Understanding Extreme Price Movements in Large-Cap NASDAQ Equities: A Microstructure and Liquidity-Focused High-Frequency Analysis

Auteur : Geudens, Nathan
Promoteur(s) : Hambuckers, Julien
Faculté : HEC-Ecole de gestion de l'Université de Liège
Diplôme : Master en ingénieur de gestion, à finalité spécialisée en Financial Engineering
Année académique : 2024-2025
URI/URL : http://hdl.handle.net/2268.2/24030

Understanding Extreme Price Movements in Large-Cap NASDAQ Equities: A Microstructure and Liquidity-Focused High-Frequency Analysis

Jury:

Supervisor:
Julien HAMBUCKERS
Reader:
Philippe HÜBNER

Master thesis by
Nathan GEUDENS

For a Master's degree in Business Engineering, specialization in Financial Engineering

Academic year 2024/2025

Acknowledgments

I would like to express my sincere gratitude to my supervisor, Mr. Julien Hambuckers, first for the quality and usefulness of his financial courses, and then for his advice and trust throughout the completion of this thesis. I am also thankful to Mr. Philippe Hübner for agreeing to be the reader of this work, for his availability, and for the interest he has shown in this research.

My heartfelt thanks go to my family, close ones and friends for their constant support, with a special thought for my mother whose presence has been invaluable, to my girlfriend for her kindness and care, and to my colleagues at Engelwood Asset Management for their encouragement.

Contents

List of Figures

List of Tables

List of Abbreviations

Abbreviation	**Meaning**
ACF	Autocorrelation Function
AUC	Area Under the Curve
AD	Average Depth
BLM	Bi-dimensional Liquidity Measure
CBOE	Chicago Board Options Exchange
CCF	Cross-Correlation Function
CDF	Cumulative Distribution Function
CR	Cancellation Rate
CSV	Comma Separated Values
ELP	Endogenous Liquidity Provider
EMH	Efficient-Market Hypothesis
EPM	Extreme Price Movement
FN	False Negative
FNR	False Negative Rate
FP	False Positive
FPR	False Positive Rate
FR	Fill Rate
GLM	Generalized Linear Model
HF	High-Frequency
HFT	High-Frequency Trader / Trading
KDE	Kernel Density Estimate / Estimation

Abbreviation	**Meaning**
LASSO	Least Absolute Shrinkage and Selection Operator
LOB	Limit Order Book
LOBSTER	Limit Order Book System – The Efficient Reconstructor
LOFI	Limit Order Flow Imbalance
LOR	Limit Order Ratio
MedRV	Median Realized Volatility
NASDAQ	National Association of Securities Dealers Automated Quotations
NDLOV	Net Dollar Limit Order Volume
PIN	Probability of Informed Trading
PQS	Percent Quoted Spread
QS	Quote Slope
ROC	Receiver Operating Characteristic
TN	True Negative
TP	True Positive
TPR	True Positive Rate
TSRV	Two-Scale Realized Volatility
VIX	Volatility Index
VPIN	Volume-synchronized Probability of Informed Trading

Chapter I

Introduction

The Flash Crash of May 6, 2010 remains a clear reminder of how quickly modern electronic markets can fall apart. This day, financial markets saw the largest one-day decline in the history of the Dow Jones Industrial Average, with a drop of 9.985%, accompanied by a particularly high trading volume, before regaining almost two-thirds of its drop in value. E-mini S&P500 future contracts also saw a huge sharp intraday decline. As highlighted by Easley et al. (2011), explanations for the flash crash were numerous: a fat-finger trade triggering a cascade of stop-losses, technical reporting difficulties, currency movements, a large purchase of put options by a hedge fund, the sale of many E-mini S&P500 futures contracts by an asset manager, or quote stuffing. However, Easley et al. (2011) argue that the main explanation is linked to liquidity. In the days and hours before the flash crash, a particular situation has been developing: trading volume was high and imbalanced, while liquidity supply was already low, with order flow being increasingly toxic to market makers, causing them to leave the market.

That context shapes the focus of this study: examining intraday Extreme Price Movements (EPMs) through a microstructure lens, with a focus on liquidity. For regulators and stock exchanges, detecting early signs of an increased probability of intraday extreme price dislocation, via liquidity stress indicators, can improve safeguards and enhance monitoring. For risk managers and market makers, such indicators can help guiding decisions on how to set inventory limits, how to avoid adverse selection, and whether to stay active on the supply side of the market. In short, the ability to detect and anticipate intraday EPM risk is an important skill that can ultimately offer, to every stakeholder, financial markets and financial exchanges that are more resilient and more stable.

This study will adopt a microstructure perspective that treats EPMs first and foremost as liquidity phenomena. The liquidity provision of contemporary markets is concentrated in specialized companies, called high-frequency trading firms, or HFT firms. These firms ultimately make money by gaining a thin margin, the bid-ask spread, on an incredibly high volume of trades. Collectively, they shape liquidity by thinning or widening spreads, submitting or canceling limit orders, and managing adverse selection risk based on the perceived toxicity of order flow by sometimes disappearing from their liquidity provider role. The state of market liquidity, held in part responsible for the Flash Crash, is exactly the one this study seeks to capture empirically. Thankfully, the multi-dimensional concept of liquidity is directly observable by looking in-depth at the limit order book (LOB). A precise view and comprehension of the LOB is thus essential, and understanding LOB dynamics is crucial to clearly depict the liquidity supplying state of the market at a given time.

Positioned at this intersection between EPMs and liquidity, the study focuses on intraday EPMs in large-capitalization NASDAQ stocks, using high-frequency LOB data and a set of multi-dimensional liquidity variables that distinguishes liquidity demand from liquidity supply. The working hypothesis is simple: conditional on market-state controls, the current state and recent evolution of the limit order book contain incremental information about the probability of an EPM occurring, with EPMs being labeled, like in Brogaard et al. (2018), as the 99.9^{th} percentile of the distribution of 10-second absolute mid-quote returns.

This perspective leads to two precise research questions that organize the empirical work:

1. *"Among the selection of liquidity and market state predictors, encompassing all liquidity dimensions, which ones most strongly explain an increased probability of a 10-second EPM?"*
2. *"Do liquidity variables deliver incremental predictive power beyond market-state controls such as time-of-day, short-horizon volatility and trading intensity?"*

By addressing these research questions, the primary goal of this master thesis is to offer a clear, microstructure-based view of how liquidity and market-state conditions jointly contribute to the short-horizon detection of Extreme Price Movements, and equally important, to determine if liquidity plays a clear role in EPMs. Beyond only establishing whether liquidity adds incremental predictive power, the study will also seek to reveal the co-movements and tensions among indicators so that potential EPM risk indicators can be interpreted together rather than simply flagged. The intended outcome is a set of signals that practitioners, market makers, and risk managers can monitor in real time. It also aims to provide insights for regulators on the dynamics of liquidity, so they can better assess how safeguards might be calibrated to preserve market integrity during stress episodes.

The study is structured as follows. Chapter II situates the study within market microstructure, moving from the core functions of markets to information asymmetry and adverse selection. It explores liquidity across all its dimensions and link liquidity to concepts such as volatility or order flow toxicity. At the end, it gives a final comprehensive picture of the literature view on microstructure concepts, liquidity, and EPMs. Chapter III presents the exact process of data selection and management. It also details the choice of chosen stocks and study period, and explains rigorously how data is transformed from raw data to ready-to-use 10-second series. Chapter IV then develops the empirical framework, detailing the construction of predictors (market state, liquidity demand, liquidity supply, and hybrid measures), the model specifications, and the in-sample and out-of-sample evaluation strategy. Chapter V reports and interprets the results, evaluating which models, with and without liquidity variables, deliver superior performance, and identifying which liquidity signals are most informative for short-horizon EPM risk. Finally, Chapter VI concludes and discusses implications for practitioners seeking to better understand the root causes of EPMs, together with limitations and directions for future research.

Chapter II

Literature Review

1 Understanding Financial Markets and Their Participants

1.1 The Role of Financial Markets

Financial markets have existed in some form for centuries. The institutional foundations of modern financial markets were established in the early 17th century with the creation of the Amsterdam Stock Exchange, often regarded as the first formal equity market. There, shares in the Dutch East India Company were traded, and concepts such as limit orders, short selling, and market making began to emerge. Over the next few centuries, stock exchanges proliferated across Europe and North America. Over the centuries, the architecture of these markets has continually adapted to innovations in communication, regulation, and trading technology. With the advent of computers and telecommunications in the late 20th century, trading began to shift from physical floors to electronic order books, accelerating price formation, and changing the profile of market participants. This transformation culminated in the rise of high-frequency and algorithmic trading in the early 2000s, creating markets that now operate at millisecond and microsecond frequencies.

Despite these technological evolutions, the core functions of financial markets have remained constant. Markets exist primarily to perform two main tasks: price discovery and liquidity provision (O'Hara, 2003).

First, price discovery is the process through which buyers and sellers interact on a market to establish the equilibrium price of a security. Financial markets aggregate dispersed private and public information through the trading process, leading to the emergence of prices that reflect collective expectations about future payoffs.

Second, financial markets allow participants to enter and exit positions with minimal delay and cost, a function known as liquidity provision. Liquidity enables and encourages market participation by reducing transaction costs and providing flexibility and immediacy for a broad range of market participants, such as portfolio managers. A liquid market allows participants to adjust positions without significantly affecting prices, and thus enhances financial stability. However, liquidity provision is not a characteristic that is constant in its magnitude, as it can vary dramatically across time and assets, and it depends critically on several factors, such as the structure of the market, the state of the market, or the macroeconomic evolution.

The price discovery and liquidity provision functions are related concepts (O'Hara, 2003). In a logical way, liquid markets facilitate trading, and trading, in turn, facilitates price discovery. Indeed, liquid markets allow new information to be absorbed more rapidly, while informationally efficient prices encourage participation and liquidity provision in return. Understanding these microstructure mechanisms is thus essential to interpret price behavior, especially during extreme price movements, where liquidity might be withdrawn from the market and where the price discovery function might not be efficient anymore.

In short, financial markets are dynamic systems in which agents interact under uncertainty, shaping prices through their trades. The efficient functioning of markets, as well as their occasional breakdowns, can only be understood through a deep understanding of both their microstructure design and the profiles of their participants, which leads us to our next section.

1.2 Market Participants

The functioning of financial markets relies on the interactions of different types of agents with different roles, objectives, and information sets. Theoretical and empirical studies on market microstructure emphasize that understanding market dynamics requires identifying who is at the end of each trade and why. Indeed, many asset pricing models in the microstructure theory consider more than one type of trader, such as Easly et al. (2012).

One of the most fundamental distinctions is made between informed and uninformed traders. Informed traders are participants who possess private or superior information about asset values. Their trades play a key role in the price discovery process, as their willingness to buy or sell reflects information not yet fully incorporated into market prices. In contrast, uninformed traders, also referred to as noise traders, operate without an informational edge. Their motivations include liquidity needs, rebalancing, speculation, or behavioral factors such as overconfidence. O'Hara (2003) suggests that behavioral finance plays a key role in understanding why noise traders would trade against better-informed counterparts, who are making profits thanks to their edge.

In addition to these traders, financial markets rely on endogenous liquidity providers (ELPs), such as market makers, who continuously quote bid and ask prices and stand ready to supply liquidity on both sides of the market (the ask side and the bid side). Nowadays, financial markets still function with market makers, but those act as such without having been designated (Brogaard et al., 2019). Their activity helps to ensure that orders can be executed quickly and with minimal price impact. Market makers take on inventory risk and adverse selection risk, as they may be trading against better-informed participants. To manage this exposure, they adjust the width of bid-ask spreads based on expected risks and trading conditions (O'Hara, 2003). Indeed, ELPs can sometimes also be considered as uninformed traders and want compensation for the risk of trading against better-informed agents. This remuneration of market makers is known as the bid-ask spread: the difference between the best ask and bid quotes, which could be considered as a middleman compensation (Glosten & Milgrom, 1985). On the other hand, this bid-ask spread constitutes a transaction cost for the liquidity-demanding side of the trade.

The last two decades have seen the emergence of high-frequency traders (HFTs) as dominant participants in modern electronic markets. These participants use sophisticated trading algorithms and state-of-the-art technological infrastructures to trade rapidly across multiple markets. Although HFTs are often considered liquidity providers, their behavior is fundamentally different from traditional market makers. As shown by Brogaard et al. (2018), HFTs supply liquidity in normal times but tend to reduce their activity during periods of stress or market-wide EPMs (when multiple stocks undergo EPMs simultaneously, i.e., co-EPMs), potentially contributing to market instability. Their ability to withdraw quickly from the market raises questions about the resilience of liquidity during shocks and the role of HFTs in amplifying or dampening extreme price changes.

In addition, the classification of market participants often includes institutional investors, such as mutual funds or pension funds, and retail investors. Institutional investors typically trade in larger volumes versus retail traders. Although these categories are less central to high-frequency studies, they contribute to overall market activity.

Understanding the diverse motives and constraints of market participants is crucial to understanding how prices adjust, how liquidity evolves, how liquidity variables can be interpreted, and how extreme price events arise.

2 Understanding a Key Point to Market Microstructure Theory: Information Asymmetry

2.1 Price Discovery Under Asymmetric Information

As mentioned briefly in the first section of this literature review, financial markets serve the crucial function of price discovery. In theory, the price discovery process should lead asset prices to reflect all available information efficiently and continuously (i.e. the efficient-market hypothesis or EMH). This concept is central to financial economics: in an ideal world without friction, prices would adjust instantaneously to the news, and observed prices would match the fundamental value of the asset at all times. In real-world markets, however, this process is imperfect and noisy, as prices emerge from a decentralized mechanism involving agents with unequal access to information and differing incentives.

One of the most important frictions that limits price discovery is asymmetric information. In any market, some traders are better informed than others. These informed traders may act on private signals about an asset's value, while other participants, such as liquidity traders or market makers, do not have the same informational edge. Brogaard et al. (2019) state that trades arising from investors' market orders reveal private information, while quotes from ELPs' limit orders reveal the public information available. Brennan et al. (2018) illustrate well the fact that some investors are better informed than others. Indeed, they find that a significant part of corporate announcements, being unscheduled (M&As, seasoned equity offerings or dividend initiations) or scheduled (earnings announcements), are incorporated into prices before the announcements. This highlights the fact that some private investors act on information before it becomes public.

The price discovery process is more than just informed traders incorporating their information into the price of the concerned stock. Indeed, information-based trades linked to announcements not only concern the specific firm, but can also occur in the competitors' stocks. When an information event occurs for a dominant, large market-share firm, informed traders prefer trading the competitors' stocks which are more vulnerable to information shocks and thus more attractive to trade on (Tookes, 2008). Moreover, the EMH and asset pricing theory suggest that public announcements directly affect the price of securities. However, Love and Payne (2008) show that information is not instantaneously incorporated, and that one third of the information is incorporated via the continuous trading process, especially via the component called order flow (i.e., the difference between buy market orders and sell market orders).

2.2 Adverse Selection Arising From Asymmetry

As outlined in the model of Glosten and Milgrom (1985), the informational imbalance described above gives rise to the classic problem of adverse selection: liquidity providers risk trading at a disadvantage when facing better-informed counterparts. Adverse selection can even be present when information is public, such as newswire from Bloomberg or Thomson Reuters (Riordan et al., 2013). Glosten and Milgrom (1985) state that specialists (i.e., ELPs) fear the adverse selection problem. Since an agent has agreed to trade (with a market order) on the opposite side of the market (the ELP's limit order), he might be doing it because they have informational advantage. This scenario is represented in Figure II.1.

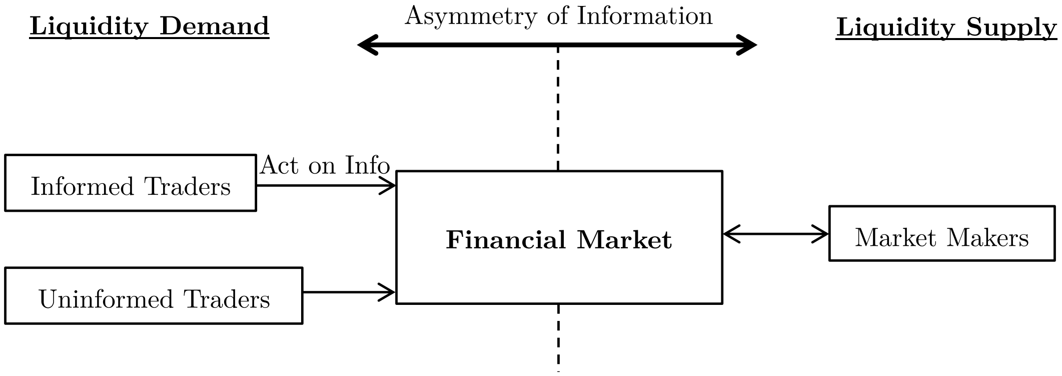


Figure II.1: Information Asymmetry Between Market Agents

Since market makers want to recoup the potential losses of adverse selection, they make gains through the bid-ask spread to compensate for adverse selection. The bid-ask spread is thus in great part an informational phenomenon. The magnitude of the spread therefore depends greatly on the estimation by ELPs of the proportion of informed traders in the market at a given time. The higher this probability that the market maker will transact at a loss due to informational asymmetry, the higher the bid-ask spread. This view on spreads widening has been confirmed many times in the literature, such as in Engle and Russel (1998).

Even if adverse selection is considered a core market friction, the presence of better-informed traders is not only detrimental to market functioning. In contrast, Back and Baruch (2004) conclude that trades move prices precisely because of the possibility that traders may possess superior information. If market participants were known to be uninformed, liquidity providers would simply absorb trades by constantly supplying liquidity. Thus, it is the possibility that a trade comes from an informed trader that drives the price discovery process. Fernandes and Ferreira (2009) even state that it is insider trading that contributes to the fast and reliable incorporation of information into prices, improving market efficiency. However, the presence of those insiders might also lead to a "crowding-out effect": when outside investors perceive that insiders have an edge, they rationally reduce their effort to collect and process information. This weakens the overall informativeness of prices and may lead to a decline in market efficiency (Fernandes & Ferreira, 2009).

To fight against the adverse selection problem, many exchanges have adopted design features designed to discourage informed trading. As discussed by Brolley and Cimon (2020), these features include inverse pricing mechanisms, dark trading venues, and retail order segmentation facilities. A more recent innovation is the introduction of latency delays, where a deliberate pause of milliseconds or microseconds is imposed between the receipt and execution of an order. Brolley and Cimon (2020) find that liquidity improves on the delayed exchange because informed traders have migrated to other exchanges, allowing market makers to reduce their spread.

2.3 Impact of Information Asymmetry on Observed Prices and Expected Return

Information asymmetry does not only impact price efficiency or liquidity provision by ELPs, but also affects how assets are priced and what returns investors expect. Higher liquidity costs (i.e., transaction costs) resulting from information asymmetry can depress asset prices. Indeed, uninformed traders, aware of their disadvantage, demand higher compensation for the risk of holding securities where there is a higher probability of adverse selection, contributing to the equity premium. Crucially, this information risk cannot be diversified away, as it is embedded in the idiosyncratic component of asset returns (O'Hara, 2003). Marshall (2006) further cites Easley and O'Hara (2000) and Easley et al. (2002) to argue that private information increases the risk of holding an asset, leading to higher expected returns.

Stocks with a higher probability of information-based trading have higher information risk which in turn creates a risk premium due to asset prices being depreciated. These findings highlight that observed prices reflect not only fundamentals, but also the risks and costs of trading under asymmetric information. This mechanic is represented in Figure II.2 below.

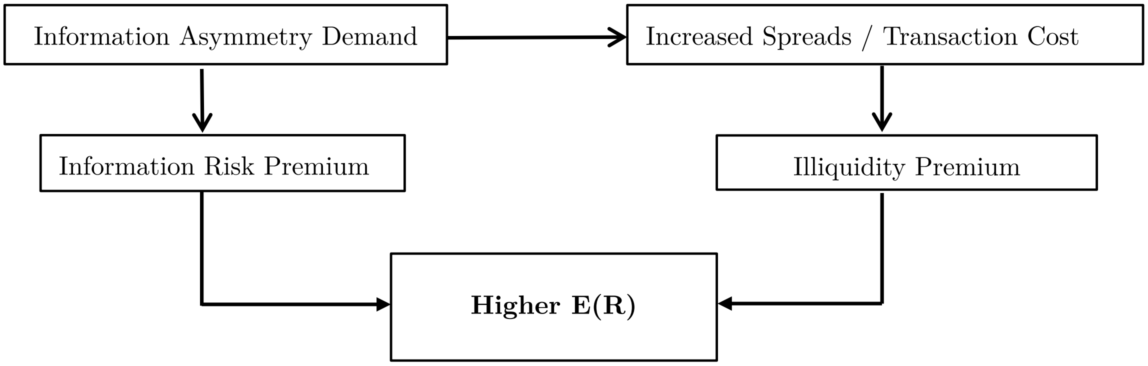

Information Asymmetry Demand
Increased Spreads / Transaction Cost
Information Risk Premium
Illiquidity Premium
Higher E(R)

Figure II.2: Impact of Information Asymmetry on Expected Return

3 High-Frequency Environment and Extreme Price Movements

3.1 The Impact of HFTs on Market Stability and Price Dynamics

The emergence of high-frequency traders (HFTs) as dominant actors in modern markets has substantially transformed how liquidity is supplied and how prices evolve over time horizons. These professional traders are typically classified as endogenous liquidity providers (ELPs) or market makers. Their activity has generated both praise for improving market efficiency and concern for potentially exacerbating market fragility.

One of the core contributions of HFTs is their role in enhancing price discovery. Brogaard et al. (2014) show that HFTs trade in the direction of permanent price changes and against transitory pricing errors, helping price efficiency. This behavior requires the ability to distinguish between informative and non-informative trades. Notably, HFTs predominantly use indicators on liquidity-demanding orders to be able to react and anticipate price moves. As such, they dynamically adjust their strategies depending on market conditions and perceived information content.

However, price discovery is not limited to market orders. In fact, it increasingly occurs through limit orders. Brogaard et al. (2019) highlight that the bulk of limit order activity originates from HFTs and that these passive quotes contribute significantly to the price discovery process, especially under normal market conditions. Even more, they show that HFTs' limit orders contribute more to price discovery than their market orders, whereas non-HFTs market orders contribute more to price discovery than their limit orders.

Yet, when volatility rises, the relative importance of market orders increases, a shift driven by HFT behavior adapting to elevated risk levels, shifting from a standard liquidity supplying standpoint to a liquidity demanding standpoint (Brogaard et al., 2019). The direction of HFTs market orders is correlated with public information (e.g. macro news announcements), market-wide price movements, and limit order book imbalances (Brogaard et al., 2014). Nevertheless, it is not clear whether HFTs withdraw completely from providing liquidity when volatility and uncertainty increase. Indeed, Brogaard et al. (2014, 2018) show that HFTs still supply liquidity during volatile episodes and macro-news releases, absorbing trade imbalances that might otherwise destabilize markets. This stabilizing role is most visible during isolated extreme price movements (EPMs) affecting a single asset, where HFTs trade against the direction of the shock. However, this support diminishes when stress is systemic: during cross-asset EPMs or market-wide dislocations like the 2010 Flash Crash, HFTs tend to withdraw liquidity and begin demanding it. Such behavior underscores the dual nature of HFTs. Although they often act as market makers instead of position takers under normal conditions, as also observed by Easley et al. (2011), they may shift their strategies when faced with elevated information risk or order flow toxicity.

In this context, the strategic behavior of HFTs can either enhance or hinder market stability. This finding points to a conditional impact of HFTs on market dynamics: stabilizing when shocks are isolated and information risk is manageable, while destabilizing the market when uncertainty is widespread or order flows become toxic. This conditionality is particularly relevant for understanding the genesis of extreme price movements, which is the next section of this study.

3.2 Theoretical Views on the Origin of EPMs

Extreme price movements (EPMs) are typically defined as intraday returns that fall within the 99.9$^{\text{th}}$ percentile of the absolute return distribution. These extreme events are characterized by sharp price shifts, often accompanied by higher trading volume and bid-ask spreads (Brogaard et al., 2018). Interestingly, Liu and Yang (2017) show that there is an asymmetric pattern in EPM co-movements. Stock prices tend to fall at the same time but not rise at the same time. Indeed, extreme positive returns occur mainly due to individual stock information release, while extreme negative returns occur mainly due to market-wide pessimistic investor sentiment (Liu & Yang, 2017). Nevertheless, the behavior of positive and negative EPMs appears largely symmetric in terms of their magnitudes, spreads, and trading volumes (Brogaard et al., 2018). It suggests that similar microstructure dynamics may be responsible for both upward and downward EPMs and that, ultimately, studying EPMs can be done regardless of the sign of the absolute return.

A key insight from the literature is that EPMs can arise from two distinct mechanisms: the arrival of new information or imbalances in trading behavior. The first scenario aligns with the classical view of efficient markets, where relevant news leads to rapid, largely permanent adjustments in asset prices (Brogaard et al., 2018; Rif and Utz, 2021). In contrast, the second mechanism attributes EPMs to large, often uninformed trade imbalances, such as when impatient traders aggressively execute large size orders. These volume-driven price shifts are typically transitory and tend to revert quickly as market makers and other participants correct them. Rif and Utz (2021) show that roughly 31% of negative EPMs reverse within the following minute, consistent with the interpretation that many are driven by non-informational shocks. Both of these mechanisms are characterized by heightened trading activity and order flow imbalance, which is represented in Figure II.3.

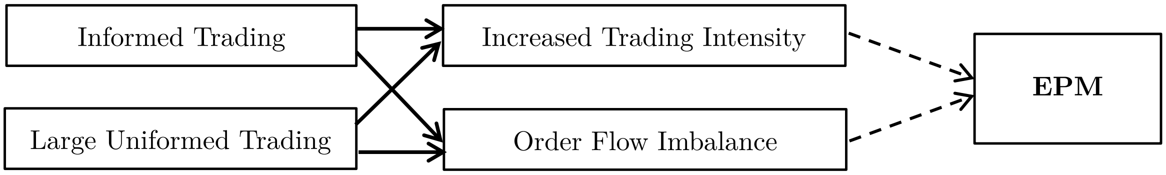


Figure II.3: The Two Types of Events at the Origin of EPMs

Market timing also appears to influence the likelihood of EPMs. Brogaard et al. (2018) report that nearly half of EPMs occur during the first hour of trading, with a moderate concentration towards market close. This aligns with findings by McInish and Wood (1992), who show that the variability of returns shows a U-shaped pattern over the day. These fluctuations can be linked to heightened uncertainty at market open when overnight news is being incorporated, and at the end of the day when traders rebalance portfolios or close positions.

Together, these findings highlight the multifaceted origins of EPMs: some consistent with efficient information processing, others reflecting temporary dislocations. The next section will explore how EPMs relate to other microstructure factors such as spread, volatility, and, most importantly, liquidity dynamics.

