0.0	0.001204	0.001098	0.001003	0.000918	0.00082	...
09:32:00	0.001639	0.001195	0.001103	0.0	0.0	0.001204	0.001098	0.001003	0.000918	0.00082	...
09:33:00	0.001639	0.001195	0.001103	0.0	0.0	0.001204	0.001098	0.001003	0.000918	0.00082	...
09:34:00	0.001639	0.001195	0.001103	0.0	0.0	0.001204	0.001098	0.001003	0.000918	0.00082	...
09:35:00	0.001639	0.001195	0.001103	0.0	0.0	0.001204	0.001098	0.001003	0.000918	0.00082	...
...	...	...	...	...	...	...	...	...	...	...	...
15:55:00	0.001639	0.001195	0.001103	0.0	0.0	0.001204	0.001098	0.001003	0.000918	0.00082	...
15:56:00	0.001639	0.001195	0.001103	0.0	0.0	0.001204	0.001098	0.001003	0.000918	0.00082	...
15:57:00	0.001639	0.001195	0.001103	0.0	0.0	0.001204	0.001098	0.001003	0.000918	0.00082	...
15:58:00	0.001639	0.001195	0.001103	0.0	0.0	0.001204	0.001098	0.001003	0.000918	0.00082	...
15:59:00	0.001639	0.001195	0.001103	0.0	0.0	0.001204	0.001098	0.001003	0.000918	0.00082	...

389 rows × 364 columns

FIGURE 4 – Bipower Variation

The bipower variation estimator has the following properties :

— It is consistent for integrated volatility in the presence of jumps

— It is asymptotically normally distributed under certain conditions

— It isolates the quadratic variation into a continuous part (the bipower variation) from the jumps

Finally, we take care of the seasonality of the returns on a minute level. We explore the estimation of the non-parametric periodicity estimator proposed by Huang, W., Lehalle, C.-A., & Rosenbaum, M. "Simulating and analyzing order book data : The queue-reactive model" which is a modification of Boudt, K., Croux, C., & Laurent, S. "Robust estimation of intraweek periodicity in volatility and jump detection" with respect to the Weighted Standard Deviation component (W_i).

$$f_i = \frac{W_i}{\sqrt{T^{-1} \sum_j W_{i-j}^2}}, \quad \text{where} \quad W_i = \sqrt{1.081 \frac{\sum_j^{n_i} \Theta\left(-\hat{r}_{j,i}^2 + x\right) \hat{r}_{j,i}^2}{\sum_j^{n_i} \Theta\left(-\hat{r}_{j,i}^2 + x\right)}}$$

In the equation above, $\hat{r}$ represents $\frac{r}{s}$.

Hour	2019-01-02	2019-01-03	2019-01-04	2019-01-05	2019-01-06	2019-01-07	2019-01-08	2019-01-09	2019-01-10	2019-01-11	...
09:31:00	3.473611	3.473611	3.473611	3.473611	3.473611	3.473611	3.473611	3.473611	3.473611	3.473611	...
09:32:00	3.069539	3.069539	3.069539	3.069539	3.069539	3.069539	3.069539	3.069539	3.069539	3.069539	...
09:33:00	3.034337	3.034337	3.034337	3.034337	3.034337	3.034337	3.034337	3.034337	3.034337	3.034337	...
09:34:00	2.550325	2.550325	2.550325	2.550325	2.550325	2.550325	2.550325	2.550325	2.550325	2.550325	...
09:35:00	2.630382	2.630382	2.630382	2.630382	2.630382	2.630382	2.630382	2.630382	2.630382	2.630382	...
...	...	...	...	...	...	...	...	...	...	...	...
15:55:00	0.987941	0.987941	0.987941	0.987941	0.987941	0.987941	0.987941	0.987941	0.987941	0.987941	...
15:56:00	0.752817	0.752817	0.752817	0.752817	0.752817	0.752817	0.752817	0.752817	0.752817	0.752817	...
15:57:00	0.737104	0.737104	0.737104	0.737104	0.737104	0.737104	0.737104	0.737104	0.737104	0.737104	...
15:58:00	0.544137	0.544137	0.544137	0.544137	0.544137	0.544137	0.544137	0.544137	0.544137	0.544137	...
15:59:00	0.722956	0.722956	0.722956	0.722956	0.722956	0.722956	0.722956	0.722956	0.722956	0.722956	...