3.3 Market Conditions and Liquidity Dynamics Around EPMs

3.3.1 Temporal, Volatility and Trading Activity Patterns

Extreme price movements (EPMs) tend to cluster during specific times of the trading day, particularly at the market open and close. As described above in section 3.2, Brogaard et al. (2018) find that nearly 50% of EPMs occur during the first trading hour, a period characterized by heightened uncertainty and the processing of overnight information. This pattern aligns with the U-shape pattern of McInish and Wood (1992).

However, this U-shape is not limited to price variability. According to Hautsch (2012), nearly all high-frequency market variables, such as trading volume, spreads, and depth, exhibit a similar intraday periodicity pattern. Trading intensity is typically highest right after the market opens and before it closes, which contributes to the clustering of EPMs at those times. Empirical studies also confirm a positive correlation between trading volume and absolute returns, as documented by Karpoff (1987) and Gallant et al. (1992) (both cited in Hasbrouck & Seppi, 2001), suggesting that heightened trading activity creates favorable conditions for EPMs to occur.

Volatility is another crucial variable. Zhou (1996) states that high-frequency financial data exhibit heavy-tailed return distributions and significant volatility heteroskedasticity, meaning that volatility is not constant over time but clusters. They show that this heavy-tail distribution mainly comes from changing volatility. Thus, these volatility bursts often coincide with EPMs, which are located at the tails of the return distribution. Furthermore, volatility is closely tied to trading intensity in the process of increasing the likelihood of EPMs: Engle and Russell (1998) demonstrate that shorter price durations (i.e., the time between successive significant price changes) are associated with higher transaction rates and higher volatility. These findings reinforce the view that EPMs are most likely when the market is both volatile and active.

3.3.2 Liquidity Dynamics and EPMs

While volatility and trading intensity help explain the timing of EPMs, liquidity conditions determine their severity and persistence. One of the most consistent empirical findings is that EPMs are accompanied by deteriorating liquidity: bid-ask spreads widen significantly during such events (Brogaard et al., 2018). Indeed, EPMs tend to occur when spreads are wider following a high transaction rate. On the other hand, narrower spreads will lead to weak price movements even if the transaction rate is high, as liquidity traders are still dominant in that case.

When multiple assets experience EPMs simultaneously, market makers and high-frequency traders tend to reduce their presence in the order book, causing liquidity to dry up just when it is needed most. Hasbrouck and Seppi (2001) argue that just like volatility, illiquidity can become systemic. Historical events such as the 1987 and 1989 equity crashes and the 1998 debt market crisis serve as examples of widespread liquidity breakdowns, where many securities experienced simultaneous price dislocations (i.e. EPMs) due to a collective withdrawal of liquidity.

As a general rule, illiquid securities are more likely to undergo an EPM. As shown by Amihud (2002), expected illiquidity is positively associated with ex ante excess returns, suggesting the existence of an illiquidity premium. Investors demand compensation for holding assets that are harder to sell, especially under stress. This premium is theoretically justified by the work of Amihud and Mendelson (1986) and Vayanos (1998) (both cited in Marshall, 2006), who argue that anticipated future transaction costs lead investors to rationally discount illiquid securities.

The microstructure link between liquidity and EPMs does not stop at the liquidity supply point of view. Market order flow (i.e., liquidity demand) and EPMs are also closely linked. Wu et al. (2020) identify both liquidity variables and market order imbalances as central drivers of EPMs, confirming that many of them arise not only from the inability of the liquidity supply to absorb aggressive trading pressure, but also from information shocks driving liquidity demand. Easley et al. (2011) go further by showing that the VPIN measure of order flow toxicity, a proxy for informed trading pressure measuring order flow imbalance, can predict EPMs better than traditional risk measures like the VIX. Finally, Deuskar and Johnson (2011) find a strong positive correlation between order flows and returns.

In sum, the emergence and severity of EPMs are deeply influenced by both structural market conditions, such as volatility, time of the day, and trading intensity, and evolving liquidity dynamics encompassing both liquidity demand and liquidity supply. Deuskar and Johnson (2011) summarize well the link between liquidity and EPMs: the risk of an asset return is the joint effect of net trade/liquidity demand (flow-driven risk caused by market orders) and the price impact of that demand (illiquidity of the market supply).

This sets the stage for this study's methodology: understanding which factors can influence and predict best EPMs, while verifying the view that liquidity supply and demand mechanisms are strong drivers for the emergence of these EPMs. To do that, the concept of liquidity and the different liquidity variables present in the literature must be looked at in more depth.

4 Liquidity as a Multi-Dimensional Concept

4.1 Liquidity Definition

Liquidity, though intuitively understood by market participants, is a multi-faceted and complex concept in market microstructure theory. One of the earliest formal definitions can be traced back to Keynes (1930, as cited in Hautsch, 2012), who described a liquid asset as one that can be traded quickly, in large quantities, and with little impact on its price. This early definition remains central today, as it captures the practical essence of what market participants seek when referring to a 'liquid' market.

Black (1971, as cited in Cobandag et al., 2022) further refined this view by emphasizing two key aspects of market liquidity: continuity and efficiency. A continuous market is one in which securities can be bought or sold almost instantaneously. An efficient market, on the other hand, ensures that small trades can be executed very near the prevailing market price, while larger trades, when spread out over time, also converge to the current market price. These early definitions stress that liquidity is both a temporal phenomenon and a cost-related phenomenon.

The core idea that unites all liquidity definitions is the notion of low transaction costs. Harris (1990, as cited in Aitken & Comerton-Forde, 2003) defines a liquid market as one in which securities can be converted into cash with minimal cost and delay. Engle and Lange (2001) also describe liquidity as the ability to trade at low cost, which is the cost implied by the bid-ask spread. Indeed, the spread captures the cost-related side of liquidity definition and serves as a proxy for market tightness, one of the dimensions of the next subsection.

However, the spread measure becomes insufficient when considering larger trades. Larger orders typically consume liquidity beyond the best bid and ask prices quotes, resulting in greater price impact and execution cost (Engle & Lange, 2001; Aitken & Comerton-Forde, 2003). This highlights the need to consider other dimensions of liquidity than only the tightness of the market.

4.2 Liquidity Dimensions

While the general notion of liquidity relates to the ease and cost of transacting in financial markets, a more precise understanding requires breaking it down into several components. One of the most influential frameworks in this regard was introduced by Kyle (1985), who defined market liquidity as a combination of three distinct dimensions: tightness, depth, and resiliency.

- Tightness refers to the cost of executing a trade immediately and is typically proxied by the bid-ask spread. More specifically, it captures the cost of reversing a position over a short period and reflects the compensation demanded by liquidity providers to cover adverse selection and inventory risks. As discussed in the previous subsection, the bid-ask spread is a widely accepted metric for this dimension.

- Depth indicates the number of shares that can be traded at the best available prices without affecting market quotes. In practical terms, it reflects the market's ability to absorb large orders without changing the price. Engle and Lange (2001) define it as the number of shares that can be bought or sold within at a given price. It is often measured by the sum of the volumes posted at the best bid and ask quotes in the order book. However, as noted by Deuskar and Johnson (2011), the presence of hidden orders can make depth appear more limited than it truly is, potentially distorting empirical estimates.

- Resiliency captures how quickly prices return to their prior levels after temporary shocks that are unrelated to fundamental value changes. A resilient market corrects itself rapidly following such disturbances, indicating that liquidity should allow markets to be efficient.

Building on Kyle's foundational model, recent literature has expanded the dimensional view of liquidity. Cobandag et al.(2022) propose two additional dimensions that are especially relevant in high-frequency trading environments:

- Breadth refers to the overall volume traded across the market. It provides insight into the market's general capacity to support transactions and reflects broader participation.

- Immediacy is the speed with which a trade of a given size can be executed at a specified cost. It addresses the temporal aspect of liquidity, a feature present in some definitions from the previous subsection.

These five dimensions offer a comprehensive framework for analyzing liquidity and are represented in Figure II.4. They highlight that liquidity is not a unidimensional concept, but rather a complex one. In the next subsection, we explore how these dimensions are measured in practice using high-frequency data.

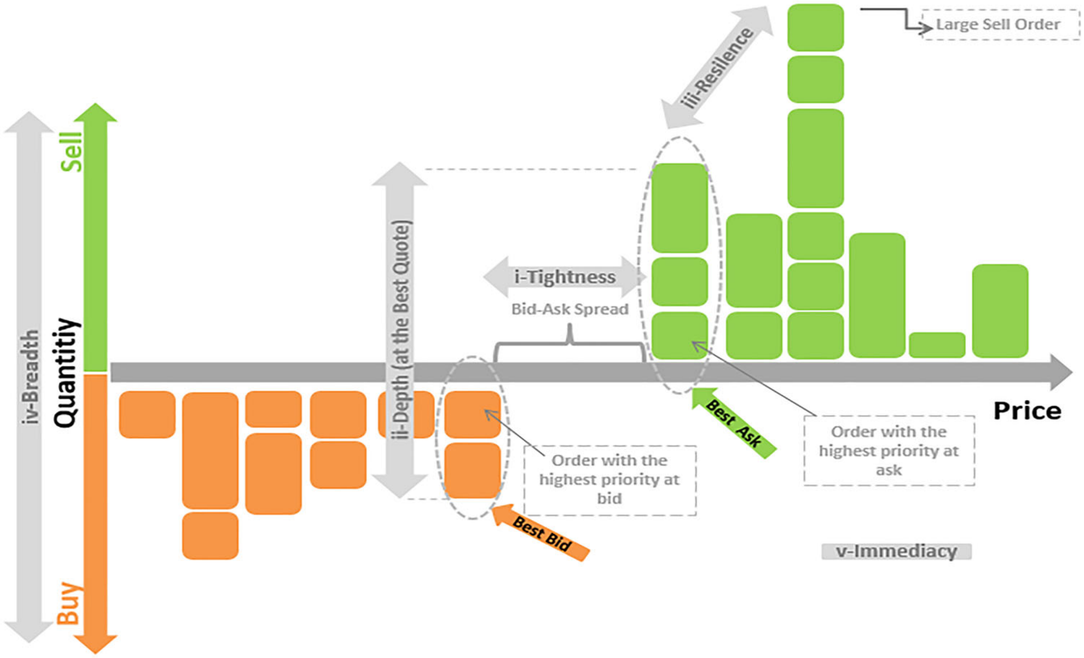


Figure II.4: The 5 Liquidity Dimensions
(Reprinted from Cobandag et al., 2022, p.56)

4.3 Liquidity Measurement

Despite the central role of liquidity in financial markets, there remains little agreement in the literature on how it should be measured (Aitken & Comerton-Forde, 2003; Cobandag et al., 2022). As seen previously, the key reason for this lack of consensus is that liquidity is inherently multi-dimensional, and no single measure can fully capture its complexity.

Liquidity indicators can differ in several respects: they may signal either liquidity or illiquidity; capture one or more dimensions; provide a point-in-time snapshot or reflect liquidity over a period; and finally, they can be ex post, describing past conditions, or ex ante, forecasting expected future liquidity. While many academic studies rely on ex post metrics, ex ante measures are preferred by practitioners as they help anticipate trading costs before execution (Cobandag et al., 2022).

Liquidity measures generally fall into two main categories: trade-based and order-based (Aitken & Comerton-Forde, 2003). Trade-based measures, such as trading volume or order flow imbalance, reflect liquidity demand and are inherently ex post. However, recent literature cautions that trading volume may also capture investor sentiment or attention rather than actual liquidity (Baker et al., 2012, cited in Cobandag et al., 2022). In contrast, order-based measures, derived from the limit order book (LOB), represent the liquidity supply side and are generally regarded as more accurate proxies of true market liquidity. LOB data not only offer the ability to compute current tightness and depth but also enable ex ante estimation of the price impact of a trade, as highlighted by Hautsch (2012).

Given the complex nature of liquidity, this study will adopt a broad approach to liquidity measurement, incorporating a variety of metrics that together aim to capture all key dimensions of liquidity. It will cover both the supply and demand sides of liquidity in order to provide a comprehensive and robust view of how liquidity behaves right before EPMs.

5 Literature Conclusion and Research Objectives

5.1 Reminder of the Core Microstructure Mechanism

As seen in the previous four sections, the microstructure of financial markets is shaped by a complex interplay of informational asymmetries, trading behavior, liquidity dynamics, and ultimately price formation. This literature review follows a narrative that begins with the two fundamental roles of financial markets: price discovery and liquidity provision. It then presents the different market participants: informed and uninformed traders, who act as position taker versus market makers (also called endogenous liquidity providers, often coming from HFT firms) who supply liquidity to the market.

At the core of market functioning lies information asymmetry, a structural imbalance in which informed traders possess superior knowledge about asset values. Their activity creates fundamental friction in the price discovery process. The presence of such informed agents introduces adverse selection risk for uninformed liquidity providers (ELPs), who face the possibility of trading at a loss when interacting with informed market orders. When informed traders begin to trade on their informational advantage, one of the first observable effects is an order flow imbalance, typically caused by a series of aggressive market orders in the same direction. This order flow imbalance can be used as a proxy for the probability of informed trading, as formalized by Easley et al. (2012), presenting their VPIN metric.

5.2 Dependencies and Correlations Between the Main Concepts

Information, trading intensity and volatility. According to Easley and O'Hara (1992), such informational shocks lead to a rise in trading activity and intensity, as more informed trades are initiated. Engle and Russell (1998) also support this idea and note that trading intensity rises following an information event due to the increased number of informed traders, leading in turn to an increase in volatility. Similarly, Engle and Lange (2001) emphasize that the intensity of trade activity can be driven by information asymmetry, citing Foster and Viswanathan (1995), who found that the pace of trading is correlated with both the proportion of informed agents and with volatility. Finally, Hautsch (2012) confirms that large quantities traded often reflect the presence of information and highlights the strong link between information, trading volume, and volatility.

Information, trading intensity, volatility and order flow influence on liquidity. The emergence of informed trading and the resulting order flow imbalances have direct implications for market liquidity. When market makers suspect that aggressive trades originate from superior information, they face heightened adverse selection risk and compensate this risk via the bid-ask spread, which functions as a pricing mechanism for asymmetric information. As shown by Hasbrouck (1988) and Schwartz (1988) (cited in McInish & Wood, 1992), large trades are more likely to carry information, prompting liquidity providers to widen bid-ask spreads in order to protect themselves. In this sense, spread dynamics serve as a response to informational risk. However, even though trading activity indicates a higher probability of informed trading, spreads can also tighten during periods of high trading intensity. Copeland and Galai (1983) argue that higher volumes can reduce the spread through economies of scale in trading (cited in McInish & Wood, 1992). Thus, spread dynamics can go both ways during high trading activity. When increased volume coincides with informational shocks that lead to order flow toxicity, they are more likely to expand.

Volatility further compounds these liquidity dynamics. Deuskar and Johnson (2011) show a positive relationship between price volatility and illiquidity, especially when volatility stems from fundamental uncertainty. Moreover, they observe that net market order selling increases illiquidity, suggesting that liquidity supply becomes more fragile when facing directional pressure from liquidity demand. While higher volume can temporarily enhance liquidity, there is evidence of a lagged deterioration in liquidity supply, implying that sustained trading pressure eventually forces market makers to withdraw from their supplying position (Deuskar & Johnson, 2011).

Final picture (excluding EPM). In sum, the literature reveals a consistent pattern: information-driven order flow imbalances amplify trading intensity and volatility, which, in turn, induce defensive responses from liquidity providers. The result is a deterioration in liquidity, visible through wider spreads and thinner depth, an effect also noted by Engle and Lange (2001), who find that market depth declines in periods of high volume and volatility. This mechanism, summarized in Figure II.5, demonstrates how adverse selection risk, initiated by information asymmetry, ultimately materializes through the erosion of liquidity in the market, creating a fertile ground for EPMs to emerge.

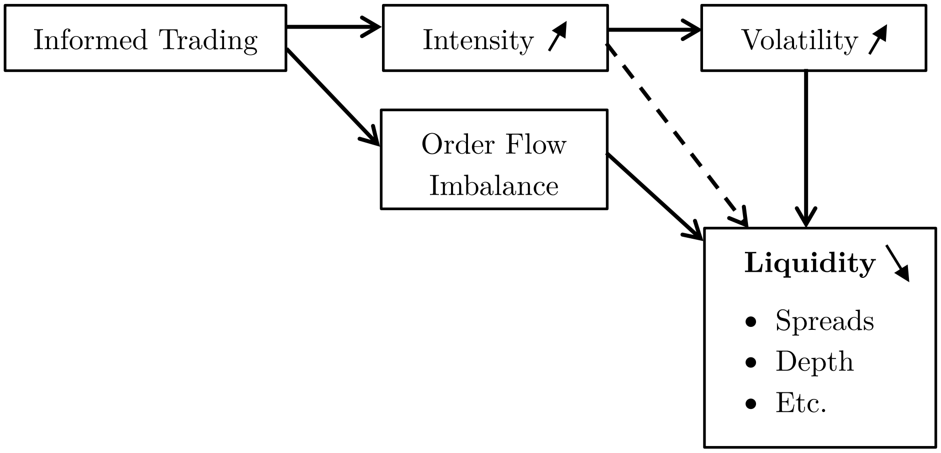


Figure II.5: Interactions Between the Main Concepts

5.3 Literature View on What Causes EPMs and Research Objectives

Building on the insights from section 5.2 above, it is now possible to bring all elements together and synthesize the key factors identified in the literature as potential drivers of EPMs, drawing in particular from the findings discussed in sections 3.2 *Theoretical views on the origin of EPMs* and 3.3 *Market Conditions and Liquidity Dynamics Around EPMs.*

Liquidity demand side. The first drivers of EPMs come from the liquidity demand side of the market. The literature identifies two principal mechanisms responsible for the occurrence of extreme price movements (EPMs): the arrival of new information and imbalances in trading behavior. While distinct in nature, both mechanisms share a common feature: they generate substantial order flow imbalances and periods of trading intensity. This intensity correlates with another important EPM driver: volatility. Indeed, HF return data exhibit heavy-tailed distributions and heteroskedasticity, indicating that extreme returns are happening during periods of heightened volatility. Finally, EPMs tend to cluster during the opening and closing of the market, a time when these mechanisms are more likely to occur.

Liquidity supply side. Beyond activity, order flow, and volatility, liquidity conditions are also critical. EPMs are typically accompanied and influenced by a deterioration in liquidity, as seen, for example, through the widening of bid-ask spreads. When market participants aggressively consume liquidity, market makers react defensively, widening spreads, or pulling back quotes altogether.

Final picture (including EPM). Ultimately, both liquidity demand variables and liquidity supply variables have been shown to be stronger drivers of EPMs likeliness. The literature consistently highlights the joint influence of information, order flow, trading activity, volatility, and ultimately liquidity on each other, as well as on EPMs. All these factors are represented in Figure II.6, the final diagram that summarizes this literature review.

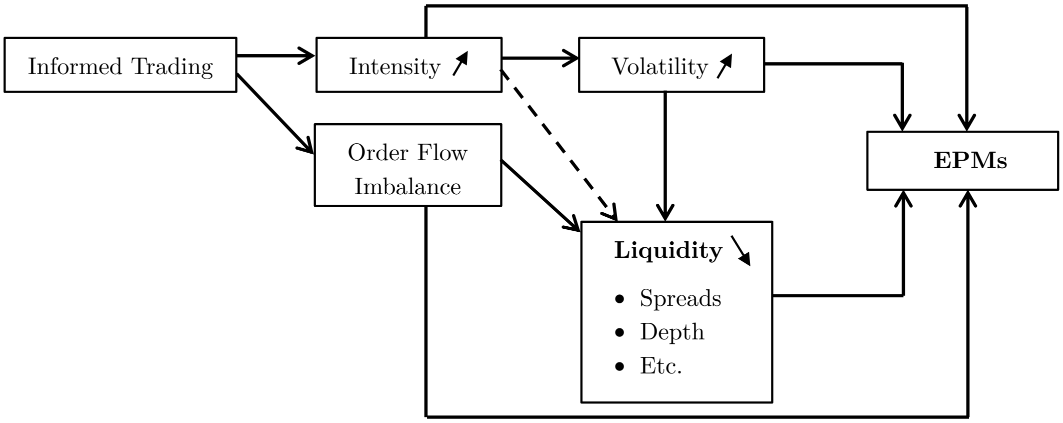


Figure II.6: Interactions Between the Main Concepts and EPMs

This study will build on those findings to further investigate the microstructural determinants of EPMs, with a particular focus on liquidity. The objective is to empirically assess which liquidity variables have the strongest influence on the probability of EPM occurrence. By combining a high-frequency data approach with a multi-dimensional liquidity framework, the research seeks to identify the specific conditions under which heavy price dislocations occur. This will contribute not only to the academic understanding of EPMs, but also to practical insights for market participants and regulators aiming to mitigate the systemic risk arising from these market movements.

Chapter III

Data Collection, Cleaning and Management

1 Data Collection

1.1 Stock Choice

This study is going to focus on the 10 largest market capitalization stocks traded on the NASDAQ (since our data comes from the NASDAQ limit order book). The stocks chosen for this analysis are, by descending market capitalization: Apple (AAPL), Microsoft (MSFT), Nvidia (NVDA), Alphabet Inc. Class C (GOOG), Alphabet Inc. Class A (GOOGL), Amazon (AMZN), Facebook (FB), which is the old name and ticker for Meta Platforms (META), Broadcom (AVGO), Tesla (TSLA), Netflix (NFLX) and Costco (COST).

These stocks have been chosen based on their current market capitalization in April 2025, which introduces a strong look-ahead bias since the studied period begins in 2020. However, it can be noted that most of these stocks (e.g., the GAFAM stocks) were already the strongest market capitalization in the world in the beginning of 2020, reducing this bias inducing potential overperformance in the stock returns.

The focus on large-cap stocks is consistent with some prior literature on EPMs and high-frequency data, including Brogaard et al. (2018), who based their approach on similar considerations from Andersen et al. (2001). Their analysis focused on the 40 largest NASDAQ stocks due to the infrequency and insufficiency of trading activity in mid-cap and small-cap stocks. As noted in their work, "Medium and small stocks trade rather infrequently, and there are usually insufficient observations to draw statistically robust conclusions about HFT and nHFT activity." (Brogaard et al., 2018, p.255).

By selecting heavily traded securities, we ensure using a rich dataset with frequent trade order observations, which is key to studying high-frequency microstructure dynamics. However, this approach contains some limitations, other than the fact that only large-cap companies are used. The selected stocks predominantly come from the technology sector, even if Tesla, Netflix, and Costco can be considered coming from, respectively, the electric vehicle, entertainment, and consumer staples industries. Moreover, all companies are based in the United States, which limits the geographical diversification of this study. Despite these limitations, the use of highly liquid, large-cap stocks offers a robust and practical starting point for analyzing market microstructure and liquidity effects at ultra-high frequency.

1.2 Study Period Choice

When selecting a study period, one must carefully select it as it might influence the validity and relevance of the findings. Several factors can be taken into account before making that choice.

First, the period must be sufficiently rich in data to have better significance in the potential conclusions. On the other hand, selecting too many data points might be heavily consuming considering eventual time and computer power limitations.

Second, one should be aware of structural changes in the financial markets (e.g., one should consider the growing effect that high-frequency traders and trading firms might have on the market) because it could limit the comparability of results with other periods (i.e., studying periods too far away in the past might bring conclusions that are inapplicable nowadays).

Third, market regimes should be taken into account, whether the regime is marked by stability or crisis, as models estimated in one regime might perform poorly in other regimes. Additionally, attention should be paid to some seasonal effects existing in the financial markets, such as the January effect or earnings announcement cycles.

All the above-mentioned attention points should ensure that any conclusions drawn are robust and meaningful in the current market environment. To try to respect them, the period chosen for this study is the full calendar year 2020 as a primary sample period. Indeed, 2020 was a year market by an extraordinary volatility peak in the market following announcements linked to COVID-19 and lockdown. The VIX, the CBOE Volatility Index, also called “fear index”, a measure to represent the 30-day expected volatility of the United States stock market, reached its highest level since the 2008 subprime crisis during March 2020. This makes 2020 one of the most volatile periods in recent financial markets history, while also containing periods of low to standard volatility before and after the March crash. The year 2020 thus provides an opportunity to study EPMs during an exogenous informational shock linked to unprecedented political measures.

Linked to the attention points cited above, choosing a complete calendar year allows us to mitigate seasonality biases by covering all four seasons. Furthermore, choosing a recent period ensures that the findings of this study will still hold some relevance today. The year 2020 also deals with periods of high market stress combined with more calm periods, making it a great candidate for studying EPMs regardless of the market regime and mitigating for effects it could have on findings.

The sample will be divided into an in-sample period from January to October used for model development and estimation, and a pseudo-out-of-sample period from November until the end of the year reserved for testing the model’s predictive performance. Easley et al. (2021) highlight in their study that: “As most empirical research in the market microstructure literature follows an in-sample procedure, without out-of-sample cross-validation, it is possible that some established empirical results are artificial.". This study will thus focus more heavily on out-of-sample results rather than just in-sample results, for which the high accuracy is generally due to model overfitting.

1.3 Database Description

To be able to study liquidity factors around EPMs at ultrahigh-frequency precision, the dataset chosen should be as granular as possible. The empirical analysis in this study is thus based on ultra-high frequency limit order book data provided by LOBSTER (Limit Order Book System - The Efficient Reconstructor), a data provider for the academic community. LOBSTER provides tick-by-tick data for all NASDAQ traded stocks. For each stock and each trading day, LOBSTER delivers two CSV files: a "message" file and an "orderbook" file. Both files are precisely timestamped in seconds after midnight with nanosecond resolution, enabling detailed intraday analysis. As explained by Huang et al. (2015), understanding LOB dynamics is essential, as it helps regulators develop effective policies, enables market makers to offer liquidity at lower costs, and allows investors to reduce transaction expenses. This is the reason why the following sections will focus on thoroughly understanding the “message” and “orderbook” file provided by LOBSTER.

Description of the message file. In the "message" file, each row corresponds to a specific action taken by an individual on the NASDAQ stock exchange (each CSV file concerns one specific stock on a specific day). Each row has a "Type" associated, which is basically the type of action made by the individual.

Consideration should be given to these five types:

- Type 1: Submission of a new limit order
- Type 2: Cancellation (partial deletion) of a limit order
- Type 3: Deletion (total deletion) of a limit order
- Type 4: Execution of a visible limit order
- Type 5: Execution of a hidden limit order (it was not possible to see the order in the order book)

For each row, there is also a precise time in seconds after midnight, an Order ID to which the action refers (a Type 5 row has an Order ID 0 associated), the size of the action in number of shares, the share price associated to the action, and finally the direction of the action.