389 rows × 364 columns

FIGURE 5 – Periodicity Estimator

The following plot showcases the periodicity estimator. As mentioned above, it is constant at the minute level.

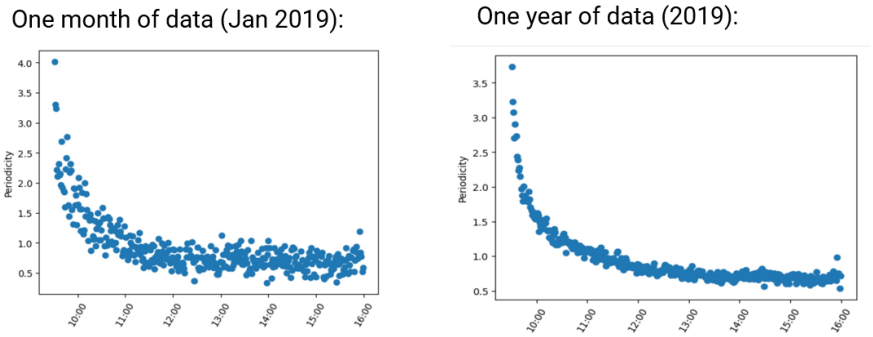

FIGURE 6 – Periodicity Estimator

For the OHLC pipeline, we only changed the consistent average volatility factor to the Glassman-Klass equation which is a robust estimator of volatility for this type of data.

$$GKHV = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\frac{1}{2}\left(\ln\frac{h_i}{l_i}\right)^2 - \frac{1}{N}\sum_{i=1}^{N}(2\ln 2 - 1)\left(\ln\frac{c_i}{o_i}\right)^2}$$

Finally, we get everything together to calculate the Jump Score J.