The direction is an important concept to understand for accurate data management. The direction always corresponds to a limit order, not a market order. The reported direction by LOBSTER can either be 1, indicating that we are observing a buy limit order, or -1, indicating that we are observing a sell limit order. However, the interpretation of this sign depends on the "Type" of action. When considering a submission or deletion of a limit order, it is straightforward that a direction 1 (resp. -1) is a submission/deletion of a buy (resp. sell) limit order. On the other hand, an execution of a limit order of direction 1 (resp. -1) means that an already existing buy (resp. sell) limit order was executed thanks to a new seller (resp. buyer) on the other side of the market. Thus, an execution of a buy (resp. sell) limit order should be reported as a sell trade (resp. buy trade) coming from a market order. This will be of great importance when computing trading intensity variables as well as other liquidity variables.

Thanks to the precision of the LOBSTER data, there will be no need to infer trade directions by using classification algorithms such as the ones discussed by Jukartis (2022), who himself was able to compare different algorithms against data that already contained exact trade directions (similar to LOBSTER data). This will even allow us to further calculate liquidity metrics with greater precision. For example, the VPIN metric, which normally uses a Bulk Volume Classification algorithm for the direction of trades (cf. Easly et al., 2012), will be computed using precise and reliable trade direction data.

Description of the "orderbook" file. The "orderbook" file has exactly the same length as the corresponding "message" file (same stock and same day). Each row in the orde rbook file thus corresponds to the actual state of the limit order book right after the action of the "message" file. The state of the order book is simply all the best ask and bid prices with their corresponding size (in number of shares).

A key methodological decision concerns the depth of the order book included in the analysis. LOBSTER allows to select the number of price levels on both the bid and ask sides. For this study, data were downloaded at level 5, which means that the five best bid and five best ask prices and their corresponding volumes are recorded. This choice constitutes a trade-off between informational richness of the data and availability of computational power. While deeper levels (which can go up to 200 in LOBSTER files) offer additional detail, most liquidity variables of interest are sufficiently captured within the top five levels. In short, limiting the depth to five levels allows for an already detailed and dense dataset while accounting for the limited computational resources. It also has to be noted that the more levels one chooses, the higher the proportion of Submission and Deletion of orders inside the dataset there will be compared to Execution of orders.

2 Data Cleaning and Management

Before any analysis can be performed, raw data should be transformed into a cleaner, more reliable, and usable format to manage. This is even more the case when data is initially sampled at ultra-high frequency. In our case, it could even be said that the data is non-sampled as it is delivered tick-by-tick. The precision and richness of LOBSTER data make it highly valuable, but also demand robust data handling protocols. This master thesis applies a structured cleaning and management framework primarily based on Brownlees and Gallo (2006).

This framework consists of two main parts: data cleaning and data management. The primary objective of this section is to construct a regular and accurate time series of trades, minimizing the influence of anomalies and data outliers.

2.1 Outlier Detection Using the Brownlees and Gallo (2006) Algorithm

As emphasized by Brownlees and Gallo (2006), "the higher the velocity in trading, the higher the probability that some error will be committed in reporting trading information". In order to minimize this risk of error arising in our final dataset, we will apply the algorithm designed in their paper to our raw trade data. Raw trade data corresponds to data already filtered in order to only account for trades (i.e. executions of limit order thanks to the submission of a market order). Their algorithm is the following:

$$(|p_i - \bar{p}_i(k)| < 3s_i(k) + \gamma) = \begin{cases} \text{true :} & \text{observation } i \text{ is kept,} \\ \text{false :} & \text{observation } i \text{ is removed.} \end{cases}$$

The purpose of this filter is to eliminate observations that seem incompatible with the market activity around them. The overall idea of the formula is to detect and remove outliers in transaction prices by comparing each trade to its neighborhood of transactions. The method involves three key parameters:

- k (window size or number of neighborhood observations): This determines how many previous (k/2) and subsequent (k/2) trades are considered when evaluating whether a given trade is abnormal. A smaller k is appropriate for stocks with lower trading activity, while a larger k is appropriate for highly liquid stocks. Indeed, if a stock is not traded frequently, having a too large window size will compare the specific trade with some trades that may have occurred too far away in time. Given that this master thesis focuses on the 10 largest NASDAQ stocks by market capitalization, we adopt a moderately large value of k = 50.
- γ (granularity parameter): This parameter sets the allowable deviation from the local mean price. The lower γ, the more sensitive the algorithm is to small price deviations. Brownlees and Gallo recommend setting γ as a multiple of the minimum tick size. On NASDAQ, most large-cap stocks trade with a \$0.01 minimum price increment, so we follow their recommendation by setting $\gamma = \$0.02$, which means that trades deviating by more than two ticks from the local mean are flagged as outliers (considering a sample standard deviation of the neighborhood of 0).
- δ (trimming percentage of the neighborhood): The trimming parameter allows the removal of a fixed proportion of potential extreme values of the neighborhood before computing the mean. Since a windowed local comparison is used rather than a global approach, and given the relatively high quality of LOBSTER data and the reduction in the frequency of reporting errors over time, we choose a conservative value of $\delta = 5\%$, in line with Brownlees and Gallo's suggestion to choose this value based on the expected frequency of outliers (the lower the probability of outlier, the lower the percentage of trimming).

Overall, these parameter values choices are quite close to the ones used in Brownlees and Gallo (2006). Indeed, they looked at the GE stock in April 2002 (frequently traded stock), used a δ of 10%, and reported that the choice of k = 60 and $\gamma = 0.02$ was more optimal than some other parameter values.

Furthermore, even if this algorithm falsely detects outliers, it erases trades that do not correspond to a true fundamental price change of the underlying asset. Indeed, this algorithm will have a tendency to determine as outliers some trades that are not following the trend of the neighborhood. In other terms, it can select some transitory price movements that do not reflect real shifts in asset value. This notion is well-supported in the literature. For instance, Lee and Mykland (2012) decompose the observed asset price into an equilibrium price and market microstructure noise, emphasizing that observed price jumps often include components unrelated to fundamental valuation. The selected outliers in our data thus probably arise from microstructure noise as they deviate too much from the trend. O'Hara (2003) suggests that this microstructure noise mainly comes from uninformed trades (or noise traders), while long-lasting price movements coming from fundamental price discovery are due to informed traders. Finally, Wu et al. (2020) stress that observed price movements can be decomposed into permanent and transitory components, with the transitory component, corresponding to a temporary deviation from the fundamental value, coming from causes like inventory adjustments or traders' overreactions.

However, a notable limitation of this approach is that some legitimate transitory EPMs may also be filtered out, such as those studied by Brogaard et al. (2018). Indeed, Brogaard et al. differentiates permanent EPMs from transitory ones, the latter being defined as those who revert by more than two-thirds in the next 30 minutes. This trade-off between constructing a clean price series and discarding possibly real but temporary sudden market movements must be acknowledged.

2.2 Creating a Proper Time Series of Trade Prices From Cleaned Trade-by-Trade Data

Now that cleaned trade data is available, some further data management techniques should be applied to ensure adequate data for econometric analysis. The goal of this step is to construct a time series of price evolution suitable for return computation, liquidity variables, and other types of measures. The following aspects need to be handled: simultaneous observations, bid-ask bounce, opening/closing procedures, and the irregularly spaced nature of the data (Brownlees & Gallo, 2006).

Simultaneous Observations. When multiple trades occur simultaneously (with exactly identical timestamps), these are aggregated by computing the median transaction price. It should be noted that, as further aggregation will occur in order to sample the data, which will reduce the frequency of observations, the choice between median price and other aggregation methods is less relevant (Brownlees & Gallo, 2006). The trading volume is naturally defined as the sum of the simultaneous trades' trading volumes. Type 4 observation (visible trade execution) is privileged over Type 5 if both Types are recorded simultaneously. On the other hand, the direction chosen if both trade directions are recorded is 0 (neither 1 or -1) to indicate that both directions occurred at this time point.

Bid-ask Bounce. Another essential correction addresses the bid-ask bounce, a phenomenon in which transaction prices mechanically alternate between the best bid and ask quotes, creating the illusion of price changes that are not driven by fundamental information. To eliminate this problem, the mid-quote price (i.e, the average of the best bid and the best ask) is used instead of raw transaction prices, which will lead to more accurate returns computation (Brownlees & Gallo, 2006; Brogaard et al., 2018), once again reducing the noise induced by market microstructure.

Opening and Closing procedures. To avoid price dislocations related to market opening and closing procedures, the first and last five minutes of each trading day are excluded (the trading day spans from 9:30:00 to 16:00:00), consistent with Brogaard et al.'s (2018) methodology.

It should be noted that eliminating the first five minutes of the trading day might come with a limitation: many information events tend to happen overnight, and a great portion of that information will be incorporated into the prices within those five first minutes. This exclusion of the five first minutes might thus eliminate a lot of potential EPMs, at the advantage of excluding those that are created by opening procedures.

Handling of the irregularly-spaced component of the data. As a reminder, LOBSTER data is recorded at irregular intervals with event-time granularity (a random time increment is separating two observations), which, even if highly precise, is not directly compatible with most statistical models that require equally spaced observations.

To resolve this, cleaned trade data is transformed into regular time series using 10-second intervals, a choice that will enable us to compute 10-second returns. Each trading day is segmented into non-overlapping windows of 10 seconds, with the first window being [9:35:00 ; 9:35:10[and resulting with timestamp 9:35:10. Within the same time window, trade sizes (number of shares traded) are again aggregated, and the last value of the mid-quote price is kept. The last state of the order book is also kept. The justification for using the "last" method over other methods presented by Brownlees and Gallo (2006) (such as "maximum") is that it allows to have the best temporal proximity to the interval's end point and thus to the associated timestamp. For example, the mid-quote price value at timestamp 9:35:00 is the exact last one observed before 9:35:00, meaning that at any precise timestamp, the price associated with it is the last trade having occurred. Furthermore, the choice of the price interpolation scheme for highly liquid stocks tends to yield similar results (Brownlees & Gallo, 2006).

The overall choice between fixed-interval sampling as in this study and other sampling methods is debatable. Brownlees and Gallo (2006) strongly advocate for fixed-interval sampling in high-frequency contexts, noting that "the problem of outliers is less severe when data are sampled at a fixed interval" and that sampling schemes such as those used in computing price durations are prone to errors because outliers will signal movements above a certain threshold. On the other hand, some microstructure variables might require other sampling methods, like volume-based sampling for the VPIN variable (Easley et al., 2012). These variables using other sampling methods will obviously be computed using their initial sampling schemes, and the last value of the variable of interest before a certain timestamp will be retained.

Through the series of transformations described above, the raw LOBSTER data is refined into a reliable high-frequency time series. This processed dataset forms the basis of this study for constructing predictive variables and identifying EPMs.

Chapter IV

Methodology

This Methodology section will be divided into two main parts. The first one, *Variables Computation*, details how the dependent variable and all explanatory variables are computed. The second one, *Methods*, outlines the statistical and machine learning techniques employed to model the occurrence of EPMs and evaluation strategies used to assess predictive performance. Together, these two components provide a comprehensive framework for analyzing the drivers of EPMs in high-frequency trading environments.

1 Variables Computation

As our former irregularly spaced data has been transformed into a time series of 10-second intervals for which the 5 first and last minutes have been taken away, each full trading day is represented by 2,280 intervals, meaning there is 2,280 observations per day. All variables will be calculated for each of those intervals, with the first and last intervals being [9:35:00 ; 9:35:10[and [15:54:50 ; 15:55:00[respectively. On a full trading year (252 trading days) without any trading halt, this amounts to 574,560 observations.

The next sections will describe in depth how the EPM independent variable and variables have been chosen and computed.

1.1 Extreme Price Movements

The cleaned series of prices, as elaborated in the section Data Cleaning and Management, allows to compute the dependent variable and main focus of this study: the Extreme Price Movement.

The procedure for computing EPMs is applied stock by stock, meaning that each stock will go through the same procedure independently and follows a straightforward yet robust methodology. For each stock, absolute mid-quote log-returns[1] are computed between two regularly spaced observations in the cleaned price series, specifically using 10-second intervals as mentioned previously. An EPM is then defined as any such log-returns falling within the top 0.1% of the

[1] Note that the words “log-returns” and “returns” will be used interchangeably in this master thesis. Every computed return is still a log-return.

empirical absolute return distribution (the 99.9th percentile) for that particular stock. This quantile-based threshold is computed using the full dataset (i.e., the entire calendar year), regardless of how data are subsequently split into training and testing sets. As a result, the proportion of EPMs may differ across these sets, which is a realistic feature of financial data given the heteroskedasticity of absolute asset returns.

The dependent variable Y is constructed as a binary indicator: it takes the value 1 if the absolute mid-quote return for the next 10-second interval (using the same interval return would have been extremely biased as independent variables are computed on this interval) exceeds the computed 99.9th percentile and 0 otherwise. By construction, this leads to a highly imbalanced dataset, with EPMs being rare events: a 1 in a 1000 event.

This methodology corresponds to the first of three approaches considered by Brogaard et al. (2018). In their study, the authors propose three definitions for EPMs:

1. Absolute return threshold: The baseline method used here flags EPMs based on absolute mid-quote returns that exceed the 99.9th percentile.
2. Return-adjusted residuals: Returns are regressed on their own lagged values and market-wide returns (S&P 500), and EPMs are identified as residuals in the 99.9th percentile. This method filters out return variation.
3. Lee and Mykland (2012) method: An advanced approach that incorporates estimates of local volatility to detect statistically significant jumps.

Brogaard et al. note that the first and second methods are "agnostic to volatility conditions", which means that they may disproportionately identify EPMs during periods of elevated market volatility. While this characteristic may be seen as a limitation, it aligns with the core objective of this research. This study does not aim to isolate liquidity from volatility, but instead seeks to examine their joint influence on the likelihood of EPMs. In contrast, adopting the second or third method would effectively normalize out return-driven and volatility-driven effects, making it more suitable for studies focused solely on liquidity.

Finally, one known limitation of this approach, also discussed by Brogaard et al. (2018), is that it implicitly assumes that each stock is equally likely to undergo an EPM. In reality, structural differences between stocks could affect their EPM likelihood. Nonetheless, doing this study on a stock-by-stock basis mitigates this concern to a large extent by tailoring the definition of “extreme” to the individual stock’s behavior.

1.2 Independent Variables Choice and Computation

This section aims at presenting the different variables which will be part of our different models to investigate EPMs. It will start with a preliminary remark on quantity computing. It will then present the four different types of variables that are going to be computed on our data: market state variables, liquidity supply variables, liquidity demand variables, and variables that encompass both liquidity supply and demand.

Several methodological choices guide our variable construction. First, due to the regular time sampling applied to the data, we refrain from using duration-based metrics, which are more suited to irregular event-based sampling. Moreover, many of the theoretical concepts behind duration measures are already captured by their fixed-interval counterparts. For instance, trading frequency is closely approximated by volume, or cancellation speed is reflected in the cancellation rate.

Second, for variables that are sensitive to the chosen time interval, particularly those for which 10 seconds of data would be too short, we implement a 5-minute rolling window approach. This approach ensures that such variables retain sufficient responsiveness to local market conditions while smoothing out excessive noise. Choosing a window size is an inevitable tradeoff one has to make between the increased noisiness of estimators when those are based on fewer data points (Hautsch, 2012) and the increased loss of information when ultra-high frequency data is aggregated (Engle, 2000). For that reason, we believe 5-minute intervals reduce much of the noisiness while still keeping the estimators relevant for the predictions of 10-second EPMs.

Third, given that this study seeks to model both positive and negative EPMs in a unified framework, no variable included in the models contains directional information. This choice avoids introducing variables that would interfere with a symmetric understanding of EPMs.

Lastly, care is taken to avoid redundancy between variables. Including too many highly correlated variables that capture the same underlying liquidity dimension could undermine the reliability and interpretability of the results. A precise set of representative variables is therefore favored over an exhaustive list.

At the end of this section, a summary table will present all the variables used, classified by category, and accompanied by short descriptions.

1.2.1 Preliminary Remark: Measuring Quantities in the Limit Order Book

When computing liquidity variables such as trading volume or average depth, a key methodological decision is whether to measure quantities in number of shares or in dollar value. While the number of shares is directly observable, it is naturally tied to the share price, which itself is arbitrary and depends on the number of shares outstanding, an accounting decision that carries no economic meaning. As such, measures based on dollar volume offer a more standardized and meaningful basis for comparison between assets with different share price levels. This approach also aligns with the literature. For example, the well-known Amihud illiquidity measure, though it is not used here due to its daily frequency, relies on a dollar-volume framework (Amihud, 2002). Likewise, Easley et al. (2021) aggregate their high-frequency data into dollar-volume bars. In this study, dollar-volume computations will be preferred where appropriate, as they serve the broader purpose of normalizing liquidity metrics across assets with a heterogeneous number of outstanding shares.

1.2.2 Market State Variables

Time. As seen in the literature review, the time of the day is linked to the probability of EPMs. Indeed, the beginning and end of the days tend to have more EPMs, with around 50% during the first trading hour (Brogaard et al., 2018). The data from this study (the current 10 biggest NASDAQ traded market capitalization during year 2020) also highlights a prominence of EPMs in the earlier part of the day, as seen in Figure IV.1 which represents the kernel density estimation (KDE) of the intraday time on the whole dataset as well as on the dataset filtered for Y=1.

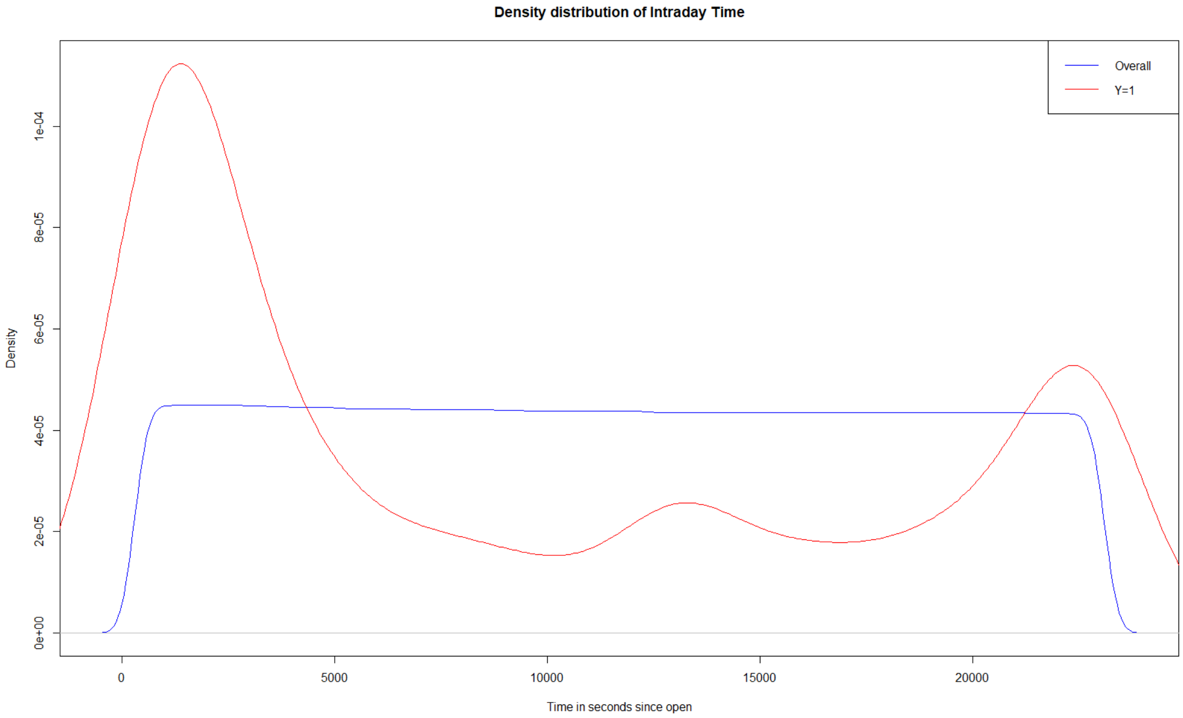


Figure IV.1: KDE of Intraday Time on the Whole Stock Dataset and Filtered for Y=1

For this reason, it was decided to model time as the number of seconds since the opening. This is based on the principle that the closer the time is to the opening, the more likely an EPM should take place. However, this has the limitation of focusing only on EPMs at the beginning of the day. An approach with dummy variables per time of the day section could also have been coherent.

Absolute Return. Past absolute returns possibly contain information on future absolute returns. The absolute return of the 10-second interval will be computed as well as the average of the last intervals using a 5-minute rolling window.

Volatility. When selecting a volatility estimator for high-frequency microstructure data, it is critical to account for market noise, jumps, and short time windows, all of which characterize the environment in which this study is evolving. While standard realized volatility (RV), computed as the sum of squared intraday returns, is a popular and straightforward non-parametric estimator of integrated volatility, it is highly sensitive to market microstructure noise when computed from high-frequency data. Indeed, Zhang et al. (2005) state that "It has been found empirically that the realized volatility estimator is not robust when the sampling interval is small." Therefore, researchers have developed more robust alternatives specifically designed for high-frequency financial applications.

Three notable estimators, which have been considered for this study, address these challenges in distinct ways. The Two-Scale Realized Volatility (TSRV), introduced by Zhang et al. (2005), mitigates i.i.d. microstructure noise by computing realized volatility across multiple sampling frequencies and then correcting for bias between them. While effective in longer time frames with dense data, TSRV underperforms in small samples due to its dependence on subsampling. The pre-Averaging Estimator, developed by Jacod et al. (2009), uses kernel-weighted moving averages of returns to smooth out noise before squaring. However, it introduces the complexity of kernel selection and assumes that noise is i.i.d. (i.e., constant variance in the noise), which might not be the case (Bandi et al., 2013). Lastly, the Median Realized Volatility (MedRV), proposed by Andersen, Dobrev, and Schaumburg (2012), captures volatility through the median of products of adjacent absolute returns, offering robustness to both noise and jumps without requiring tuning or subsampling.

The MedRV estimator is particularly well suited for short rolling windows in high-frequency settings. Its use of the median, instead of the mean, allows it to downweight extreme values caused by price jumps. Moreover, unlike TSRV and pre-averaging, MedRV requires no tuning parameters, making it computationally efficient and non-parametric. In our context of using 5-minute rolling windows over 10-second returns, the estimator performance remains stable even with just 30 observations. This reliability under constrained sample sizes, combined with its jump-robust construction, makes MedRV an appropriate and theoretically sound choice.

The MedRV is defined as:

$$\mathrm{MedRV}_N = \frac{\pi}{6 - 4\sqrt{3} + \pi} \cdot \left(\frac{N}{N-2} \right) \cdot \sum_{i=2}^{N-1} \mathrm{med}\left(|r_{i-1}|,\ |r_i|,\ |r_{i+1}| \right)^2$$

where r_i are the log-returns.

In our case, at the end t of a 10-second interval, MedRV is defined as:

$$\mathrm{MedRV}_t = \frac{\pi}{6 - 4\sqrt{3} + \pi} \cdot \frac{30}{28} \cdot \sum_{i=t-29}^{t-1} \mathrm{med}\left(|r_{i-1}|,\ |r_i|,\ |r_{i+1}| \right)^2$$

1.2.3 Liquidity Demand Variables

Trading intensity. Trading intensity stands at the frontier between market state variables and liquidity demand variables. It is linked to the state of the market because the variable represents the current activity intensity of the market. However, this variable is solely driven by the liquidity demanding side of the market, which is why we categorized it in this section.

Trading intensity is simply measured as traded volume during the 10-second interval in dollars. It will also be computed as the average of the last 5-minutes rolling window.

$$\text{Trading Intensity}_t = \text{Dollar Market Volume}_t = \sum_{i=t-10s}^{t} V_{buy,i} + \sum_{i=t-10s}^{t} V_{sell,i}$$

where:

- $V_{buy,i}$ is the dollar traded volume of buy market orders at time i.
- $V_{sell,i}$ is the dollar traded volume of sell market orders at time i.

Probability of informed trading. As seen in the literature review, the more the amount of informed traders who are trading on the market, the more likely some market agents can be adversely selected. The core idea behind the probability of informed trading (PIN) model is that the order arrival process gives information for subsequent price movements (Easley et al., 2012). This order arrival process, or order flow, can become "toxic" when this order flow goes strongly in one direction and it adversely select market makers who are unaware they provide liquidity at loss. This is why the PIN is represented mathematically by the order flow imbalance, meaning the imbalance between buy market orders and sell market orders.

Like many models evolving in an informationally asymmetric world, the PIN and VPIN models consider liquidity providers against traders / position takers. The model supposes a flow of uninformed traders who trade continuously and informed traders who trade when there is an information event. From this model is derived the probability that a market order arrives from an informed trader, or probability of informed trading (PIN):

$$\text{PIN} = \frac{\alpha\mu}{\alpha\mu + 2\varepsilon}$$

where:

- α is the probability that an information event occurs during period i.
- μ is the arrival rate of informed traders.
- ε is the arrival rate of uninformed buyers and uninformed sellers.

This model also describes the bid-ask spread as follows:

$$\Sigma = \left(\frac{\alpha\mu}{\alpha\mu + 2\varepsilon}\right) \cdot (\overline{S}_i - \underline{S}_i) = \text{PIN} \cdot (\overline{S}_i - \underline{S}_i)$$

where $\overline{S}_i$ (resp. $\underline{S}_i$) is the price the informed trader knows the asset will reach at the end of the trading period when the information event is positive (resp. negative).

This notion of bid-ask spread is logically based on the principle that the higher the probability of informed trading, the more liquidity providers will protect themselves against adverse selection by widening the spread. A bad anticipation in a rise of PIN from market makers will result in losses for them (Easly et al., 2012).

The VPIN measure, or volume-synchronized probability of informed trading, is simply a volume-based approach of the PIN, meaning that the measure is updated in volume time rather than clock time (at first glance inconsistent with our fixed time interval approach). The VPIN measure is thus recalculated every time a constant amount of volume is exchanged on the market. It can be noted that the literature has had some heated debates around VPIN. Andersen and Bondarenko (2014) argue that VPIN is a poor predictor of volatility and extreme returns, as its predictive power would simply come from its relationship with the underlying trading intensity. Easley et al. (2014) replied, stating that the methodology they used was different.

The first step in computing VPIN is to compute the order imbalance. Indeed, Easley et al. (2008) (cited in Easley et al., 2012) showed that

$$\mathbb{E}\left[\|VS_\tau - VB_\tau\|\right] \approx \alpha\mu,$$

and thus:

$$\text{VPIN} = \frac{\alpha\mu}{\alpha\mu + 2\varepsilon} \approx \frac{1}{nV}\sum_{\tau=1}^{n} \left|VS_\tau - VB_\tau\right|, \quad \text{where:}$$

- $\|VS_\tau - VB_\tau\|$ is the absolute difference between the buy volume and sell volume (i.e., market orders volume) in one volume bucket τ. [2]

- n is the number of buckets used to approximate trade imbalance (to be chosen).

- V is the volume of one bucket (to be chosen).