$$J_t = \frac{r_t}{\sigma_t f_t}$$

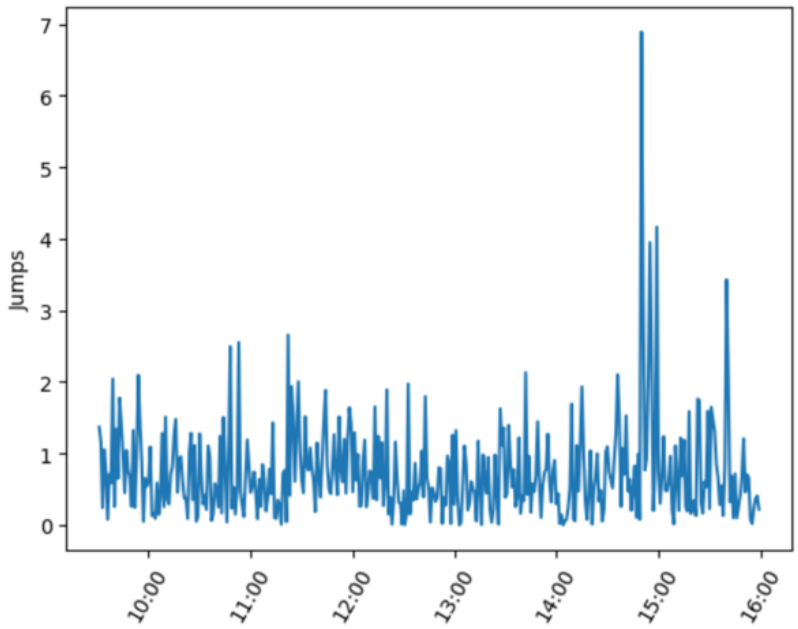

FIGURE 7 – Jump

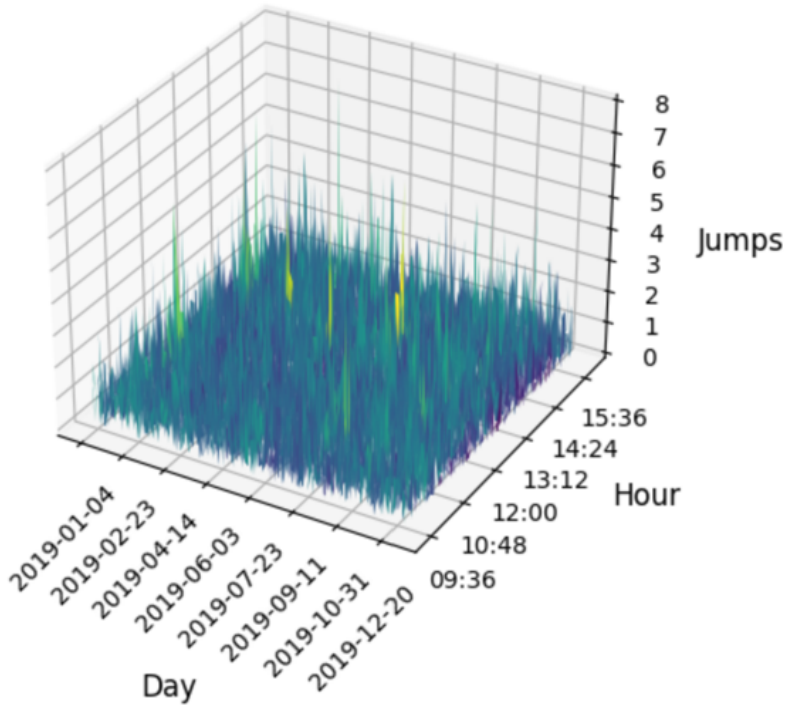

FIGURE 8 – Jumps

At this point, we use the extreme value theory in order to identify jumps for further classification. We achieve this with the following formula and which identified 230 jumps during 2019.

$$|J_\iota| > C_K - S_K \log\left(\log \frac{1}{1-\alpha}\right) \approx 4.36, \quad (K = 390)$$

The ML classification encompasses the uses of a diverse set of features that reflect the volatilify profiles of both type of jumps. The following is a comprehensive list that we are using. Also, it is important to mention that we have models such as logistic and probit regression almost ready to be tested for accuracy.

Main Features:

- Instantaneous jump-score: $|J_t|$
- Exponential moving average of past excess volatility, defined (k – decay measure)

  $\Sigma_t = \kappa|J_t| + (1-\kappa)\Sigma_{t-1}$
- Normalized past price trend: $T_t = \kappa J_t + (1-\kappa)T_{t-1}$
- Binarized past price trend: $B_t = \kappa \frac{J_t}{|J_t|} + (1-\kappa)B_{t-1}$
- Instantaneous average LOB sparsity, defined using the 2 best limit prices:

  $s_t = \max\left[\frac{p_t^a - p_t^{b+1}}{\psi(1+\log V_t^b)}, \frac{p_t^{a+1} - p_t^b}{\psi(1+\log V_t^a)}\right]$

FIGURE 9 – Features

0.4 Challenges

Conducting research is an intricate process fraught with numerous challenges. A significant hurdle we encountered pertained to data acquisition. Our research required access to a news database from Bloomberg. Despite exploring multiple avenues, we soon realized that we lacked the appropriate license to procure this dataset. After weeks of diligent exploration, considering vendors like Ravepack, RapidAPI, and Factiva, we finally identified the ideal source : StockNewsAPI.

A major concern of our research was the inconsistency in the outcomes from the periodicity estimator, which serves as a benchmark for anticipated results. Ideally, the periodicity should yield objective results, unaffected by certain testing environment variables like the chosen period or stock. To illustrate this discrepancy, the subsequent figure displays the Periodicity estimator exhibiting an unexpected spike around 13 :00 hours, deviating from our expectations.

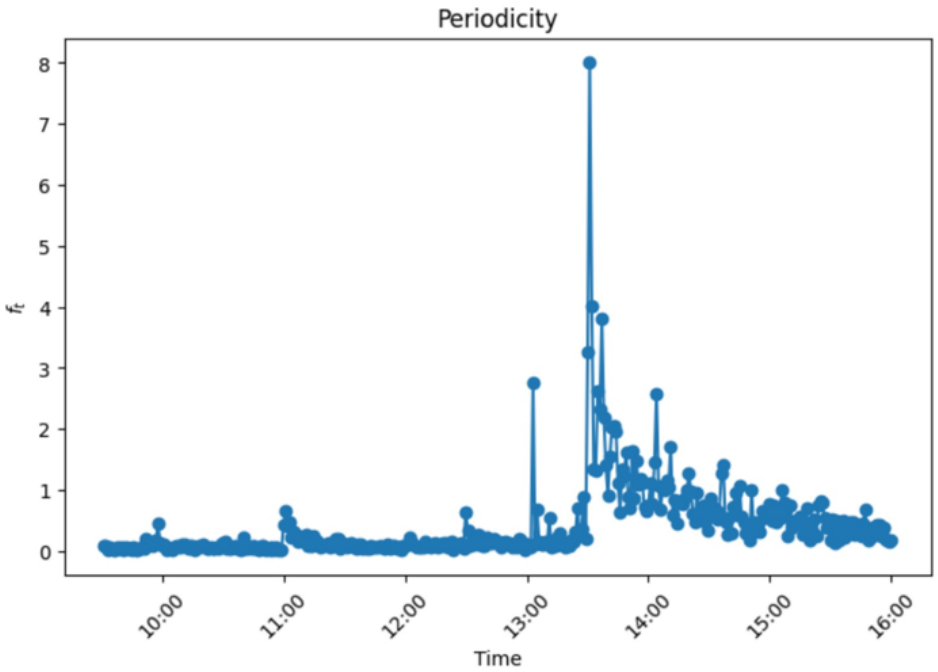

FIGURE 10 – Wrong Periodicity Estimator

We adopted a top-down approach to pinpoint the error in our process. First, we looked at our calculation and algorithm. Yet, after several weeks without discernible progress, we opted for a more granular examination. We decided to explore the data by comparing it with Yahoo finance. By juxtaposing our data with that of Yahoo Finance, it became evident that our data was flawed. A meticulous investigation revealed that the high-frequency data sourced from Databento had an incorrect time zone.

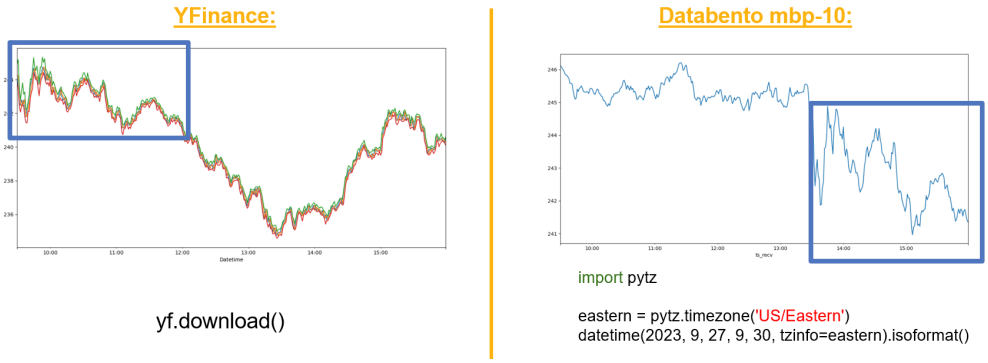

FIGURE 11 – Wrong Time Zone

Databento's API operates on Universal Time Coordinated (UTC), the primary global standard for regulating clocks and time. Consistent worldwide, UTC remains unaffected by seasonal changes. Initially, our requests did not specify a time zone, defaulting to UTC-0 :00. However, once we adjusted to the Nasdaq Exchange's time zone, "US/Eastern", we resolved the discrepancy, aligning with UTC-4 :00.