In the paper by Easly et al. (2012), Volume V is calculated as 1/50th of the average daily volume and n = 50 (meaning that the rolling window size to calculate VPIN is 50). This corresponds to a daily VPIN on a day of average volume.

This study wants to remain as non-parametric as possible, meaning that it does not want to compute variables using data in the future. However, measuring the average daily volume in order to choose V requires the entire stock dataset. Also, it has a focus on high frequency and returns are calculated on a short time frame (10 seconds). For these reasons, the probability of informed trading (PIN) will not be computed with volume sampling, but rather with standard time sampling. It will simply be calculated as the rolling 5 last minutes trade order imbalance. This approach is similar to Ranaldo (2000)'s Order Ratio approach. This approach still maintains the logic of the market volume imbalance, but does not take into account trading intensity to update the measure.

[2] Easley et al. (2012)'s paper uses probabilistic volume classification to get the expected order imbalance whereas our data enables us to have the exact order imbalance, facilitating the calculation process.

1.2.4 Liquidity Supply Variables

Liquidity supply variables are all computed by looking at the state of the limit order book at an exact instant. Together, these variables try to represent in the best possible way a clear picture of the order book. To make the most out of the LOBSTER data, which contain the 5 best levels of bid and ask prices of the order book, each variable will be computed twice: once using the first level of the order book and once using the 5 best levels of the order book.

The use of the five levels will fall into 2 categories in the most appropriate manner: variables using only the fifth level and variables using level 1 through level 5. For example, the spread will be calculated between the fifth level of the ask curve and the fifth level of the bid curve, and the depth will be measured using all five levels.

Spread. Being one of the most important concepts of microstructure theory, the spread will be part of the independent variables of this study. In its most fundamental definition, the spread is the difference in dollar between the best ask and best bid quotes of the market. To be able to compare this variable between stocks and between periods, it has to be represented as a percentage of the price. Two main solutions are presented in Cobandag et al. (2022), the percent quoted spread, which is the standard difference between ask and bid divided by the mid-quote; and the percent effective spread, being twice the absolute difference between the trade price and the mid-quote price divided by the mid-quote. Other spread-based measures exist, like the implementation shortfall and the percent price impact. This study will use the standard percent quoted spread variable as it presents the advantage of being an ex ante variable. For the sake of result readability, this variable is computed times a hundred so that it represents a percentage.

$$PQS_t = \left(\frac{P_{a,t} - P_{b,t}}{P_{m,t}}\right) \cdot 100$$

where:

- $P_{a,t}$ is the ask price at time t.
- $P_{b,t}$ is the bid price at time t.
- $P_{m,t}$ is the mid-quote price at time t.

Average Depth. The average depth (Mann & Ramanlal, 1996) (cited in Cobandag Guloglu & Ekinci, 2022) is simply the average of the depth at the best bid and best ask price, computed in our case in dollar volume.

$$AD_t = \frac{V_{a,t} + V_{b,t}}{2}$$

where:

- $V_{a,t}$ is the ask volume at the first level of the order book at time t, calculated as the number of shares times the ask price (i.e., in dollar volume).
- $V_{b,t}$ is the bid volume at the first level of the order book at time t, calculated as the number of shares times the bid price (i.e., in dollar volume).

Quote Slope. Another measure of liquidity supply, which captures both tightness and depth, is the quote slope introduced by Hasbrouck and Seppi (2001). The quote slope is, more precisely, a measure of the liquidity supply curve. The authors define it as the spread over the available quantities on the best bid and ask quotes. They use logarithmic transformations of the available quantities $V_{a,t}$ and $V_{b,t}$. They either use normal prices $P_{a,t}$ and $P_{b,t}$ (i.e., the quote slope) or log prices $\ln(P_{a,t})$ and $\ln(P_{b,t})$ (i.e., the log quote slope):

$$\text{Quote Slope}_t = \frac{P_{a,t} - P_{b,t}}{\ln(V_{a,t}) + \ln(V_{b,t})}$$

$$\text{Log Quote Slope}_t = \frac{\ln(P_{a,t} - P_{b,t})}{\ln(V_{a,t}) + \ln(V_{b,t})}$$

In this study, the standard quote slope is calculated and the quantities are expressed in dollar volume. Figure IV.2 below is a good representation of the measured liquidity supply slope. A steeper slope indicates worse liquidity, while a flatter slope suggests better liquidity. Indeed, if more quantity (i.e. more depth) is provided at the ask or bid price or if the spread narrows (i.e. more tightness), the slope becomes flatter.

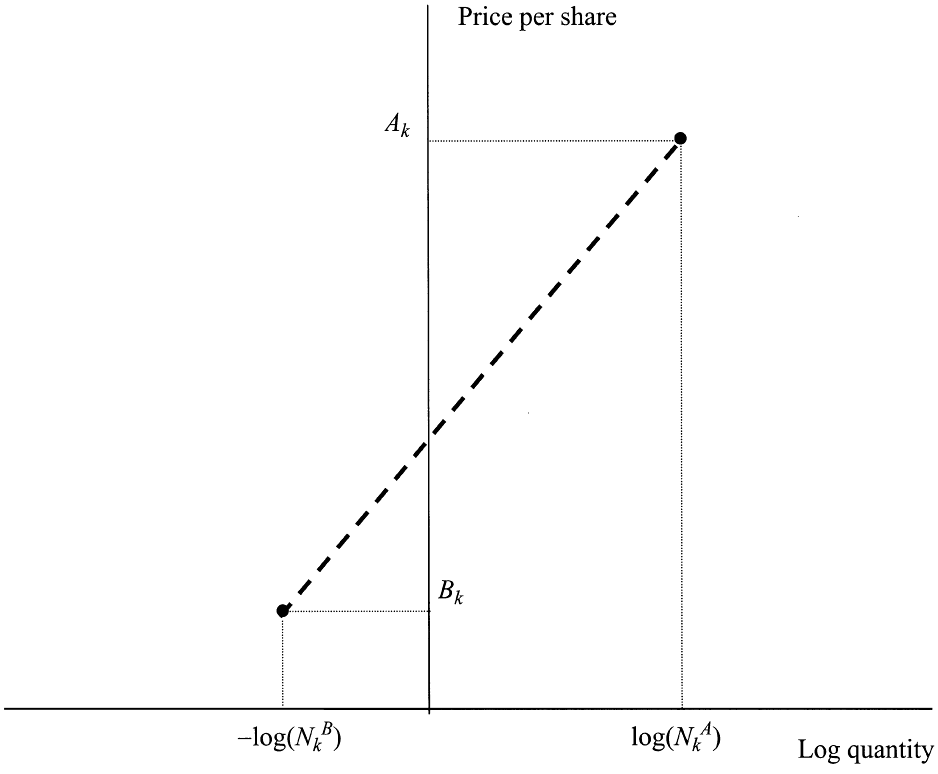


Figure IV.2: Illustration of the Quote Slope (i.e. the liquidity supply curve)
(Reprinted from Hasbrouck & Seppi, 2001, p.403)

Bi-dimensional Liquidity Measure. The Bi-dimensional liquidity measure by Pascual et al. (2004) (cited in Cobandag Guloglu & Ekinci, 2022) is bi-dimensional as it looks at both the tightness and the depth of the market. It is measured as the corrected ratio of quoted depth by time, divided by the time-weighted relative spread. For simplification purposes, this study will look at this measure as simply the average depth divided by the percent quoted spread (both measures being explained above):

$$BLM_t = \frac{AD_t}{PQS_t}$$

This measure makes perfect sense, knowing that, if either the depth goes up or the spread goes down, the BLM will go up and indicate a better liquidity supply on the market.

Limit Order Ratio. The Limit Order Ratio has not been taken from the literature, although this variable could possibly be found in the literature in some form. This measure looks at the standing limit order imbalance on the liquidity supply side of the market. More specifically, Limit Order Ratio is a measure of the LOB depth imbalance at the best bid and ask:

$$\text{LOR}_t = \frac{|V_{a,t} - V_{b,t}|}{V_{a,t} + V_{b,t}}$$

A high limit order ratio indicates that one side of the liquidity supply is stronger than the other, whereas a low limit order ratio indicates stability in the limit order book supply.

Net Dollar Limit Order Volume. The Net Dollar Limit Order Volume has also not been taken from the literature. It represents the intensity in the arrival of new limit orders inside the LOB. The name of this variable comes from the fact that it is netted from cancellations, meaning that the arrival of limit orders is calculated by the difference between the submitted and canceled limit orders, in dollar volume. Hence, this measure can take negative values when more limit orders are canceled than created. This measure will be measured considering all 5 levels of the LOB only, as data from LOBSTER do not indicate to which level of the order book a submitted or canceled order belongs to.

$$\begin{aligned} NDLOV_t &= \left(\sum_{i=t-10s}^{t} V_{a,i}^{sub} - \sum_{i=t-10s}^{t} V_{a,i}^{can} \right) + \left(\sum_{i=t-10s}^{t} V_{b,i}^{sub} - \sum_{i=t-10s}^{t} V_{b,i}^{can} \right) \\ &= \sum_{i=t-10s}^{t} V_{a,i}^{net} + \sum_{i=t-10s}^{t} V_{b,i}^{net} \end{aligned}$$

where:

- $V_{a,i}^{sub}$ (resp. $V_{a,i}^{can}$) is the dollar volume at time i that is submitted (resp. canceled) as an ask limit order to the LOB.
- $V_{b,i}^{sub}$ (resp. $V_{b,i}^{can}$) is the dollar volume at time i that is submitted (resp. canceled) as a bid limit order to the LOB.

Limit Order Flow Imbalance. The third variable that has not been taken from the literature is the Limit Order Flow Imbalance. In the same way that the PIN represents the market order imbalance looking at the liquidity demand side of the market, Limit Order Flow Imbalance will capture the imbalance in limit orders coming to the order book during the 10-second interval.

$$LOFI_t = \left| \frac{\sum_{i=t-10s}^{t} V_{a,i}^{net} - \sum_{i=t-10s}^{t} V_{b,i}^{net}}{\sum_{i=t-10s}^{t} V_{a,i}^{net} + \sum_{i=t-10s}^{t} V_{b,i}^{net}} \right|$$

1.2.5 Variables Representing Liquidity Supply and Demand

Fill Rates. Holden et al. (2014) (cited in Cobandag Guloglu & Ekinci, 2022) computes three different variables representing the breadth and immediacy component of liquidity: the partial fill rate, the complete fill rate, and the cancelation rate.

The partial (resp. complete) fill rate is the number of orders partially (resp. completely) executed on the total number of submitted orders. LOBSTER data do not identify orders as belonging to a particular individual. Thus, when observing a trade execution in the message file, no information is directly given to indicate whether the executed limit order has been executed partially or completely (e.g., if a 20 stocks order is executed, it does not tell whether the limit order from the specific individual was 20 or 100). This is the reason why this study will use a simple fill rate, defined as the number of executed orders (i.e., trades coming from market orders) over the number of submitted orders (i.e., new limit orders providing liquidity to the order book). The size of the order will also be taken into account in dollar volume:

$$\text{Fill Rate}_t = \frac{\sum_{i=t-10s}^{t} V_{buy,i} + \sum_{i=t-10s}^{t} V_{sell,i}}{\sum_{i=t-10s}^{t} V_{a,i}^{sub} + \sum_{i=t-10s}^{t} V_{b,i}^{sub}}$$

A higher fill rate indicates a worsening, while a lower fill rate indicates a strengthening of liquidity.

In a similar manner, the cancellation rate will be used and computed as follows in this study, with a higher cancellation rate representing a worsening of liquidity:

$$\text{Cancellation Rate}_t = \frac{\sum_{i=t-10s}^{t} V_{a,i}^{can} + \sum_{i=t-10s}^{t} V_{b,i}^{can}}{\sum_{i=t-10s}^{t} V_{a,i}^{sub} + \sum_{i=t-10s}^{t} V_{b,i}^{sub}}$$

It has to be noted that, for clarity purposes, both rates are presented in this section, while the cancellation rate only represents the liquidity supplying side of the market, which will be corrected in the summary table.

Kyle's Lambda. Kyle's Lambda is a key market microstructure measure that quantifies price impact, specifically the sensitivity of asset prices to order flow. Introduced by Kyle (1985), it reflects how much the price moves in response to a net buying or selling pressure over a given interval. Formally, it is estimated as the slope coefficient in a regression of price changes on signed trade volume:

$$\Delta P_t = \lambda \cdot Q_t + \varepsilon_t$$

where:

- $\Delta P_t = P_t - P_{t-1}$ is the price change over interval t.
- Q_t is the signed trade volume (positive for buy-initiated trades, negative for sell-initiated trades).
- λ or Kyle's lambda is the price impact per unit of signed volume.
- ε_t is the error term.

Kyle's Lambda will be estimated each time on a regression using the last 5 minutes of data (30 intervals). A higher lambda indicates that even small imbalances in order flow can cause significant price changes, suggesting low liquidity and a high cost of trading. Conversely, a low lambda implies that markets can absorb large trades with minimal price disturbance.

1.3 Variables Summary

Table IV.1 summarizes all the variables present in the models of this study, explaining how the variable is computed, what the variable represents, and the range of values the variable can take.

The dependent variable, market state variables, and liquidity demand variables are only represented under the "Using only 1 depth level of the LOB (when applicable)" (title of Table IV.1) but do not use any LOB bid and ask quotes data. Thus, this title is not applicable per se to them.

The dependent, market state, and liquidity demand variables, including Kyle's Lambda and excluding Time, are computed either:

- A first time during the 10-second interval, and a second time as an average of the 30 last intervals.
- Only once taking the 5 last minutes of data, without being computed as an average.

The liquidity supply variables, including the Fill Rate, also exhibit two different patterns of computation:

- A first time as a picture of the first level of the LOB at the end of the interval, and a second time as a picture of the five first LOB levels at the end of the interval.
- A first time during the 10-second interval on the five first LOB levels, and a second time as an average of the 30 last intervals still using the five first LOB levels.

In total, 26 explanatory variables will be the model inputs for the analysis of EPMs determinants.

Variable	How is this Variable Computed							What it represents	Range
	Using only 1 depth level of the LOB (when applicable)				Using the 5 depth levels of the LOB				
	End of interval (picture at an instant t)	During the 10-sec interval	With 5-min data (one 5-min interval)	With a 5-min rolling window average (30x 10-sec intervals)	End of interval (picture at instant T) on 5 LOB levels	During the 10-sec interval on 5 LOB levels	With a 5-min rolling window average (30x 10-sec intervals) on 5 LOB levels		
DEPENDENT VARIABLE									
EPM	X	V	X	X	X	X	X	Absolute Returns of the next interval above the 99,9th percentile of the distribution	0 or 1
MARKET STATE VARIABLES									
Time	V	X	X	X	X	X	X	Intraday seconds since market open	[310;23100]
Absolute Log Return	X	V	X	V	X	X	X	Absolute change of midquote price	[0, +∞[supposedly small
Median RV	X	X	V	X	X	X	X	Realized Volatility	[0,+∞[
LIQUIDITY DEMAND VARIABLES									
Trading Intensity (Dollar Market Volume)	X	V	X	V	X	X	X	Market orders Volume / Transactions Intensity	[0,+∞[
PIN	X	X	V	X	X	X	X	Market order flow imbalance	[0,1]
LIQUIDITY SUPPLY VARIABLES									
Percent Quoted Spread (PQS)	V	X	X	X	V	X	X	Liquidity supply tightness	]0,100] supposedly small
Average Depth (AD)	V	X	X	X	V	X	X	Liquidity supply depth	]0,+∞[
Quote Slope (QS)	V	X	X	X	V	X	X	Liquidity supply curve (tightness & depth)	]0,100] supposedly small
Bi-dimensional liquidity measure (BLM)	V	X	X	X	V	X	X	Average depth / Percent quoted spread	]0,+∞[
Limit Order Ratio	V	X	X	X	V	X	X	LOB's standing limit orders imbalance	[0,1]
Net Dollar Limit Order Volume	X	X	X	X	X	V	V	Limit orders Volume / Liquidity Supply Intensity	]-∞,+∞[
Limit Order Flow imbalance	X	X	X	X	X	V	V	Limit order flow imbalance	[0,1]
Cancellation Rate	X	X	X	X	X	V	V	Limit orders cancellation rate	[0,+∞[
LIQUIDITY SUPPLY & DEMAND VARIABLES									
Fill Rate	X	X	X	X	X	V	V	Limit orders execution rate	[0,+∞[
Kyle's lambda	X	X	V	X	X	X	X	Sensibility of returns (and thus liquidity supply) to market orders imbalance (liquidity demand)	Supposed to be small [0, +∞[but can be negative if liquidity supply is strong on the side with more market orders

Table IV.1: Variables Summary

2 Methods

To begin the empirical investigation, a descriptive and exploratory data analysis will be conducted in order to better understand the distributional properties, temporal patterns, and interrelationships among variables prior to model estimation. For the sake of clarity, conciseness, and efficiency, the descriptive analysis will not be performed on a stock-by-stock basis but rather on the full dataset. Then, stock-specific logistic regressions will be estimated in four variants: a market-state only GLM, a full all-variables GLM, an Elastic Net tuned by cross-validated ROC–AUC, and an unpenalized GLM on the Elastic-Net-selected subset. In-sample, unpenalized specifications are contrasted via a fit–parsimony trade-off. Out-of-sample, classification performance is evaluated using a recall-weighted cutoff, and calibration is compared pairwise via a loss-based test.

2.1 Descriptive and Exploratory Data Analysis

The descriptive analysis will first look at the frequency and distribution of EPMs. They will be summarized across stocks and over time, with a breakdown by month to explore potential seasonal clustering. While the direction of EPMs is not retained, their occurrence will be examined relative to calendar time to reveal patterns in their temporal concentration. The descriptive analysis will then focus mainly on the variables and their interactions using the following methods.

2.1.1 Kernel Density Estimates and Summary Statistics

For each variable X_j, this study will look at its unconditional density distribution (over the entire sample) and, at the same time, at its conditional distribution given that an EPM occurs over the next interval:

- The unconditional distribution $f_{X_j}(x)$ represents the density of X_j over the entire sample.
- The conditional distribution $f_{X_j|Y=1}(x)$ represents the density of X_j conditioned on $Y = 1$.

This allows to compare the general behavior of the variable to a potential behavior shift right before an EPM happens. The distribution of the variable will be estimated using a Kernel Density Estimate (KDE), a method used to estimate the real probability density function of a random variable based on a finite sample. It provides a smoothed version of the sample histogram. The KDE at a point x is given by:

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

where:

- $\hat{f}_h(x)$ is the estimated density at point x.
- n is the number of data points (by default 512 in the R `stats` package).
- h is the bandwidth, a smoothing parameter, by default Silverman's rule.
- $K(\cdot)$ is the kernel function, by default the Gaussian kernel.
- x_i are the observed data points.

To ensure better granularity, n has been set to a minimum of 65,536 (2^{16}). Some variables can even have a higher n depending on their granularity needs, particularly coming from the width of the range of values where the density is estimated.

Moreover, descriptive statistics will also be calculated for each , separately over the full sample and over the sub-sample conditioned on an EPM occurring, in the same way it has been done for the density distribution. These include mean, median, standard deviation, and interquartile ranges.

Finally, to assess the statistical significance of any distributional shifts, the following statistical tests will be applied to each variable on both the full sample and the conditioned sub-sample.

Kolmogorov-Smirnov Test. This test assesses whether two samples come from the same continuous distribution. It is sensitive to differences in location, scale, and shape. It compares the empirical cumulative distribution functions (CDFs) of both samples.

- Null hypothesis (H_0): Both samples come from the same continuous distribution, i.e., the CDFs of the variable are the same between the two samples:

$$H_0 : F_{X_j}(x) = F_{X_j|Y=1}(x), \quad \forall x \in \mathbb{R}$$

- Assumptions: Observations are independent between the two samples, and i.i.d. within each sample. The variable is continuous. No assumption on the shape of the distribution.

Mann-Whitney U Test. This non-parametric test evaluates whether the distributions of a continuous variable differ in terms of their central tendency (typically the median) between the two groups. It is applied to assess whether the distribution of a variable X_j differs in central tendency between the full sample and the sub-sample where $Y = 1$.

- Null hypothesis (H_0): The two samples come from populations with the same median:

$$P(X_j > X_{j|Y=1}) = P(X_j < X_{j|Y=1})$$

- Assumptions: Observations are independent between the two samples, and i.i.d. within each sample. The variable is at least ordinal. While not a strict assumption, it is better that the two distributions have a similar shape for the test to indicate a difference in medians and not a difference in distributions.

Levene's Test. Levene's test evaluates whether the variances of a variable are similar between two samples. It is robust against departures from normality and useful for detecting heteroskedasticity.

- Null hypothesis (H_0): The variances of the variable are equal across the two groups:

$$H_0 : \hat{\sigma}^2_{X_j} = \hat{\sigma}^2_{X_{j|Y=1}}$$

- Assumptions: Observations are independent between the two samples, and i.i.d. within each sample. The variable is continuous or interval-scaled. Normality is not required, as the test is more robust to non-normality.

It has to be noted as a limitation for those 3 tests that the two samples are not independent, as one is a sub-sample of the other. Also, the distributions of these two groups might not have similar shapes. The results of these statistical tests should therefore be interpreted as exploratory rather than confirmatory.

2.1.2 Autocorrelation and Cross-Correlation Between Variables

To explore the temporal dynamics of the variables, autocorrelation functions (ACF) will be computed for each variable over 10 lags, representing 100 seconds of market time, allowing for the detection of short-term persistence or cyclical behavior. Variables computed as a 5-minute rolling window average or on the 5 last minutes of data will not have their ACF computed. Indeed, those variables are by definition highly autocorrelated because they share a huge portion of underlying data between each lag.

Additionally, cross-correlation functions (CCF) will be computed for selected variable pairs that are theoretically linked in the microstructure literature, such as trading intensity and volatility. These will be used to examine possible lead-lag relationships, relevant for better understanding of the market microstructure dynamics in place.

Finally, a Spearman correlation matrix will be generated to assess associations between variables. Spearman correlation is chosen over Pearson as it is better suited for capturing non-linear monotonic relationships, which might be the case between the variables chosen. This approach provides a more robust measure of association in the presence of skewed distributions or outliers. Spearman's rank correlation coefficient is calculated by converting the original values into ranks and then computing the Pearson correlation between these ranks.

2.2 Regression Type Models

2.2.1 Regression Framework and Model Specification

Overview. To investigate the determinants of Extreme Price Movements (EPMs), this study adopts a regression-based framework applied on a stock-by-stock basis, a methodology in line with the way EPMs have been computed. Recognizing that microstructural dynamics might vary significantly across individual stocks, separate models are estimated for each stock independently. For every stock, the dataset is chronologically split into a training set (January to October 2020) and a test set (November to December 2020). Regressions will then be run only on the training set, the test set being left for model testing and comparison between models.

Logistic Regression. Since this study's dependent variable is a binary variable, logistic regression is the appropriate modeling choice. To understand how it functions, let's first understand how it extends from the basic linear regression model. At its core, linear regression models seek to explain a continuous dependent variable Y as a linear combination of explanatory variables. In its simplest form, the model is written as:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i$$

where:

- $\mathbf{x}_i$ is the vector of predictors for observation i.
- $\boldsymbol{\beta}$ is the vector of coefficients to estimate.
- ε_i is the error term assumed to be normally distributed with constant variance.

However, in many applications, including this study's modeling of EPMs, the dependent variable is not continuous, but binary. In such cases, linear regression is inappropriate because it can predict values outside the [0,1] interval and does not account for the Bernoulli distribution of the outcome.

This motivates the use of Generalized Linear Models (GLMs), which extend linear regression by allowing the dependent variable to follow a distribution from the univariate exponential family (e.g. Bernoulli, Poisson, or Gaussian) and by linking the expected value of the dependent variable to the predictors via a link function (whose inverse is called the response function). In the case of logistic regression, which is appropriate for binary outcomes, it is assumed that the dependent variable $Y_i \in \{0, 1\}$ follows a Bernoulli distribution with success probability $\pi_i = \mathbb{E}[Y_i] = P(Y_i = 1)$. The model estimates this probability as a function of the variables using a response function $h(\cdot)$:

$$\pi_i = h(\eta_i) = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)}$$

where:

- $\eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} = \mathbf{x}_i^\top \boldsymbol{\beta}$ is the linear predictor.
- $h(\cdot)$ is the logistic cumulative distribution function.

This formulation ensures that predicted probabilities remain between 0 and 1 and provides a principled framework for maximum likelihood estimation of the parameters. Logistic regression is thus a natural and interpretable tool for estimating the probability that an EPM occurs within a given time interval, conditional on a set of observed variables.

Standardization of Variables. Before estimating the logistic regression models, all independent variables are standardized within the training set. This transformation is critical in the context of regularized regression techniques because these models apply penalization terms directly to the regression coefficients, which are sensitive to the scale of the input variables. Standardization ensures that all variables contribute equally to the penalization process, preventing variables from being differently penalized due to their initial magnitude.

For each variable x_j, the standardized variable $\tilde{x}_j$ is computed as:

$$\tilde{x}_j = \frac{x_j - \bar{x}_j}{s_j}$$

where:

- $\bar{x}_j$ is the sample mean of variable x_j within the training set.
- s_j is the sample standard deviation of variable x_j within the training set.

All transformation parameters of every variable are then stored and reused to standardize the corresponding variables in the test set, preserving consistency. This ensures that no information from the test set is used during training and that the test set is evaluated using transformation parameters derived solely from the training data. In other words, the exact model derived from the training set will be applied to the test set.

While standardization is required for the regularized models, it is applied to the dataset for all GLMs to allow for consistent comparison across model specifications, including the non-regularized GLMs.

Model Specifications. In order to examine the main factors affecting EPMs, as specified by the research aim with a specific focus on liquidity, four different logistic regression models are estimated for every stock. The approach builds on a first constrained model, consisting of solely the market state variables, forming a benchmark. The accuracy of models with the addition of liquidity variables is compared to this benchmark, so we can determine the incremental contribution of liquidity to the prediction of the occurrence of EPMs. The four models are specified as follows:

1. Restricted GLM (Market State Variables, Unpenalized):
   As mentioned above, this specification act as a benchmark for other specifications. The logistic regression of this model will only take the market state variables into account as explanatory variables, namely: Time, Absolute Return (in the normal and average form), Trading Intensity (in the normal and average form), Median Realized Volatility. It excludes all liquidity-related variables, allowing a focused analysis of how much EPM predictability can be attributed to general market dynamics alone.

2. Full GLM (All Variables, Unpenalized):
   A logistic regression including all available variables: market state variables, liquidity supply and demand measures, and hybrid indicators. This model serves as the most comprehensive benchmark, capturing many dimensions of microstructure information.

3. Elastic Net Logistic Regression (Penalized Variable Selection):
   To address multi-collinearity and perform automatic variable selection, an Elastic Net-penalized logistic regression is applied. This method estimates the regression coefficients by minimizing a penalized version of the negative log-likelihood of the logistic model.
   The Elastic Net estimator is defined as:

$$\widehat{\boldsymbol{\beta}}_{\text{EN}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \left\{ -\frac{1}{n}\sum_{i=1}^{n} \log\left(f(\boldsymbol{\beta}(x_i); y_i)\right) + \lambda\left(\alpha\|\boldsymbol{\beta}\|_1 + \frac{1-\alpha}{2}\|\boldsymbol{\beta}\|_2^2\right)\right\}$$

   where:

   - $f(\boldsymbol{\beta}(x_i); y_i)$ is the probability of observing y_i given the variables x_i and model parameters $\boldsymbol{\beta}$, under the logistic model.
   - $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p} |\beta_j|$ is the LASSO penalty.
   - $\|\boldsymbol{\beta}\|_2^2 = \sum_{j=1}^{p} \beta_j^2$ is the Ridge penalty.
   - $\lambda \geq 0$ is the regularization parameter controlling the overall penalty strength.

- $\alpha \in [0, 1]$ is the mixing parameter:
  - $\alpha = 1$: pure LASSO.
  - $\alpha = 0$: pure Ridge.
  - $0 < \alpha < 1$: Elastic Net (a balance of both).

The goal is to find the coefficient vector $\widehat{\boldsymbol{\beta}}_{\mathrm{EN}}$ that minimizes the penalized loss, balancing model fit (via the log-likelihood) and regularization (via the combined penalty).

To select the most predictive and robust Elastic Net regression model, a specific procedure is put in place. This involves tuning the two main hyperparameters: *alpha*, which balances LASSO and Ridge penalties, and *lambda*, which controls the overall strength of regularization. The selection procedure begins by defining a sequence of alpha values, ranging from 0 (Ridge) to 1 (LASSO) in increments of 0.1.

Once a specific value of alpha is fixed, the model selection procedure focuses on identifying the optimal regularization strength (lambda) for that penalty configuration. A range of lambda values is tested using k-fold cross-validation. In each fold, the model is trained on a portion of the data and evaluated on the remaining part by predicting class probabilities. The performance of each lambda is assessed using the area under the Receiver Operating Characteristic curve (AUC of the ROC). The AUC values are averaged across the k folds for each lambda, and the lambda with the highest mean AUC is selected as optimal for the given alpha. Once all alpha values have been evaluated, we select the one that led to the highest AUC overall, ultimately selecting the optimal combination of alpha and lambda. A final Elastic Net regression is applied with this optimal hyperparameter combination.

The choice of using AUC as the performance metric for lambda selection, instead of a more traditional mean-squared or deviance error measure (all those measure types are available in the `cv.glmnet` function from the R `glmnet` package, with deviance being the default method), is deliberate. AUC of the ROC compares the true positive rate (TPR) and false positive rate (FPR) across all possible thresholds, focusing (partially) on how well the model detects the positive class ($Y = 1$), which makes it more suitable in this study's highly imbalanced setting where positives (EPMs) occur only once in a thousand cases.

4. <u>Selected-Variable GLM (Post-Elastic Net Variable Selection, Unpenalized)</u>:
   The fourth and final model is a standard (unpenalized) logistic regression. The regression coefficients from the final Elastic Net regression model described above are extracted, and only the variables associated with non-zero coefficients are kept for this model. In other words, this regression a simple unpenalized logistic regression made on a subset of variables "selected by the Elastic Net procedure". This approach combines the benefits of regularized variable selection with the interpretability of standard GLM estimation, making it more suitable for direct comparison with the two other unpenalized regressions.

These four models provide a structured way to compare penalized and unpenalized frameworks, and to assess the respective contributions of liquidity and market state variables in predicting EPMs.

2.2.2 Regression Evaluation and Comparison on the Training Set

This sub-section presents the framework used to evaluate the four regression models on the in-sample results from the training set. It will explain how to interpret coefficients from standardized predictors and how to compare the three unpenalized models with the BIC, keeping in mind that the aim is to identify the extent to which liquidity variables influence EPMs, as well as which ones most influence the likelihood of EPMs occurring.

Interpreting the logistic regression with standardized variables. As explained in section 2.2.1 *Regression Framework and Model Specification*, all predictors were standardized on the training set to have a mean of 0 and standard deviation of 1. Coefficients therefore measure the effect of a one–standard deviation change in a variable, which is an important nuance compared to the usual one-unit change. Furthermore, this makes the magnitudes comparable across variables.

With standardized predictors, a coefficient β_j represents the change in log-odds of an EPM for a one-standard deviation increase in x_j. Exponentiating β_j gives the odds ratio

$$\mathrm{OR}_j = \exp(\beta_j),$$

which is the multiplicative factor applied to the odds when x_j rises by one standard deviation. For example, OR = 1.20 means that the odds are 20% higher and OR = 0.80 means that the odds are 20% lower.

In the coefficient summary tables of the unpenalized regression models, statistical significance will be summarized with significance stars placed next to each coefficient estimate. These stars correspond to the p-value from a Wald test of the null $H_0 : \beta_j = 0$ in the specific model. The p-value corresponds to the probability, if the true effect were zero, of obtaining an estimate at least as extreme as the one observed; more stars mean a smaller p-value, and thus stronger evidence against H_0. Only stars will be reported for readability purposes. These stars will be treated with caution: they support the reading of coefficient signs and magnitudes but do not, by themselves, determine the importance of the regressor for the objectives of this study.

Bayesian Information Criterion. For comparability and selection across unpenalized logistic specifications, this study reports the Bayesian Information Criterion (BIC):

$$\mathrm{BIC} = -2 \log \hat{\mathcal{L}} + k \log n,$$

where $\hat{\mathcal{L}}$ is the maximized likelihood on the estimation sample (the training set), k is the number of freely estimated parameters (including the intercept), and n is the number of training observations. BIC summarizes a trade-off between in-sample fit (the log-likelihood term) and parsimony (the complexity penalty $k \log n$), and is interpreted here as an in-sample diagnostic to compare alternative unpenalized GLMs (i.e., the Restricted, Full, and post–Elastic Net Selected-Variable GLM). BIC is not applied to the penalized Elastic Net regression model itself because the Elastic Net regression was already maximizing a penalized version of the log-likelihood, making comparison not possible between this model and the three others.

Because its definition relies on the maximized likelihood from the estimation sample, BIC is computed on the training data; with a large n (around 472,530 per stock), the $k \log n$ penalty is substantial, so BIC will often favor sparser specifications even when richer models yield slightly higher out-of-sample discrimination. Compared to the Akaike Information Criterion (AIC) whose penalty is $2k$, the BIC wants an even better maximized likelihood in the presence of more variables if the training sample size n is larger. Results are presented per stock.

2.2.3 Regression Evaluation and Comparison on the Test Set

Assessing the models on the test set is the most important part of model selection, in order to evaluate the importance and contribution of liquidity variables to the predictive ability of the four models in predicting EPMs. This can be done particularly by evaluating the out-of-sample performance of the three models that include liquidity variables, in comparison with the benchmark market-state-variables-only model.

This subsection will first look at how performance is evaluated without any threshold. To recall, the four logistic regression models estimate the probability $\pi_i = \mathbb{E}[Y_i] = \mathbb{P}(Y_i = 1)$, i.e. the probability that the next interval will be an EPM. Threshold-free evaluation will thus evaluate models without choosing a particular probability threshold above which the final prediction will be 1 (positive) and under which the final prediction will be 0 (negative).

In a second step, an optimal and economically sound method to choose a threshold in order to have confusion matrices will be presented. Then, multiple metrics of interest will be calculated on those confusion matrices. Finally, the Diebold–Mariano test is applied to each pair of models to assess which model outperforms the others in terms of forecasted probabilities calibration.

Threshold-free discrimination performance. Out-of-sample performance is first assessed without applying a classification threshold. For each regression and stock, the Receiver Operating Characteristic area under the curve (ROC–AUC) and the Precision–Recall area under the curve (PR–AUC) are computed on the test set. The ROC curve plots the True Positive Rate

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

against the False Positive Rate

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}},$$

measuring how well predicted probabilities rank EPM vs. non-EPM cases over all thresholds.

The PR curve plots

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad \text{against} \quad \text{Recall} = \text{TPR},$$

capturing the trade-off between correctly identifying positives and avoiding false alarms. While ROC-AUC is stable and widely used, it can overstate performance under the extreme imbalance of this study. Thus, PR-AUC is more informative, as it directly reflects the challenge of detecting rare EPMs with high precision.

Threshold selection on the training set. To report the metrics of the confusion matrix, a probability threshold must be first chosen. For each stock and regression, a single cutoff $\tau^\star \in [0,1]$ will be fixed using the training predictions only, and then carried unchanged to the test set. The choice of $\tau^\star$ is guided by Precision–Recall trade-offs, which are central under class imbalance.

Two "usual" composite summaries of Precision and Recall are:

- the harmonic mean (the classic F_1 score),
$$F_1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}},$$
- the geometric mean (the Fowlkes–Mallows index),
$$\text{FM} = \sqrt{\text{Precision} \cdot \text{Recall}}.$$

Within a risk management framework, the cost of missing an EPM would typically be much higher than the cost of triggering a false alarm. The latter might cause unnecessary hedges or costs, whereas the former can result in substantial unhedged losses or missed profit opportunities. For this reason, the recall-weighted F_β family is used and the metric F_2 is calculated for each threshold on the training set. In general,

$$F_\beta = \frac{(1+\beta^2)\,\text{Precision} \cdot \text{Recall}}{\beta^2\,\text{Precision} + \text{Recall}},$$

with F_1 being the special case where $\beta = 1$. Setting $\beta = 2$ weights Recall four times more than Precision ($\beta^2 = 4$), resulting in

$$F_2 = \frac{5\,\text{Precision} \cdot \text{Recall}}{4\,\text{Precision} + \text{Recall}}.$$

After having computed F_2 over each threshold on the training set, the optimal threshold that maximizes F_2 is chosen:

$$\tau^\star = \arg\max_{\tau \in [0,1]} F_2(\tau).$$

This produces a single cutoff per stock and regression, computed on the training set and then applied as-is on the test set to form confusion matrices, ensuring clean out-of-sample threshold evaluation.

Performance metrics computation on the confusion matrix. With the threshold $\tau^\star$ computed from the training set, test set probabilities are converted to class labels and a confusion matrix is formed per stock and model, having a count for TP (true positives), FN (false negatives), FP (false positives), and TN (true negatives). The following metrics are then computed:

- Recall (Sensitivity, TPR):
$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

- Precision (PPV):
$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$
- Specificity (TNR):
$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$
- F_1 score:
$$F_1 = \frac{2\,\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$
- F_2 score:
$$F_2 = \frac{5\,\text{Precision} \cdot \text{Recall}}{4\,\text{Precision} + \text{Recall}}$$
- Fowlkes–Mallows (FM):
$$\text{FM} = \sqrt{\text{Precision} \cdot \text{Recall}}$$
- Balanced Accuracy:
$$\text{BA} = \tfrac{1}{2}\big(\text{Recall} + \text{Specificity}\big)$$

FPR and FNR are omitted as they are redundant (each equals 1 minus a metric already reported). Plain accuracy is also omitted due to the extreme class imbalance.

To summarize performance across stocks, support-weighted averages are used, with s being stock indexes:

- Recall (TPR): weight by event support $(\text{TP} + \text{FN})$:
$$\overline{\text{Recall}}_w = \frac{\sum_s \text{TP}_s}{\sum_s (\text{TP}_s + \text{FN}_s)}$$
- Specificity (TNR): weight by non-event support $(\text{TN} + \text{FP})$:
$$\overline{\text{Spec}}_w = \frac{\sum_s \text{TN}_s}{\sum_s (\text{TN}_s + \text{FP}_s)}$$
- Precision (PPV): weight by predicted-positive support $(\text{TP} + \text{FP})$:
$$\overline{\text{Prec}}_w = \frac{\sum_s \text{TP}_s}{\sum_s (\text{TP}_s + \text{FP}_s)}$$
- Balanced Accuracy: combine the two weighted components:
$$\overline{\text{BA}} = \tfrac{1}{2}\left(\overline{\text{Recall}}_w + \overline{\text{Spec}}_w\right)$$
- F_1, F_2, FM: the support-weighted average versions of Precision and Recall are used and plugged into the usual F_1, F_2, and FM formulas.

Forecast comparison test. To compare the out-of-sample predictive quality of the four logistic regressions, the Diebold–Mariano (Diebold & Mariano, 2002) test is applied to each pair of models and for each stock. The test checks whether two forecasting models have the same predictive accuracy, by checking the two models' forecast errors.

The Diebold-Mariano test does not usually take as input two vectors of errors as-is; it takes a loss function of the forecast errors. Common loss functions are the squared error $L(e) = e^2$ or the absolute error $|e|$. In the case of this study, the four models produce probabilities for a binary event (EPM or no-EPM), for which a proper loss function is chosen: the Log-Loss. Under the heavy imbalance of the dataset in this study, the Log-loss function will strongly penalize overconfident no-EPM predictions that miss rare positives.

For each stock s and model m, let $\{p_{s,t}^{(m)}\}_{t=1}^{T_s}$ denote the predicted probabilities that the next 10-second interval is an EPM, and $\{y_{s,t}\}_{t=1}^{T_s} \in \{0, 1\}$ the realized labels. The log-loss for each timestamp is defined as:

$$\ell_{s,t}^{(m)} = -\Big(y_{s,t} \log p_{s,t}^{(m)} + (1 - y_{s,t}) \log \big(1 - p_{s,t}^{(m)}\big)\Big),$$

and, for a model pair (m_1, m_2), the loss-difference series is

$$d_{s,t}^{(m_1,m_2)} = \ell_{s,t}^{(m_1)} - \ell_{s,t}^{(m_2)} \qquad \text{for } t = 1 \text{ to } t = T_s.$$

For each timestamp, a positive (resp. negative) value of the loss-difference series indicates a higher (resp. lower) loss and worse (resp. better) performance for the first model.

After having computed the log-loss and the loss-difference series on our data, the Diebold-Mariano test evaluates whether the two forecast models have an equal expected loss (H_0) or a different expected loss (H_1):

$$H_0 : \mathbb{E}\Big[d_{s,t}^{(m_1,m_2)}\Big] = 0 \quad \text{vs} \quad H_1 : \mathbb{E}\Big[d_{s,t}^{(m_1,m_2)}\Big] \neq 0.$$

The test statistic is given by:

$$\text{DM}_s = \frac{\bar{d}_s}{\sqrt{\widehat{S}_s(0)/T_s}} \qquad \text{with} \qquad \bar{d}_s = \frac{1}{T_s} \sum_{t=1}^{T_s} d_{s,t}^{(m_1,m_2)},$$

where $\widehat{S}_s(0)$ is the long-run variance (using a HAC/Newey–West estimator) and T_s is the forecast sample size. With one-step-ahead probabilities, DM_s is asymptotically standard normal under H_0.

On each stock and each pair of models, the mean $\bar{d}_s^{(m_1,m_2)}$ of the loss-difference series $d_{s,t}^{(m_1,m_2)}$ is reported, along with the Diebold-Mariano test p-values, indicated by significance stars placed next to each mean loss-difference. The interpretation is the same as the one explained above for the individual values of the loss-difference series: along with a significant p-value, a positive (resp. negative) mean loss-difference represents a better predictive power and calibration of probabilities for the second (resp. first) model.

Chapter V

Empirical Results and Analysis

1 Descriptive and Exploratory Data Analysis

1.1 EPM Month-by-Month Distribution

Table V.1 presents the monthly distribution of Extreme Price Movements (EPMs) for each of the 10 (11 if we interpret Google as two different stocks) selected large-cap NASDAQ stocks during the year 2020. The frequency of EPMs exhibits a clear temporal concentration, with over 74% of all EPMs occurring in March alone. This is consistent across all stocks, indicating a systematic market-wide phenomenon rather than stock-specific anomalies. This sharp spike in EPMs aligns closely with the exceptional surge in market volatility captured by the CBOE Volatility Index (VIX), shown in Figure V.1. The VIX reached its peak in March 2020, surpassing 80, reflecting extreme investor uncertainty in response to the COVID-19 outbreak and its financial implications.

Following March, the incidence dropped, stabilizing at much lower levels for the rest of the year. Interestingly, a smaller secondary peak appears in September, which corresponds to a mild rebound in the VIX. This visual and statistical co-movement between EPMs and the VIX supports the notion that EPMs are closely tied to periods of elevated market volatility and macroeconomic uncertainty. Overall, it reinforces the need to include volatility-related variables in any model aiming to predict extreme price behavior, as it was done with the MedRV.

An important implication of this distributional asymmetry is that EPMs are not evenly spread across time. Specifically, for some stocks, the two final months of 2020, which are used as a test set in this study, contain very few or even no EPMs. This implies that the evaluation of predictive models on the test set must be interpreted with caution, as the absence of EPMs for certain stocks may affect the reliability and comparability of out-of-sample performance metrics.

Ticker	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Sum
AAPL	0	45	352	11	3	4	3	6	128	7	3	2	564
AMZN	4	28	323	44	13	5	86	3	36	3	18	1	564
AVGO	0	1	511	31	5	5	0	0	9	1	0	1	564
COST	1	13	520	20	2	0	1	1	4	1	1	0	564
FB	2	9	448	26	7	18	3	11	24	5	3	8	564
GOOG	0	9	483	25	10	5	6	0	12	7	7	0	564
GOOGL	0	7	494	21	6	4	4	0	13	8	7	0	564
MSFT	0	23	498	18	5	2	1	3	13	1	0	0	564
NFLX	7	13	359	63	10	6	58	2	22	13	11	0	564
NVDA	0	30	396	18	8	5	8	3	91	1	4	0	564
TSLA	0	95	237	13	14	2	68	15	101	0	5	14	564
Sum	14	273	4621	290	83	56	238	44	453	47	59	26	6204

Table V.1: EPMs Month-by-Month Occurrence Across Stocks

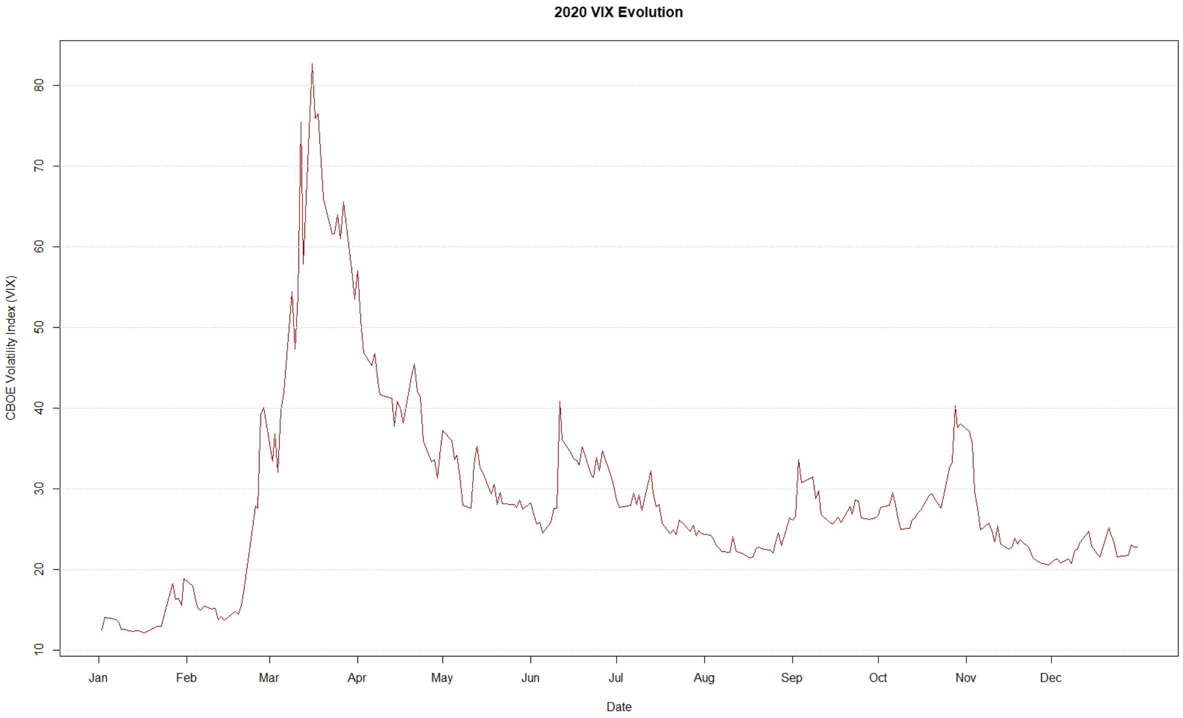


Figure V.1: CBOE Volatility Index Evolution During the Year 2020

1.2 Kernel Density Estimates and Summary Statistics

This subsection presents, for each variable, its summary statistics over the full sample and the sub-sample conditioned on an EPM occurring in the next interval ($Y = 1$). The following statistics are represented:

- Location and dispersion: minimum, 25th percentile, median, mean, 75th percentile, maximum, standard deviation.
- Distribution shape: skewness and kurtosis.
- Distributional tests between both distributions:
    - Kolmogorov–Smirnov test (test statistic D, showing the largest vertical difference in the empirical CDFs of both distributions, and p-value)
    - Mann–Whitney U test (p-value)
    - Levene's test (p-value)

These numerical results are complemented by Kernel Density Estimate (KDE) plots, showing both unconditional distributions and distributions conditional on an EPM happening in the following interval. To facilitate visual comparison between variables, the KDE plots have been uniformly truncated on both tails to include exactly 99% of the density area. This approach improves readability while preserving the relative differences in tail thickness, allowing for a clearer assessment of the extent to which certain variables exhibit heavier tails than others.

1.2.1 Market State Variables and Liquidity Demand Variables

Table V.2 and Figure V.2 display respectively the summary statistics and KDE plots of all variables belonging to Market State or Liquidity Demand (with Trading Intensity, also called Dollar Market Volume, being part of both categories). Their analysis reveals several systematic differences in the period preceding EPMs compared to the overall sample.

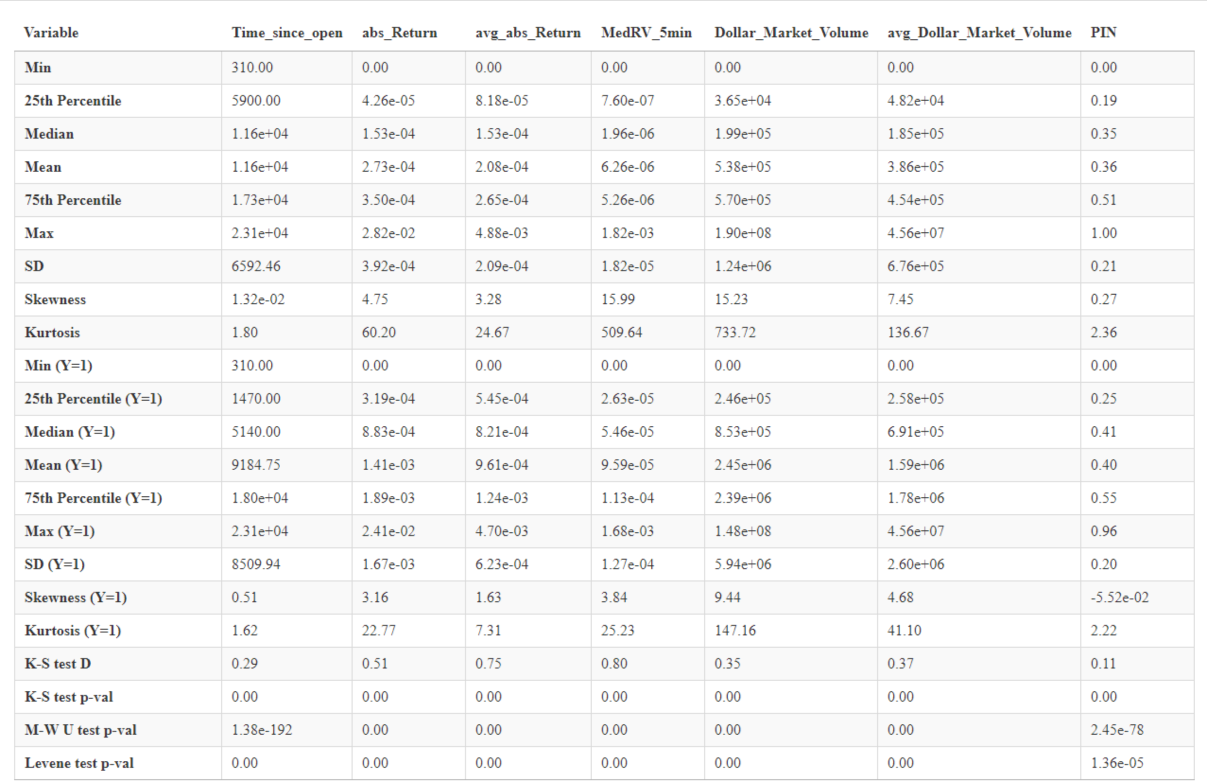
The time-of-day measure (`Time_since_open`) shows a pronounced clustering of EPMs near the market open and close, in line with intraday seasonality patterns seen in the literature review (Brogaard et al., 2018; McInish and Wood, 1992). The distribution median falls from 11,600 to 5,140 seconds in the conditioned sub-sample.

Measures of return volatility, like `abs_Return` and its smoothed counterpart `avg_abs_Return`, display substantial increases in central tendency before EPMs, with medians being five to six times higher than in the full sample. The realized volatility metric `MedRV_5min` exhibits an even more pronounced shift, with the median increasing by nearly ten times, and the entire conditional distribution moving rightward in the KDE plots. These changes indicate that EPMs tend to occur in periods of heightened volatility.

Trading activity, as measured by `Dollar_Market_Volume` and `avg_Dollar_Market_Volume`, also rises sharply in the moments preceding EPMs. The median dollar volume is more than quadrupled in the conditional sample, and the KDEs reveal both a general rightward shift and heavier tails, indicating that extreme volume spikes are more frequent. This aligns with the market microstructure literature suggesting that trading intensity is positively correlated with absolute returns (Karpoff, 1987; Gallant et al., 1992) (both cited in Hasbrouck & Seppi, 2001).

Interestingly, the `PIN` (Probability of Informed Trading) measure only displays a modest increase before EPMs. It remains consistent with the idea that informed trading pressure may build ahead of large price changes. Although the magnitude of the shift in `PIN` is smaller than for volatility and volume measures, it remains statistically significant (as highlighted by the 3 tests) and directionally consistent. However, it can be noted that the choice of computing the simple `PIN` measure instead of its `VPIN` counterpart might have had an effect on the magnitude of the influence of this measure on potential EPMs.