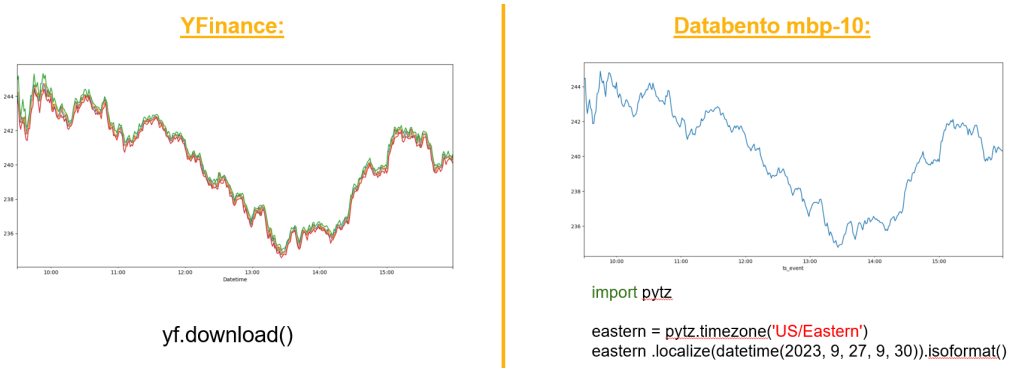

FIGURE 12 – Adjusted Time Zone

Another issue while reading the papers is that some formulas weren't clear on their implementation and we had to contact the authors in order to get some clarification.

Ultimately, we faced issues related to storage and the sheer volume of data. Given that we handle high-frequency data, the file sizes increase dramatically. For instance, the data for TSLA in 2019 alone amounted to approximately 9GB. Consequently, we are contemplating various strategies to manage this challenge as we expand our pipeline to include more stocks over the four-year span of our study.

0.5 Conclusion

This research bolsters the notion that the Efficient Market Hypothesis is not consistently upheld in real-world scenarios. Through our empirical study, we delved into the decomposition of high-frequency returns, culminating in the creation of a pivotal metric termed the "Jump Score". This enabled us to distinguish between exogenous jumps triggered by news events and endogenous jumps that occur independently of news. A crucial next step in our research is to employ the metrics outlined in the Jump Modeling section to develop our machine learning algorithm, facilitating more accurate jump classification.

As a final note, we prioritized crafting organized and modular code, housed within a GitHub repository, ensuring its accessibility and usability for anyone interested.

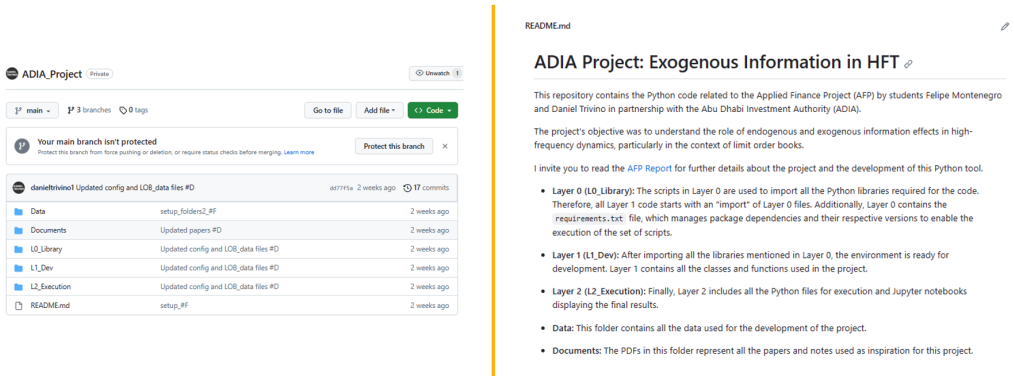

FIGURE 13 – GitHub Repo