For all variables in this group, the Kolmogorov-Smirnov, Mann-Whitney U, and Levene tests reject the null hypotheses of equal distributions, medians, or variances at conventional significance levels, confirming that observed shifts are statistically robust. Moreover, all variables show significant D statistics for the Kolmogorov–Smirnov test, especially the `MedRV_5min` with a D statistic of 0,8. Overall, the evidence paints a coherent and literature-backed picture: EPMs are most likely to emerge during high-volatility, high-volume periods clustered around the start and end of the trading day, with a modest increase in signals of informed trading.



Variable	Time_since_open	abs_Return	avg_abs_Return	MedRV_5min	Dollar_Market_Volume	avg_Dollar_Market_Volume	PIN
Min	310.00	0.00	0.00	0.00	0.00	0.00	0.00
25th Percentile	5900.00	4.26e-05	8.18e-05	7.60e-07	3.65e+04	4.82e+04	0.19
Median	1.16e+04	1.53e-04	1.53e-04	1.96e-06	1.99e+05	1.85e+05	0.35
Mean	1.16e+04	2.73e-04	2.08e-04	6.26e-06	5.38e+05	3.86e+05	0.36
75th Percentile	1.73e+04	3.50e-04	2.65e-04	5.26e-06	5.70e+05	4.54e+05	0.51
Max	2.31e+04	2.82e-02	4.88e-03	1.82e-03	1.90e+08	4.56e+07	1.00
SD	6592.46	3.92e-04	2.09e-04	1.82e-05	1.24e+06	6.76e+05	0.21
Skewness	1.32e-02	4.75	3.28	15.99	15.23	7.45	0.27
Kurtosis	1.80	60.20	24.67	509.64	733.72	136.67	2.36
Min (Y=1)	310.00	0.00	0.00	0.00	0.00	0.00	0.00
25th Percentile (Y=1)	1470.00	3.19e-04	5.45e-04	2.63e-05	2.46e+05	2.58e+05	0.25
Median (Y=1)	5140.00	8.83e-04	8.21e-04	5.46e-05	8.53e+05	6.91e+05	0.41
Mean (Y=1)	9184.75	1.41e-03	9.61e-04	9.59e-05	2.45e+06	1.59e+06	0.40
75th Percentile (Y=1)	1.80e+04	1.89e-03	1.24e-03	1.13e-04	2.39e+06	1.78e+06	0.55
Max (Y=1)	2.31e+04	2.41e-02	4.70e-03	1.68e-03	1.48e+08	4.56e+07	0.96
SD (Y=1)	8509.94	1.67e-03	6.23e-04	1.27e-04	5.94e+06	2.60e+06	0.20
Skewness (Y=1)	0.51	3.16	1.63	3.84	9.44	4.68	-5.52e-02
Kurtosis (Y=1)	1.62	22.77	7.31	25.23	147.16	41.10	2.22
K-S test D	0.29	0.51	0.75	0.80	0.35	0.37	0.11
K-S test p-val	0.00	0.00	0.00	0.00	0.00	0.00	0.00
M-W U test p-val	1.38e-192	0.00	0.00	0.00	0.00	0.00	2.45e-78
Levene test p-val	0.00	0.00	0.00	0.00	0.00	0.00	1.36e-05

Table V.2: Summary Statistics of Market State Variables and Liquidity Demand Variables

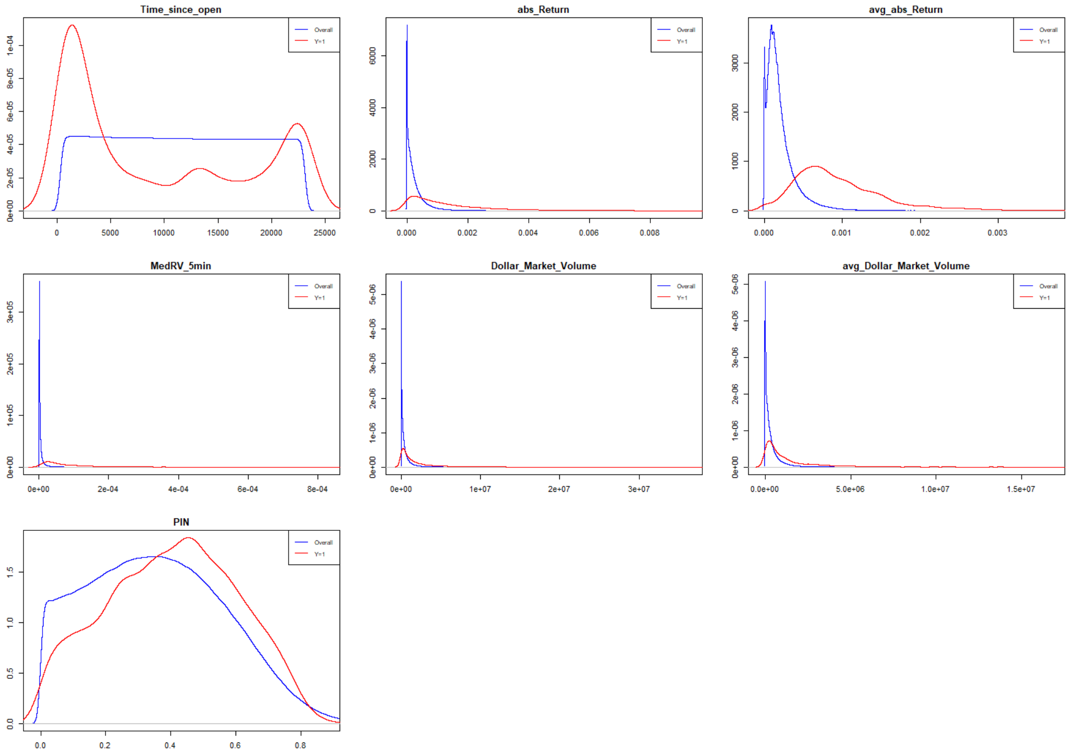


Figure V.2: KDE of Market State Variables and Liquidity Demand Variables

1.2.2 Liquidity Supply Variables Representing a Picture of the LOB

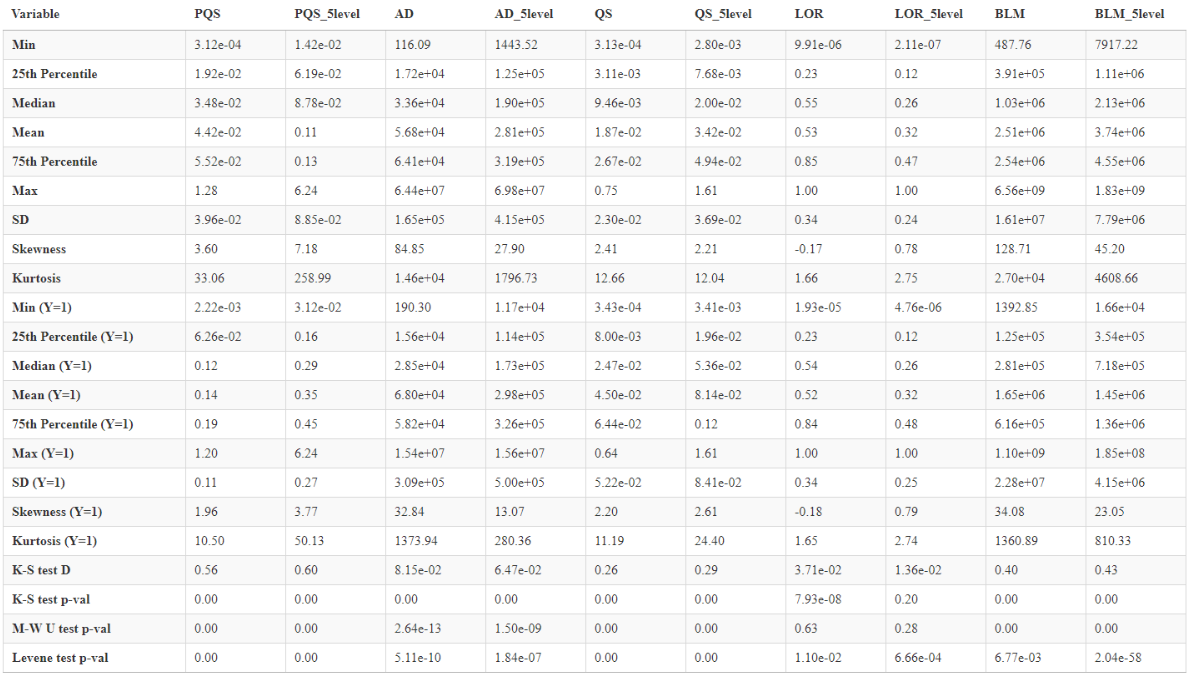
The liquidity supply variables capturing a snapshot of the LOB at a given instant (`PQS`, `PQS_5level`, `AD`, `AD_5level`, `QS`, `QS_5level`, `LOR`, `LOR_5level`, `BLM`, `BLM_5level`) show a less strong evidence of distributional differences between the unconditional sample and the sample conditioned on $Y = 1$, as shown through the KDEs of Figure V.3, even though Table V.3 shows Kolmogorov–Smirnov, Mann–Whitney U, and Levene's tests p-values that should confirm a distributional shift.

First, the Percent Quoted Spread, both with its standard measure and with the measure of the spread on 5 LOB levels, highlights that EPMs tend to happen wat more often when spreads are high. Indeed, the conditional distributions for $Y = 1$ display higher median and mean values compared to the unconditional sample, indicating that EPMs are associated with wider spreads both at the best level and across five levels. The KDE plots also show a rightward shift in mass, suggesting reduced liquidity in terms of cost to trade. This is consistent with literature findings that liquidity deteriorates ahead of EPMs, as wider spreads increase transaction costs and, most importantly, reflect higher adverse selection risk.

However, Average Depth, measured by `AD` and `AD_5level`, show almost no shift in their KDE and central tendency measures. At first glance, depth itself does not appear to be an important factor in the appearance of EPMs. Similarly, the depth imbalance, represented by `LOR` and `LOR_5level`, shows no sign of a distributional shift, with the highest values of the Kolmogorov–Smirnov test and Mann–Whitney U test throughout the whole variable set. A possible explanation for the non-shift of Average Depth could have been that liquidity supply builds up on the side of the order book not experiencing an EPM right before an EPM happens, balancing out $V_{a,t}$ and $V_{b,t}$. Unfortunately, the fact that the depth imbalance measures show no distributional shift either eliminates this theory.

Finally, the Quote Slope (`QS` and `QS_5level`) and Bi-dimensional Liquidity Measure (`BLM` and `BLM_5level`) summary statistics and KDEs show a large distributional shift, both in opposite directions. The Quote Slope follows a pattern similar to the Percent Quoted Spread variables, which is logical since the numerator of the Quote Slope is the spread $P_{a,t} - P_{b,t}$. The KDEs of `QS` and `QS_5level` display higher conditional medians and a fatter right tail, suggesting a steeper slope (less liquidity supply) before EPMs. The Bi-dimensional Liquidity Measure, measuring average depth on the percent quoted spread, has its conditional distribution shifted to the left, indicating a smaller `BLM` before EPMs. Again, this distributional shift makes perfect sense and indicates a worse liquidity before EPMs.

In general, the percent quoted spread, at the best quote and across five levels, shows the clearest pre-EPM signal, with right-shifted conditional densities and higher central moments under $Y = 1$. In contrast, Average Depth and the depth imbalance (`LOR`) measures show weak shifts, with heavy overlap between the distributions over the full sample and conditioned on $Y = 1$. Taken together, these results suggest that price-based tightness seems to matter more than depth or depth imbalance for the probability of EPMs. More interestingly, Kolmogorov–Smirnov tests' D statistics on `PQS`, `QS` and `BLM` are always higher on the variable computed on 5 LOB levels, emphasizing the meaningful nature of these variables, computed with data passing through the order book.



Variable	PQS	PQS_5level	AD	AD_5level	QS	QS_5level	LOR	LOR_5level	BLM	BLM_5level
Min	3.12e-04	1.42e-02	116.09	1443.52	3.13e-04	2.80e-03	9.91e-06	2.11e-07	487.76	7917.22
25th Percentile	1.92e-02	6.19e-02	1.72e+04	1.25e+05	3.11e-03	7.68e-03	0.23	0.12	3.91e+05	1.11e+06
Median	3.48e-02	8.78e-02	3.36e+04	1.90e+05	9.46e-03	2.00e-02	0.55	0.26	1.03e+06	2.13e+06
Mean	4.42e-02	0.11	5.68e+04	2.81e+05	1.87e-02	3.42e-02	0.53	0.32	2.51e+06	3.74e+06
75th Percentile	5.52e-02	0.13	6.41e+04	3.19e+05	2.67e-02	4.94e-02	0.85	0.47	2.54e+06	4.55e+06
Max	1.28	6.24	6.44e+07	6.98e+07	0.75	1.61	1.00	1.00	6.56e+09	1.83e+09
SD	3.96e-02	8.85e-02	1.65e+05	4.15e+05	2.30e-02	3.69e-02	0.34	0.24	1.61e+07	7.79e+06
Skewness	3.60	7.18	84.85	27.90	2.41	2.21	-0.17	0.78	128.71	45.20
Kurtosis	33.06	258.99	1.46e+04	1796.73	12.66	12.04	1.66	2.75	2.70e+04	4608.66
Min (Y=1)	2.22e-03	3.12e-02	190.30	1.17e+04	3.43e-04	3.41e-03	1.93e-05	4.76e-06	1392.85	1.66e+04
25th Percentile (Y=1)	6.26e-02	0.16	1.56e+04	1.14e+05	8.00e-03	1.96e-02	0.23	0.12	1.25e+05	3.54e+05
Median (Y=1)	0.12	0.29	2.85e+04	1.73e+05	2.47e-02	5.36e-02	0.54	0.26	2.81e+05	7.18e+05
Mean (Y=1)	0.14	0.35	6.80e+04	2.98e+05	4.50e-02	8.14e-02	0.52	0.32	1.65e+06	1.45e+06
75th Percentile (Y=1)	0.19	0.45	5.82e+04	3.26e+05	6.44e-02	0.12	0.84	0.48	6.16e+05	1.36e+06
Max (Y=1)	1.20	6.24	1.54e+07	1.56e+07	0.64	1.61	1.00	1.00	1.10e+09	1.85e+08
SD (Y=1)	0.11	0.27	3.09e+05	5.00e+05	5.22e-02	8.41e-02	0.34	0.25	2.28e+07	4.15e+06
Skewness (Y=1)	1.96	3.77	32.84	13.07	2.20	2.61	-0.18	0.79	34.08	23.05
Kurtosis (Y=1)	10.50	50.13	1373.94	280.36	11.19	24.40	1.65	2.74	1360.89	810.33
K-S test D	0.56	0.60	8.15e-02	6.47e-02	0.26	0.29	3.71e-02	1.36e-02	0.40	0.43
K-S test p-val	0.00	0.00	0.00	0.00	0.00	0.00	7.93e-08	0.20	0.00	0.00
M-W U test p-val	0.00	0.00	2.64e-13	1.50e-09	0.00	0.00	0.63	0.28	0.00	0.00
Levene test p-val	0.00	0.00	5.11e-10	1.84e-07	0.00	0.00	1.10e-02	6.66e-04	6.77e-03	2.04e-58

Table V.3: Summary Statistics of Liquidity Supply Variables
that Represent a Picture of the LOB at a Fixed Instant

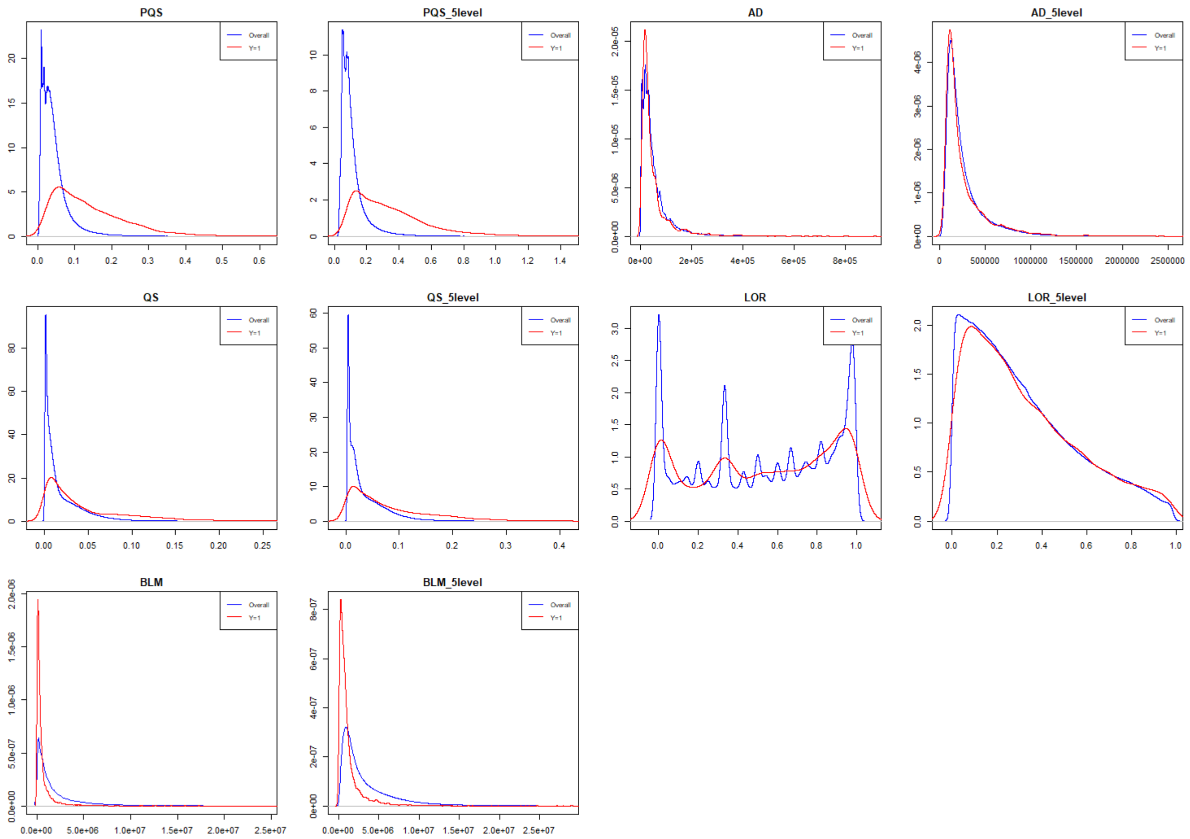


Figure V.3: KDE of Liquidity Supply Variables
that Represent a Picture of the LOB at a Fixed Instant

1.2.3 Liquidity Supply Variables Representing the Evolution of the LOB and Liquidity Supply and Demand Variables

Table V.4 and Figure V.4 report summary statistics and KDEs for variables that track how the LOB evolves over a 10-second interval (`NDLOV`, `avg_NDLOV`, `LOFI`, `avg_LOFI`, `CR`, `avg_CR`), as well as variables that combine liquidity supply with demand (`FR`, `avg_FR`, `Kyle_Lambda`). For every variable, the Kolmogorov–Smirnov, Mann–Whitney U, and Levene tests (except for the cancellation rate) reject equality between the unconditional and $Y = 1$ samples, confirming systematic pre-EPM shifts.

For the Net Dollar Limit Order Volume (`NDLOV`) and its averaged version, the $Y = 1$ conditional distributions show substantially higher central tendencies and fatter right tails than the overall sample. This highlights a behavior that is counterintuitive: before EPMs, the net liquidity supply volume (considering submitted limit orders minus canceled limit orders) is higher than usual, not lower. It tells that market makers place more limit orders before EPMs. A logical explanation might be that EPMs happen when there is more trading intensity and, therefore, during those periods of heightened market order volume, market makers have a tendency to adapt and also submit more limit orders to match the market order volume. This supposition will be confirmed if a strong correlation is found between market order volume and net limit order volume, which will be examined in a subsequent section of this study.

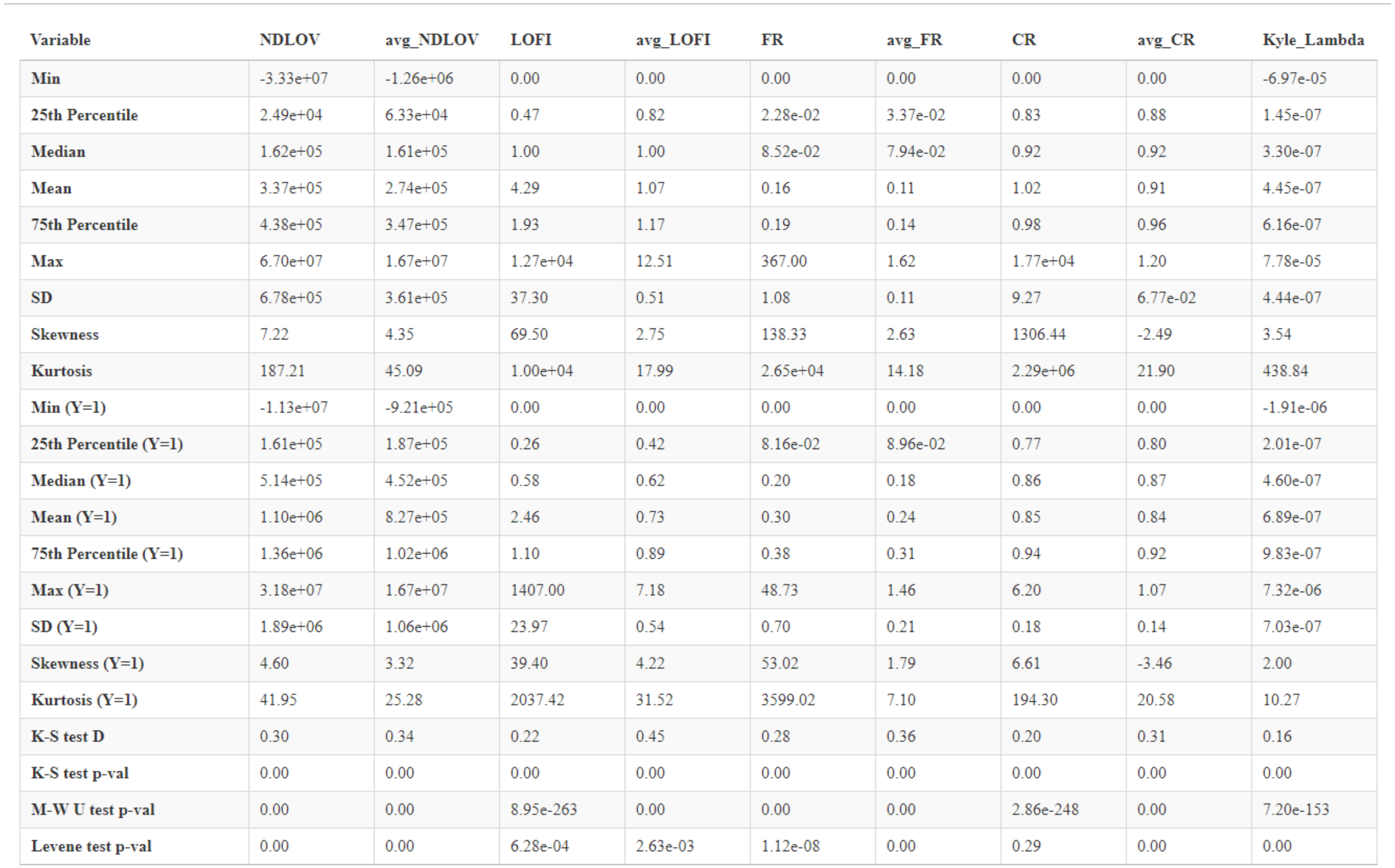
Limit Order Flow Imbalance (`LOFI`) and `avg_LOFI` also display statistics and distributional behavior that is difficult to interpret. It has an interesting median of 1 over the full sample, with a large amount of data showing a `LOFI` over 1, as highlighted by the mean of 4,29. Usually, an imbalance ratio like this one has values between 0 and 1. However, mathematically, as $V_{a,i}^{net}$ and $V_{b,i}^{net}$ can be both positive and negative, it can easily lead high `LOFI` ratio when $V_{a,i}^{net}$ is positive and $V_{b,i}^{net}$ is negative, or the opposite. While the mathematical explanation helps to understand the summary statistics on the full sample, it doesn't explain why there is a distributional shift towards the left on the conditioned sub-sample. One explanation could be that the distribution over the full sample comprises periods of really low limit order volume intensity, while EPMs tend to happen during periods of higher limit order volume intensity. A high limit order volume intensity would normally lead to both $V_{a,i}^{net}$ and $V_{b,i}^{net}$ being positive, giving a number between 0 and 1.

Execution aggressiveness, captured by the Fill Ratio (`FR`) and its average counterpart, also rises before EPMs. The conditional medians are consistently higher, and the right tail is heavier in the $Y = 1$ sub-sample. This suggests that before EPMs, liquidity is consumed at a relatively higher rate compared to its replenishment than under standard conditions, signaling some stress in the order book. The Fill Ratio remains below one (even before EPMs) most of the time because the numerator is made of executed trades at the best LOB quote while the denominator is made of new submitted limit orders at all 5 LOB levels.

In contrast, the Cancellation Ratio (`CR`) and `avg_CR` exhibit a leftward shift before EPMs, indicating that cancellations become less frequent relative to submissions. This change may reflect a market environment in which market makers see that their orders are consumed by more trades than usual. Thus, market makers would submit more limit orders than usual to face the liquidity consumption, ultimately reducing the typical Cancellation Rate. Nevertheless, a higher cancellation rate in the conditioned sub-sample would have made more sense, as we would have expected EPMs coming from liquidity providers flying away from the market.

Finally, Kyle's Lambda, a price impact measure linking order flow to price changes, rises noticeably before EPMs. Both the median and mean values are higher in the conditional sample, and the KDEs show a consistent rightward shift, indicating higher price impact per unit of traded volume when an EPM is imminent.

Overall, the results from these variables suggest that, contrary to a common literature narrative of liquidity providers withdrawing cand causing large price movements, market makers do not appear to vanish in the moments before EPMs. Instead, the data points toward a pattern in which they on average increase their presence in the LOB, potentially seeking to absorb the heightened pressure from the demand side. This behavior, reflected in higher net limit order volumes and lower cancellation rates, indicates that EPMs do not arise from a sudden evaporation of supply. This view was already described and argued by Brogaard (2018) in his scientific paper "High frequency trading and extreme price movements".



Variable	NDLOV	avg_NDLOV	LOFI	avg_LOFI	FR	avg_FR	CR	avg_CR	Kyle_Lambda
Min	-3.33e+07	-1.26e+06	0.00	0.00	0.00	0.00	0.00	0.00	-6.97e-05
25th Percentile	2.49e+04	6.33e+04	0.47	0.82	2.28e-02	3.37e-02	0.83	0.88	1.45e-07
Median	1.62e+05	1.61e+05	1.00	1.00	8.52e-02	7.94e-02	0.92	0.92	3.30e-07
Mean	3.37e+05	2.74e+05	4.29	1.07	0.16	0.11	1.02	0.91	4.45e-07
75th Percentile	4.38e+05	3.47e+05	1.93	1.17	0.19	0.14	0.98	0.96	6.16e-07
Max	6.70e+07	1.67e+07	1.27e+04	12.51	367.00	1.62	1.77e+04	1.20	7.78e-05
SD	6.78e+05	3.61e+05	37.30	0.51	1.08	0.11	9.27	6.77e-02	4.44e-07
Skewness	7.22	4.35	69.50	2.75	138.33	2.63	1306.44	-2.49	3.54
Kurtosis	187.21	45.09	1.00e+04	17.99	2.65e+04	14.18	2.29e+06	21.90	438.84
Min (Y=1)	-1.13e+07	-9.21e+05	0.00	0.00	0.00	0.00	0.00	0.00	-1.91e-06
25th Percentile (Y=1)	1.61e+05	1.87e+05	0.26	0.42	8.16e-02	8.96e-02	0.77	0.80	2.01e-07
Median (Y=1)	5.14e+05	4.52e+05	0.58	0.62	0.20	0.18	0.86	0.87	4.60e-07
Mean (Y=1)	1.10e+06	8.27e+05	2.46	0.73	0.30	0.24	0.85	0.84	6.89e-07
75th Percentile (Y=1)	1.36e+06	1.02e+06	1.10	0.89	0.38	0.31	0.94	0.92	9.83e-07
Max (Y=1)	3.18e+07	1.67e+07	1407.00	7.18	48.73	1.46	6.20	1.07	7.32e-06
SD (Y=1)	1.89e+06	1.06e+06	23.97	0.54	0.70	0.21	0.18	0.14	7.03e-07
Skewness (Y=1)	4.60	3.32	39.40	4.22	53.02	1.79	6.61	-3.46	2.00
Kurtosis (Y=1)	41.95	25.28	2037.42	31.52	3599.02	7.10	194.30	20.58	10.27
K-S test D	0.30	0.34	0.22	0.45	0.28	0.36	0.20	0.31	0.16
K-S test p-val	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
M-W U test p-val	0.00	0.00	8.95e-263	0.00	0.00	0.00	2.86e-248	0.00	7.20e-153
Levene test p-val	0.00	0.00	6.28e-04	2.63e-03	1.12e-08	0.00	0.29	0.00	0.00

Table V.4: Summary Statistics of Liquidity Supply Variables Representing the Evolution of the LOB and Liquidity Supply and Demand Variables

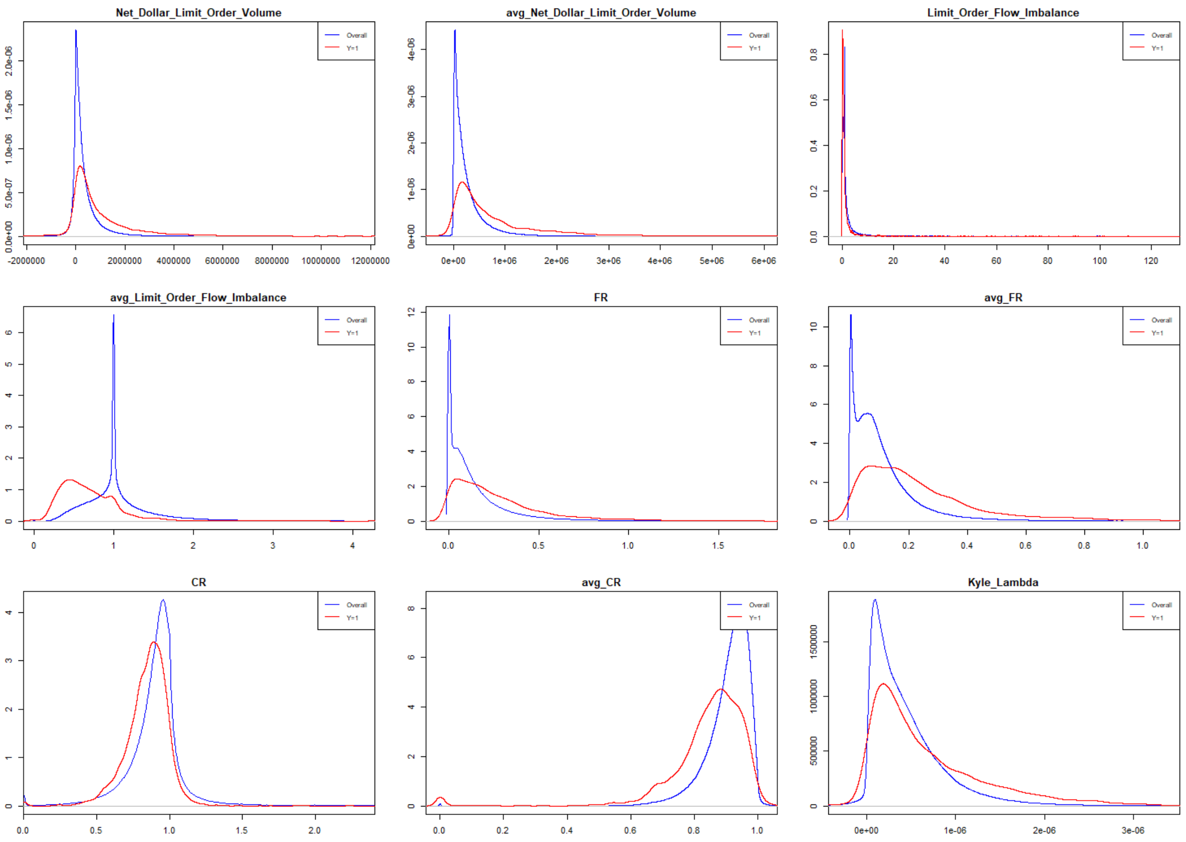


Figure V.4: KDE of Liquidity Supply Variables Representing the Evolution of the LOB and Liquidity Supply and Demand Variables

1.3 Temporal Structure of Variables

As already noted in the methodology, variables computed as 5-minute rolling averages or over the last 5 minutes of data are excluded from the autocorrelation function (ACF) analysis, since their construction inherently induces high autocorrelation due to overlapping observations used to compute those variables. This is the reason why measures such as `PIN` or Median Realized Volatility are absent from the plots displayed in Figure V.5. The ACFs are computed over 30 lags, corresponding to 5 minutes of high-frequency data.

The ACF plots reveal strong persistence patterns across almost all variables. First, market state measures like previous absolute returns and trading intensity (`Dollar_Market_Volume`) have ACF values of nearly 0.4 over all 30 lags, highlighting heteroskedasticity in the returns and cyclical behavior in the trading intensity. Net limit order volume also follows the same trend as market volume (i.e., trading intensity), with similar autocorrelation values.

LOB state measures (`PQS`, `AD`, `QS`, `LOR`, `BLM` and their 5 LOB level equivalent) display medium to sometimes extremely high persistence, with autocorrelation remaining near 0.8 even after 30 lags for the Percent Quoted Spread or Quote Slope. This indicates that the state of the order book is somewhat stable over short horizons.

In contrast, the LOFI, representing imbalance in the arrival of limit orders, and both the fill and cancellation rates, show a very weak autocorrelation compared to the other explanatory variables. Those three variables still have autocorrelation values above the 95% confidence interval of values being statistically different from zero. This is due to the very large sample comprising a full year of data on all stocks (over 6 million observations per variable). Nevertheless, while these autocorrelation values are statistically significant, they remain notably lower than those observed for the other variables.

Overall, the temporal structure indicates strong persistence in most explanatory variables, reflecting that market conditions and LOB characteristics tend to evolve gradually rather than through abrupt independent changes.

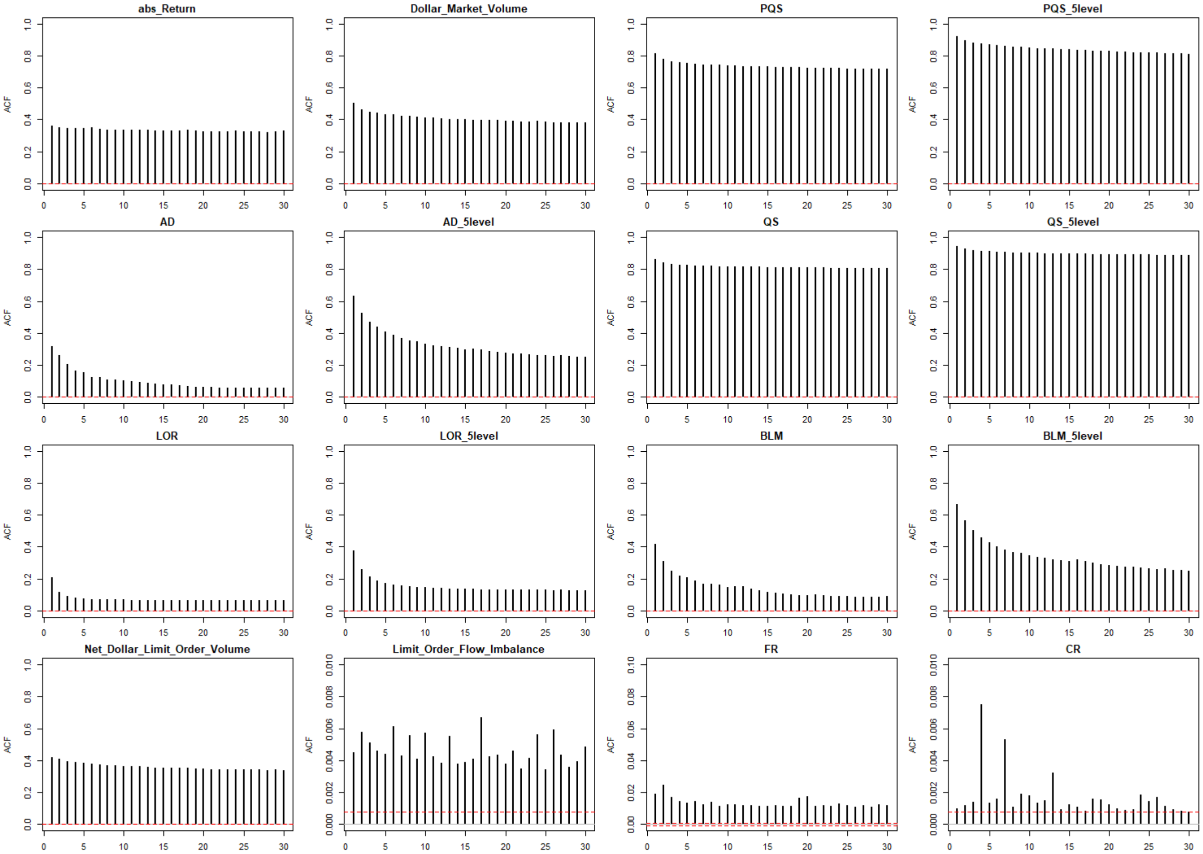


Figure V.5: Autocorrelation Function of Independent Variables Over 30 Lags

1.4 Association Structure Between Variables

1.4.1 Spearman Correlation Matrix

Figure V.6 shows a heatmap of the Spearman rank correlation matrix between variables, ordered so as to focus on variables that cluster together. The dark red colors represent strong positive correlation while the dark blue color represent strong negative correlation.

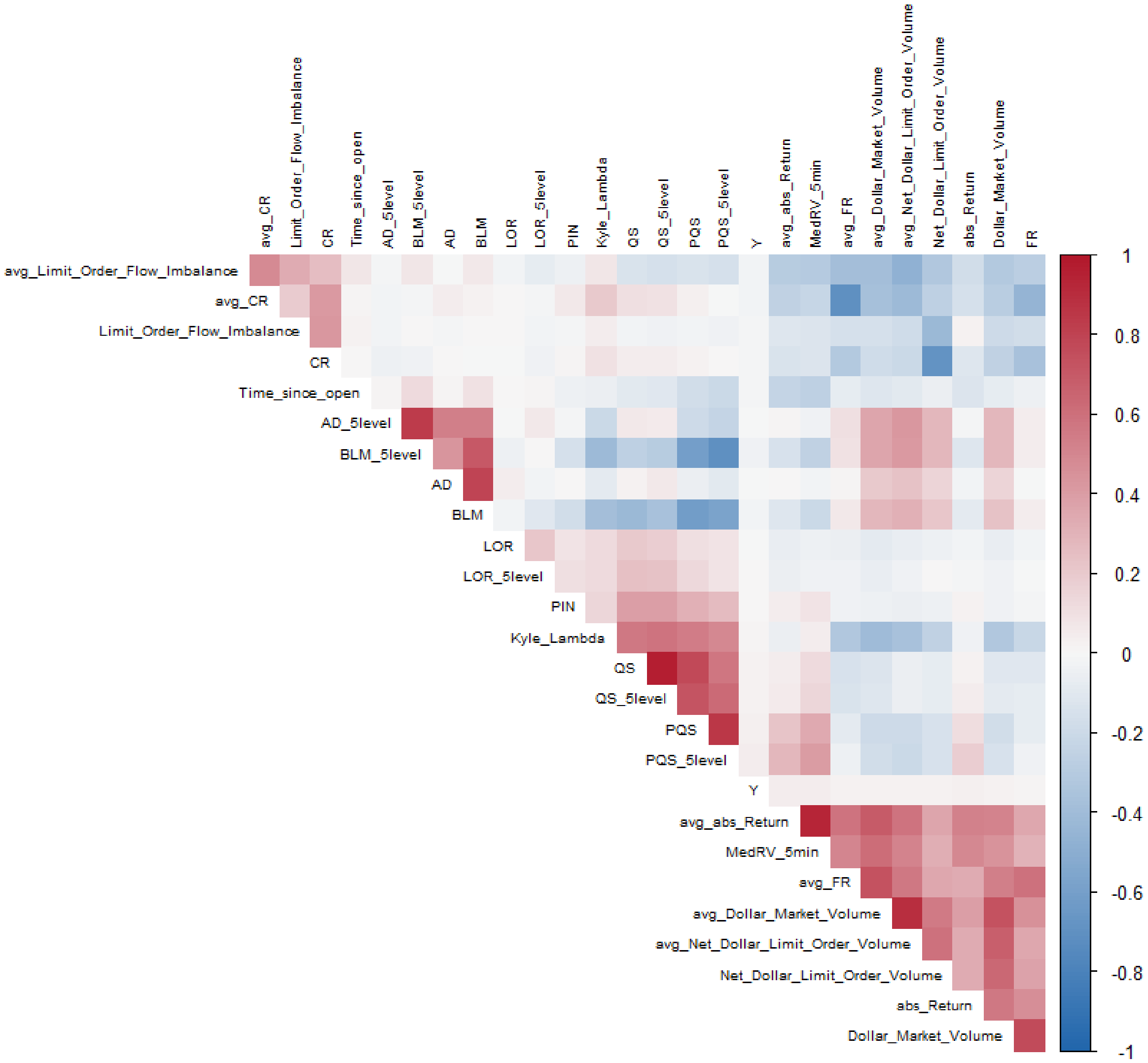


Figure V.6: Spearman Rank Correlation Matrix Between Variables

Four main clusters emerge. The first cluster could be categorized as a volatility and trading intensity cluster, grouping together Absolute Returns, Median Realized Volatility, Trading Intensity (i.e., Market Order Volume), Net Limit Order Volume, and Fill Rates. They reflect a general co-movement of volatility, market activity, and execution aggressiveness together, and support the theory stemming from the literature that trading intensity and market volatility are strongly positively correlated. More interestingly, liquidity supply net volume (`NDLOV`) is also correlated with market order volume. This can be linked to sub-section 1.2 “Kernel Density Estimates and Summary Statistics”, where it was supposed that the right-shift of the `NDLOV` distribution when conditioned on $Y = 1$ was due to `NDLOV` being correlated with market order volume.

Moreover, fill and cancellation rates are negatively correlated and share the same denominator (limit order submission volume), which helps to conclude that when market order volume rises, cancellation volume falls. Consistently, NDLOV is strongly positively correlated with the fill rate and negatively with the cancellation rate. Together, these facts show that when liquidity demand intensifies, liquidity suppliers respond by providing above-average net liquidity and canceling fewer orders.

The second cluster corresponds to (non-)tightness and price impact, with spreads, quote slopes, Kyle's Lambda, PIN, and to a lesser extent the LOB imbalance (LOR) showing positive associations. This can be meaningfully linked to the microstructure literature. One fundamental notion of the literature is the fact that spreads arise from informational asymmetry and order flow toxicity. Indeed, the Probability of Informed Trading in Figure V.6 is shown to have a significantly strong link to the other variables of this cluster and in particular spreads.

The third cluster is simply formed by the two versions of Average Depth (AD, AD_5level) and Bi-dimensional liquidity measure (BLM, BLM_5level). Interestingly, depth is slightly negatively correlated with spreads. The fourth and last cluster is represented by the two versions of Limit Order Flow Imbalance and Cancellation Rate. The fourth cluster is negatively correlated with the first one, confirming the view from sub-section 1.2 that LOFI was smaller when NDLOV was higher, and that high limit order imbalance values were due to low limit order volume.

1.4.2 Cross-Correlation Functions

Figure V.7 displays three CCFs between some main microstructure concepts seen in Figure II.5 of the literature review: the probability of informed trading (measured by the order flow imbalance), the trading intensity and volatility. Trading intensity is taken with its average version (5 minutes of data) to match the fact that both PIN and volatility are measured using 5 minutes of data. Each CCF of this sub-section is titled "X vs Y" and plots

$$\rho_{XY}(k) = \mathrm{Corr}\big(X_{t+k},\, Y_t\big)$$

for integer lags $k \in \{-180, \dots, 180\}$. Each plot thus depicts the relationship between variables with up to a 30-minute lag. By this convention:

- $k > 0$ means Y leads X by k periods.
- $k < 0$ means X leads Y by $|k|$ periods.
- $k = 0$ is contemporaneous.

The CCFs of those three explanatory variables together reveal different patterns of association. The link between trading intensity and volatility is both strong and almost symmetric across positive and negative lags, with CCF values consistently peaking around 0.5 for the contemporaneous correlation ($k = 0$), suggesting a co-movement where periods of heightened trading activity are almost invariably associated with higher median realized volatility. It can be noted that the market intensity seems to have a small lead over the volatility rather than the opposite.

In contrast, the relationships involving PIN are much weaker. Both cross-correlation values with trading intensity and volatility are below 0.06 but remain significant. The relationship between trading intensity and PIN shows persistently higher values at negative lags, indicating clearly that trading intensity tends to mildly lead PIN, with the correlation at negative lags being even higher than the contemporary correlation. On the contrary, the CCF between volatility and PIN highlights a small area of higher correlation values at positive lags close to 0, showing that PIN might be moderately leading volatility.

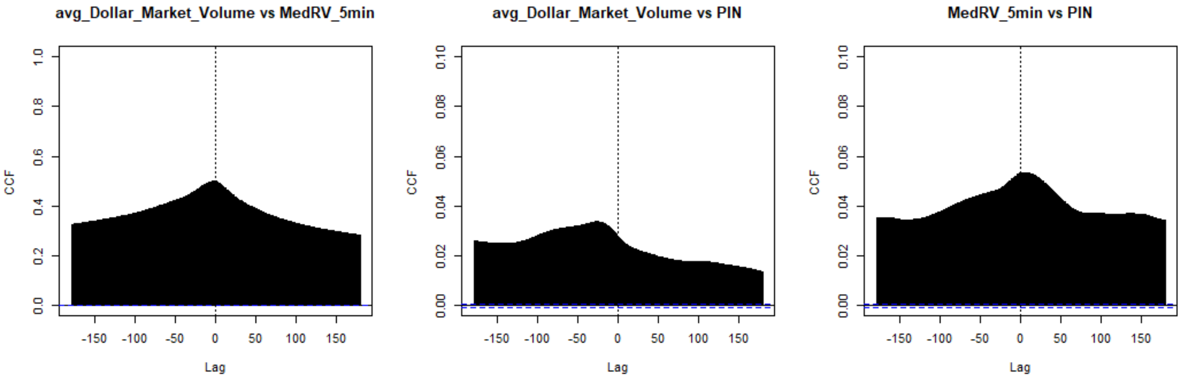
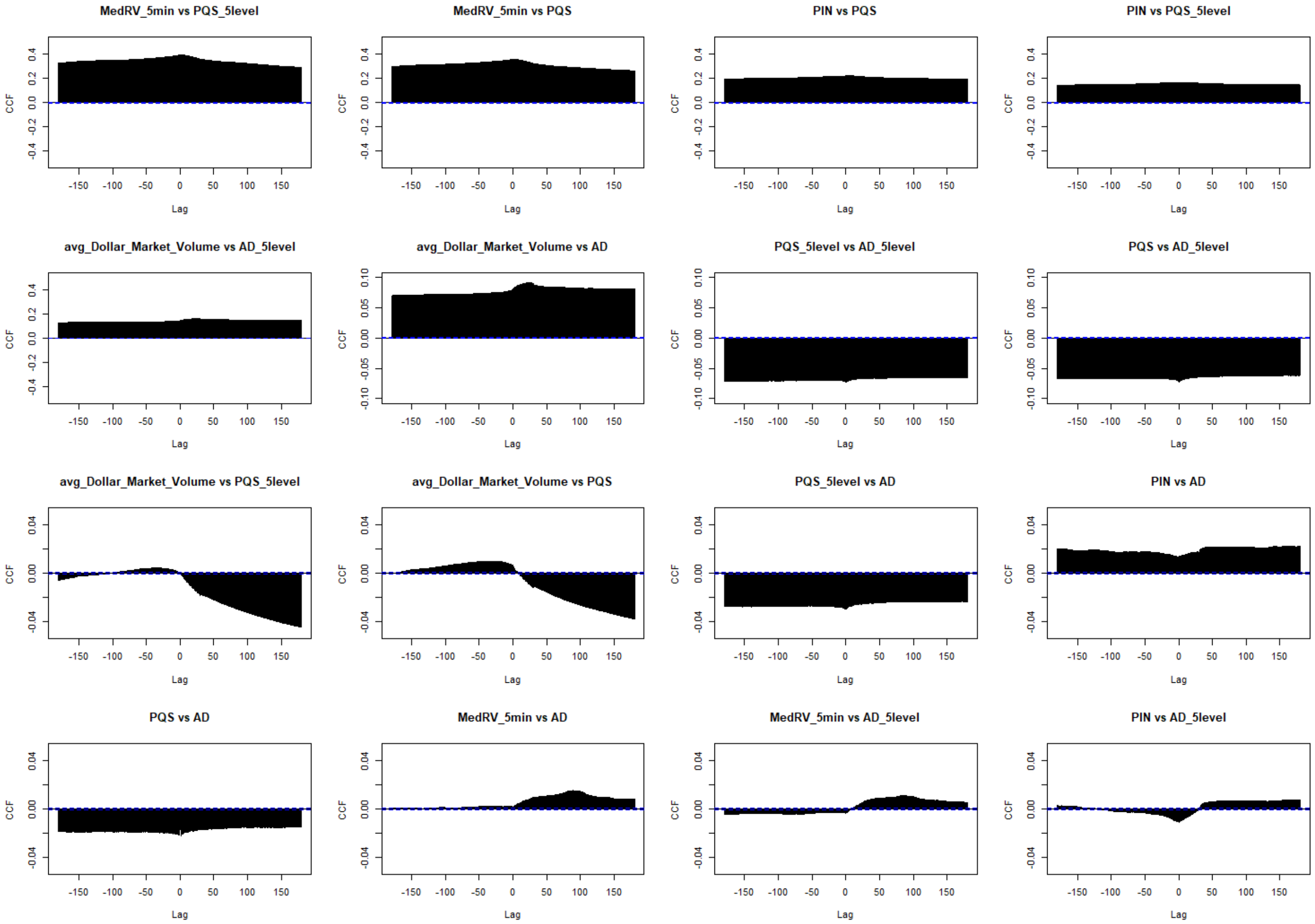


Figure V.7: CCFs Between Trading Intensity, Median Realized Volatility and Probability of Informed Trading

The relationship between the 3 microstructure concepts presented above and liquidity in the LOB was more nuanced in the literature, especially for the link between trading intensity and liquidity supply. The CCFs contained in Figure V.8 are ordered, going from the strongest correlations at the top left to the weakest correlations at the bottom right.

Even if the PIN and volatility are not strongly correlated between each other, they are both strongly correlated with spreads. This shows that PIN and volatility are telling different information to the market (by being very weakly correlated), but are both strongly influencing market makers to widen their spreads during those periods of heightened uncertainty or heightened information asymmetry. A noticeable feature of Figure V.8 is also that PIN and volatility are strongly linked to the spread whereas they are poorly linked to depth, highlighting that all liquidity supply variables do not react the same way or with the same intensity to different market dynamics. To balance this statement, spreads and depth are still logically negatively correlated, showing on average improvement or deterioration of the liquidity at the same time.

CCFs between trading intensity and liquidity supply measures are very interesting. The first noticeable feature of those plots is that trading intensity is highly correlated with average depth, telling a logical story: the more traders consume liquidity at high speed, the more market makers will on average fulfill their role as the liquidity provider by strengthening the LOB. This confirms the view that trading intensity doesn't always lead to worse liquidity in the LOB and that market makers might see heightened market activity as an opportunity to remunerate themselves with the spread rather than withdraw from the market.


MedRV_5min vs PQS_5level

MedRV_5min vs PQS

PIN vs PQS

PIN vs PQS_5level

avg_Dollar_Market_Volume vs AD_5level

avg_Dollar_Market_Volume vs AD

PQS_5level vs AD_5level

PQS vs AD_5level

avg_Dollar_Market_Volume vs PQS_5level

avg_Dollar_Market_Volume vs PQS

PQS_5level vs AD

PIN vs AD

PQS vs AD

MedRV_5min vs AD

MedRV_5min vs AD_5level

PIN vs AD_5level

Figure V.8: CCFs Between Trading Intensity, Median Realized Volatility, Probability of Informed Trading and Liquidity Supply Main Variables (Spreads and Depths)

The second noticeable feature of those CCFs is the strong lead–lag relationship between market activity and spreads. Even if the CCF values are below 0.05, the two CCFs of trading intensity with spreads tell two stories. First, high market activity tends to be followed by a small increase in spreads. Second, and most importantly, spreads seem to be able to regulate in a modest way trading intensity, with a latent market adjustment to higher trading costs.

Finally, Figure A.1 to Figure A.3 (*see Appendix A*) shows all CCFs between the dependent variable Y and all independent variables for integer lags $k \in \{-10, \dots, 10\}$, ordered from CCFs with highest correlation (at a certain lag) in Figure A.1 to those with lowest correlation in Figure A.3. Those relationships between Y and explanatory variables were already depicted in the KDE plots. However, those CCFs give some more information. First, it reaffirms that market intensity, volatility and spreads seem to have the most direct link to EPMs. Also, it confirms that data manipulation was done correctly thanks to the correlation between absolute returns and Y at lag 1. Indeed, Y has been defined as the 99.9^{th} percentile of absolute returns in the next interval, thus showing this definition at lag 1. The bump at lag 1 of the correlation with trading intensity is also noticeable. Last, those CCFs, together with the ACFs from sub-section 1.3 Temporal Structure of Variables, highlight a key fact: given the strong temporal autocorrelation across most variables, and its reflection in the CCFs with Y, predicting the exact timing of an EPM is inherently difficult. What a model can more realistically achieve is to identify windows where market conditions and the state of the liquidity supply are consistently aligned with elevated EPM probability. In essence, the market and the LOB evolve in somewhat persistent states, and these states are heavily linked to their past state.

2 Logistic Regressions Results and Comparison

2.1 Chosen Predictors of the Elastic Net Logistic Regression

2.1.1 Alpha and Lambda Tuning

Figure V.9 presents, for each stock and each α on the x-axis, the value of the cross-validated AUC (of the ROC) optimized for λ. The plotted curves remain quite flat, indicating that the performance remains stable across the whole range of alphas. The high AUC values are not necessarily synonyms of high performance. Indeed, the highly-imbalanced dataset of EPMs (positives) and non-EPMs (negatives) makes the FPR stay very small and has a tendency to "inflate" the curve, giving high AUC scores even for mediocre classifiers. Thus, those high AUC values must be interpreted with precaution.

Table V.5 gives a final picture of all (α, λ) optimal couples for each stock, with the resulting number of variables displayed. Values of α span across the 0 to 0.7 range, indicating a slight preference for the Ridge penalty, with even FB being a pure Ridge regression. Most stocks select an α in the 0.1–0.3 range. Values of λ fall in the 10^{-5}–10^{-3} range. A larger λ implies a higher shrinkage of the coefficients, which can be seen by the fact that the stronger the λ, the least amount of variables are being kept by the model on average. The amount of selected variables thus varies widely from 9 to 26 predictors kept out of 26. The FB ridge regression keeps all predictors as no coefficients are forced to 0.

Overall, the α–AUC curves and chosen (α, λ) confirm non-homogeneity across stocks: different stocks need different regularization strengths and yield materially different model sizes.

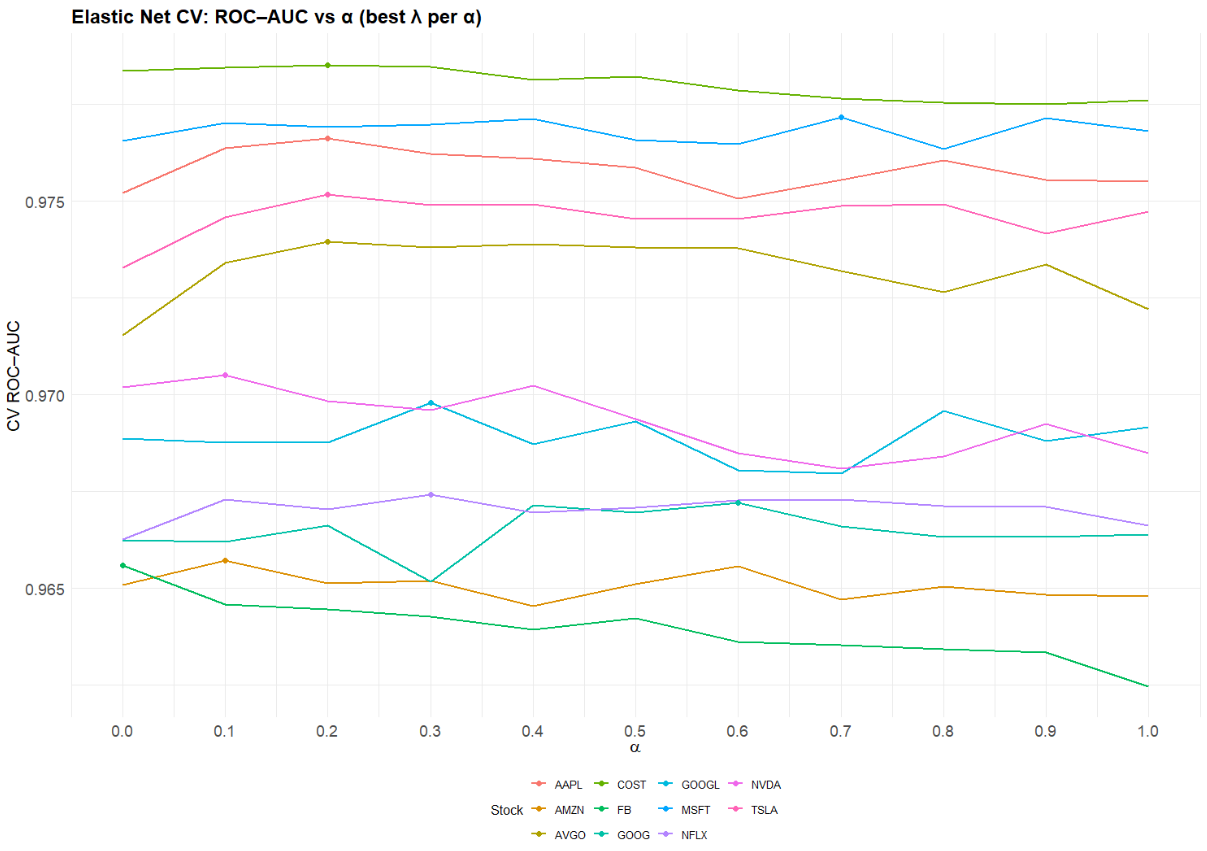
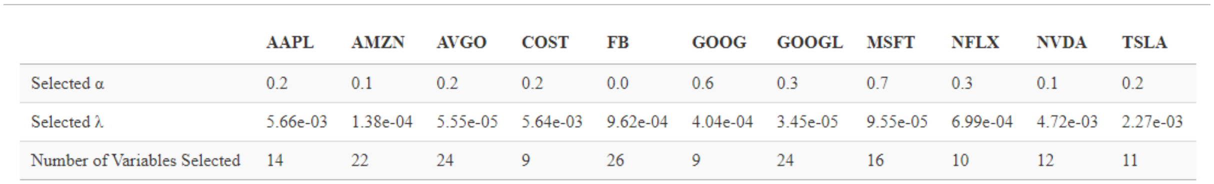


Figure V.9: (α, λ) Cross-Validated Area Under the Receiver Operating Characteristic Curve for all Stocks



	AAPL	AMZN	AVGO	COST	FB	GOOG	GOOGL	MSFT	NFLX	NVDA	TSLA
Selected α	0.2	0.1	0.2	0.2	0.0	0.6	0.3	0.7	0.3	0.1	0.2
Selected λ	5.66e-03	1.38e-04	5.55e-05	5.64e-03	9.62e-04	4.04e-04	3.45e-05	9.55e-05	6.99e-04	4.72e-03	2.27e-03
Number of Variables Selected	14	22	24	9	26	9	24	16	10	12	11

Table V.5: Final (α, λ) Selected by the Elastic Net Cross-Validation Procedure With Their Respective Number of Variables That Do Not Have Their Coefficient Shrunk to 0

2.1.2 Selected Variables

Table V.6 takes a deep dive in the specific selected variables of the Elastic Net regression across all stocks, ordering them by number of appearance. Predictors fall into 3 categories: those who are nearly always selected (9 times or more out of 11), those who are frequently selected (6 to 8 times out of 11), and those who are more rarely selected (5 or less times out of 11).

Consistently selected predictors. Across stocks, the Elastic Net almost always keeps the volatility group driven by previous absolute returns (abs_Return, avg_abs_Return), tightness measures (PQS, PQS_5level), and trading intensity (Dollar_Market_Volume, avg_Dollar_Market_Volume). This is exactly the picture established in the descriptive and exploratory analysis: market-state volatility and volume rise before EPMs (see Table V.2 and Figure V.2), with strong KS test D statistics; and spreads widen, with the variant on the 5-level LOB amplifying the signal (see Figure V.3).

avg_Net_Dollar_Limit_Order_Volume, as a variable correlated with the volatility and intensity variables just described above (see Figure V.6), is also part of the most frequently chosen variables.

Rates in their average version (avg_FR and avg_CR) are also chosen almost across every stock, coherent with their KDE plots (see Figure V.4). It is interesting to note that the measures representing averages over the last 30 intervals are always preferred than those that only picture the last 10-second interval (with the exception of abs_Return and PQS because both versions are chosen 11 times out of 11).

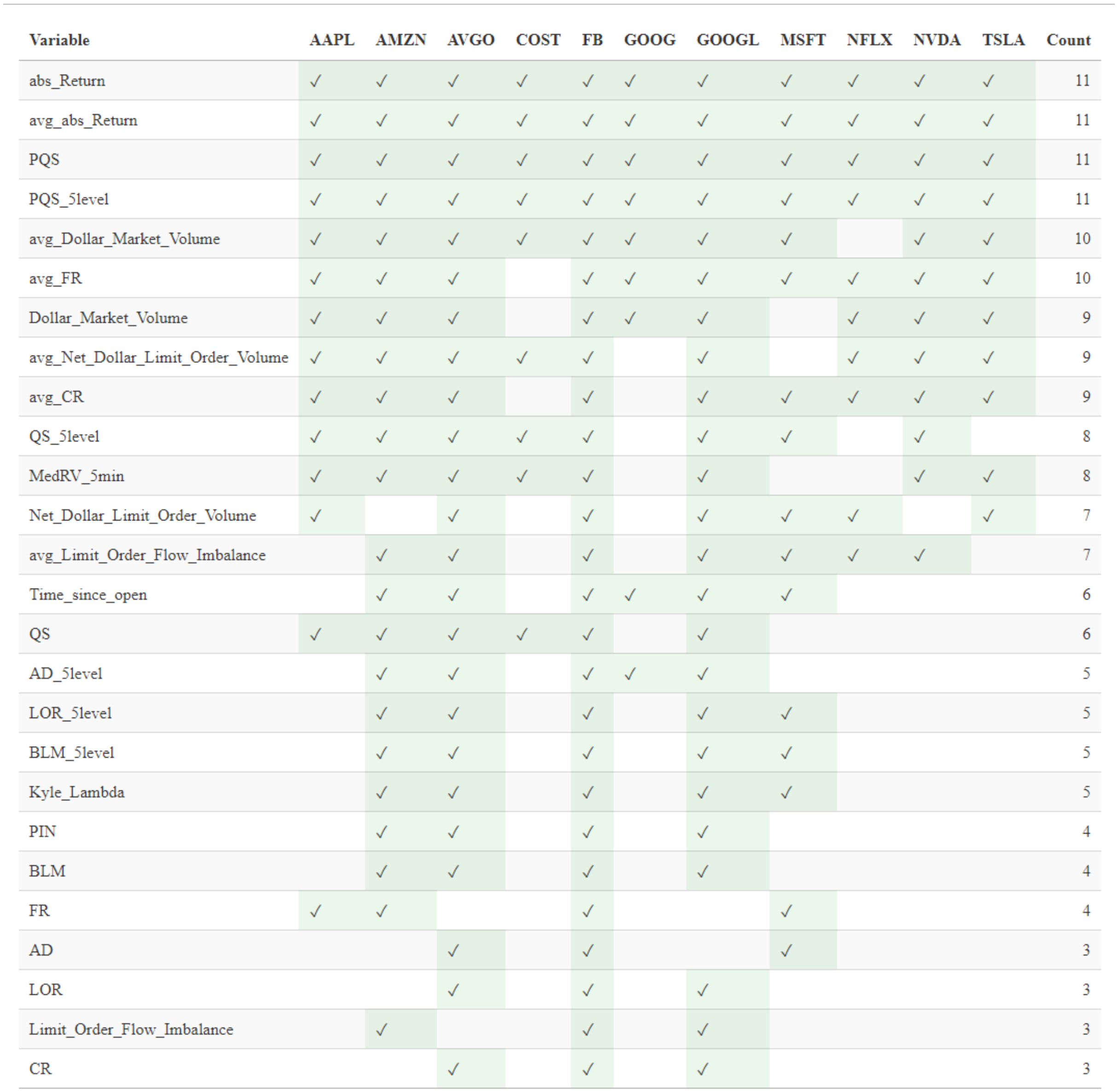
Frequently selected predictors. Beyond this core group, the model often keeps order-book slope (QS, QS_5level) together with Median RV (MedRV_5min), Net_Dollar_Limit_Order_Volume in its simple version, avg_Limit_order_Flow_Imbalance and Time_Since_Open. It becomes noticeable that more simple measures of previous absolute returns are preferred to the more mathematically complex and literature-guided measure of Median Realized Volatility by Andersen et al. (2012), an estimator which has been designed for high frequency data. Also, the Time variable does not appear to be much chosen compared to its more explicit KDE (see Figure V.2). It suggests that time in itself does not contain a lot more information than what is described by other market-state and liquidity variables.

Occasionally to rarely selected predictors. By contrast, static liquidity supply variables (AD, AD_5level, BLM, BLM_5level) and liquidity supply imbalances (LOR, LOR_5level, raw_LOFI) are only sporadically useful once tightness, slope, and execution pressure are already in the model. Furthermore, theoretically sound price impact and information-asymmetry proxies, respectively Kyle_Lambda and PIN, do not emerge as primary drivers according to this variable selection protocol. This is consistent with the Descriptive Analysis, where both measures exhibited smaller pre-EPM shifts and heavier overlap between unconditioned and conditioned KDEs (see Figure V.2 and Figure V.4).

In the same way average measures seem to be preferred, liquidity supply measures computed on five levels of the LOB are more often chosen, highlighting the fact that variables computed over more than one level of the LOB give a more complete view about the current state of this LOB.

Global Picture. Taken together, the Elastic Net paints a coherent pre-EPM mechanics. The signal is anchored in simple proxies: previous absolute returns, quoted tightness, sustained trading intensity, fill rate, and cancellation rate. Multi-level LOB aggregations add context that single-level snapshots miss, and once tightness, slope, and persistent flow are in the model, static depth and LOB imbalances contribute little incremental information and are thus wiped out of the model more often. More elaborate constructs like MedRV, Kyle's λ, and PIN are also not the main drivers of EPMs.

Finally, the main differences across stocks seem to be more linked to how regularized by λ the final Elastic Net regression model is, rather than linked to which specific variables survive. In other words, the less frequently chosen variables are chosen because the model penalizes less the coefficients, not because of differing variable preferences across stock.



Variable	AAPL	AMZN	AVGO	COST	FB	GOOG	GOOGL	MSFT	NFLX	NVDA	TSLA	Count
abs_Return	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	11
avg_abs_Return	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	11
PQS	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	11
PQS_5level	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	11
avg_Dollar_Market_Volume	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	10
avg_FR	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	10
Dollar_Market_Volume	✓	✓	✓		✓	✓	✓		✓	✓	✓	9
avg_Net_Dollar_Limit_Order_Volume	✓	✓	✓	✓	✓		✓		✓	✓	✓	9
avg_CR	✓	✓	✓		✓		✓	✓	✓	✓	✓	9
QS_5level	✓	✓	✓	✓	✓		✓	✓		✓		8
MedRV_5min	✓	✓	✓	✓	✓		✓			✓	✓	8
Net_Dollar_Limit_Order_Volume	✓		✓		✓		✓	✓	✓		✓	7
avg_Limit_Order_Flow_Imbalance		✓	✓		✓		✓	✓	✓	✓		7
Time_since_open		✓	✓		✓	✓	✓	✓				6
QS	✓	✓	✓	✓	✓		✓					6
AD_5level		✓	✓		✓	✓	✓					5
LOR_5level		✓	✓		✓		✓	✓				5
BLM_5level		✓	✓		✓		✓	✓				5
Kyle_Lambda		✓	✓		✓		✓	✓				5
PIN		✓	✓		✓		✓					4
BLM		✓	✓		✓		✓					4
FR	✓	✓			✓			✓				4
AD			✓		✓			✓				3
LOR			✓		✓		✓					3
Limit_Order_Flow_Imbalance		✓			✓		✓					3
CR			✓		✓		✓					3

Table V.6: Variables Selected by the Elastic Net Regression Across Stocks and Their Total Number of Appearance

2.2 Logistic Regression Results on the Training Set

2.2.1 Interpreting the Regression Coefficients Across Model Specifications

With the set of selected variables by the Elastic Net regression in sub-section 2.1 *Chosen Predictors of the Elastic Net Logistic Regression*, the four different models have been run on the training set. Tables B.1 to B.4 in Appendix B depict the resulting coefficients of the standardized variables for all four models: the unpenalized model with all variables, the unpenalized model with only market-state variables, the Elastic Net model, and the unpenalized model with variables chosen by the Elastic Net.

In a multivariable regression, each coefficient is a partial effect: it captures the association between x_j and the dependent variable (EPM probability) conditional on (i.e., holding fixed) all other variables in the model. Its sign therefore reflects this adjusted relationship and can legitimately differ from the marginal sign one would obtain in a univariate regression of the outcome on x_j alone, or from the intuition suggested by univariate diagnostics such as the KDEs of sub-section 1.2 *Kernel Density Estimates and Summary Statistics*. When predictors are correlated, conditioning can even produce suppression and sign reversals, so the multivariable sign should be interpreted as a conditional effect rather than a raw association. In sum, conclusions on the sign of the regression coefficients should be taken with great caution considering the multicollinearity of this study's independent variables. A prime example of this effect can be viewed in `MedRV_5min`. Its regression coefficients are often negative and strongly significant across all models in Tables B.1 to B.4 in Appendix B, whereas the summary statistics and KDE on the univariate relationship with EPMs vigorously indicated the opposite. This conditional negative effect is probably due to the predictor's high collinearity with variables such as `avg_abs_Return` or `avg_Dollar_Market_Volume`. This switch in relationship sign between `MedRV_5min` and EPMs from the univariate relationship to the multivariate relationship also gives insight into the reason why this variable has been less selected than its `abs_Return` and `avg_abs_Return` counterparts by the Elastic Net regression.

With these caveats in mind, several predictors display multivariable regression signs that coincide with the univariate evidence from the descriptive analysis, and do so with notable cross-stock stability. In particular, the simple volatility proxies, `abs_Return` and mostly `avg_abs_Return`, enter with positive and typically highly significant coefficients across all models, mirroring the right-shift observed in the KDEs. Quoted spreads, especially its 5-levels counterpart `PQS_5level`, are likewise positive and often strongly significant, consistent with widening spreads prior to EPMs. Trading intensity (`Dollar_Market_Volume`, `avg_Dollar_Market_Volume`) exhibits almost always positive effects across models and stocks, with some cases interestingly showcasing negative coefficients. Those coherent and strong predictors are also those who have been chosen regularly by the Elastic Net regression in sub-section 2.1.2 *Selected Variables*.

From those regression results, other more subtle effects can be described on those variables that have been more neglected by the Elastic Net regression:

- `Time_since_open`: When put in a multivariate regression, the intraday time has the opposite sign (positive) as the one KDE would suggest (negative). Indeed, when significant, the coefficient sign is positive, showcasing a higher tendency of EPMs happening at the end of the day when controlling for other variables. The only exception is for TSLA in the market-state regression.

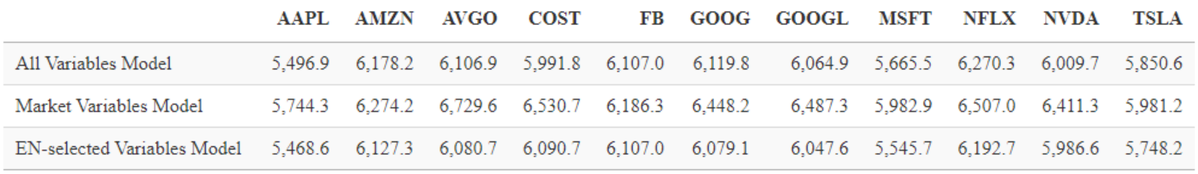
- PIN: The multivariate regression sign of PIN is mostly coherent with its KDE. It sometimes shows strongly significance, highlighting the added value of the probability of informed trading in regression models to predict EPMs.
- AD and AD_5level: Even if not chosen often by the Elastic Net model and having no obvious influence on EPMs when looking at the descriptive analysis, average depth appears to often have a strongly significant positive effect on EPM likelihood, when controlling for all other variables and selected variables. It is a big surprise, as it suggests that EPMs will happen more when the depth of the order book liquidity is stronger, with other variables already taken into account.
- BLM_5level: The bi-dimensional liquidity measure seems to be a great and coherent predictor of EPMs when controlling for the effects of other variables. The strong and significant negative signs suggest that smaller BLMs on 5 LOB levels (i.e., wider spreads in combination with less depth), representing a weaker liquidity supply considering tightness and depth combined, coincide with more EPMs. Interestingly not chosen a lot by the Elastic Net model, this liquidity indicator presents the biggest coefficients across all stocks.
- avg_Limit_Order_Flow_Imbalance: The sign of the coefficient is negative and often significant, confirming the already counterintuitive univariate effect from the descriptive analysis: EPMs tend to happen while liquidity supply is more balanced, even when controlling for both versions of Net_Dollar_Limit_Order_Volume.

2.2.2 Comparing Model Fit to Model Complexity

Table V.7 shows the Bayesian Information Criterion for the three unpenalized models. One clear pattern stands out: the unpenalized model with variables selected by the Elastic Net regression has the best balance between model fit and model complexity (on 9 out of 11 stocks). According to the BIC, it is the model that best explains the data without being overly complex. On the opposite, the market-state benchmark model performs clearly less well than the other two by several hundred BIC points. As a general rule of thumb, the reference thresholds of BIC differences, around 2, 6, and 10, are often read as positive, strong, and very strong support for the lower-BIC alternative. The differences between the BIC of the three models are therefore important in size.

This pattern makes sense considering that the training sample size is around 472,530 observations per stock. Indeed, the BIC penalty with such a sample size gives a penalty of 13 per extra variable in the model. Dropping 10 non-informative variables would "save" 130 BIC points if they do not deliver an in-sample likelihood gain. The EN-selected model often keeps a core block of variables and discards weak or redundant ones, capturing a lot of the fit of the full model with fewer regressors.

In conclusion, EN-selected unpenalized logistic regression is the preferred specification when assessing it on the BIC for the vast majority of stocks, having the best balance between fit and complexity. The market-state-only specification is not competitive, confirming that market-state variables alone are insufficient and that liquidity variables have a strong added-value to predict EPMs.



	AAPL	AMZN	AVGO	COST	FB	GOOG	GOOGL	MSFT	NFLX	NVDA	TSLA
All Variables Model	5,496.9	6,178.2	6,106.9	5,991.8	6,107.0	6,119.8	6,064.9	5,665.5	6,270.3	6,009.7	5,850.6
Market Variables Model	5,744.3	6,274.2	6,729.6	6,530.7	6,186.3	6,448.2	6,487.3	5,982.9	6,507.0	6,411.3	5,981.2
EN-selected Variables Model	5,468.6	6,127.3	6,080.7	6,090.7	6,107.0	6,079.1	6,047.6	5,545.7	6,192.7	5,986.6	5,748.2

Table V.7: Bayesian Information Criterion Values
on the Three Unpenalized Regression Models across all Stocks

2.3 Logistic Regression Results on the Test Set

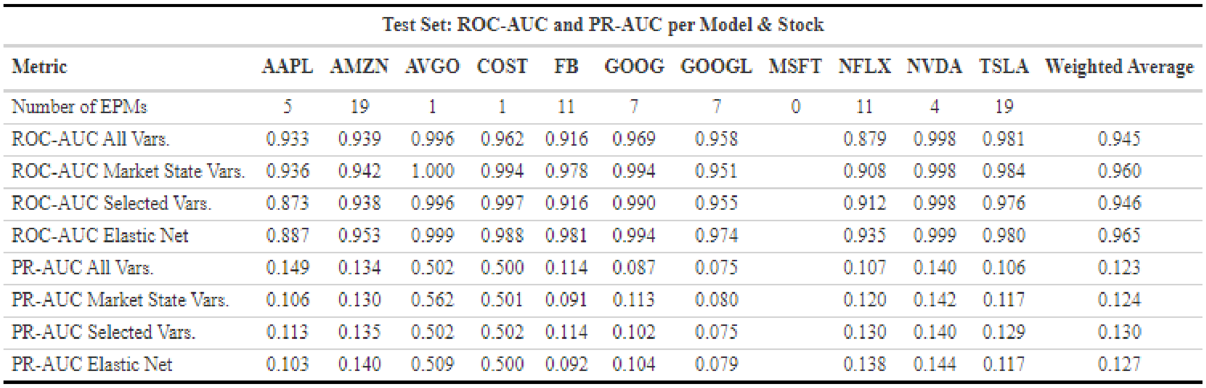
2.3.1 Threshold-Free Discrimination Performance

Table V.8 summarizes how well the four models rank EPM vs. non-EPM cases on the test set without imposing a probability cutoff. Both ROC-AUC values and PR-AUC values are averaged across all stocks by weighting for the number of true EPMs inside the test set. MSFT does not display any value in Table V.8 and for any test set results table. This is because MSFT has exactly 0 EPM recorded for the months of November and December. For illustrative purposes, Figure V.10 displays the ROC and PR curves for AMZN under the Elastic Net and the Market State models. The graph spikes because there are very few positive cases to detect in the test set.

As expected under extreme imbalance, ROC–AUCs are uniformly very high. On the weighted average, Elastic Net achieves the highest ROC-AUC, followed by the market-state benchmark, while the All-variables and EN-selected specifications are tied.

On the other hand, PR–AUC is more informative under this class imbalance because it focuses solely on a proper EPM prediction, balancing the proportion of EPMs detected with the accuracy with which they are detected. On the weighted average, the EN-selected variables unpenalized logistic regression achieves the best PR–AUC (≈ 0.130), followed by the Elastic Net regression itself (≈ 0.127). The Market-state benchmark (≈ 0.124) and the All-variables model (≈ 0.123) have slightly lower accuracy on the calibrated detection of EPMs. This indicates that liquidity variables add out-of-sample discriminatory power when chosen with parsimony.

In sum, judging by the metric that matters the most under the imbalance, PR-AUC, the EN-selected logit is the strongest threshold-free discriminator on the test set, with Elastic Net a close second. Market-state variables alone provide a strong baseline, but adding a compact set of liquidity features improves EPM detection out-of-sample.



Test Set: ROC-AUC and PR-AUC per Model & Stock												
Metric	AAPL	AMZN	AVGO	COST	FB	GOOG	GOOGL	MSFT	NFLX	NVDA	TSLA	Weighted Average
Number of EPMs	5	19	1	1	11	7	7	0	11	4	19	
ROC-AUC All Vars.	0.933	0.939	0.996	0.962	0.916	0.969	0.958		0.879	0.998	0.981	0.945
ROC-AUC Market State Vars.	0.936	0.942	1.000	0.994	0.978	0.994	0.951		0.908	0.998	0.984	0.960
ROC-AUC Selected Vars.	0.873	0.938	0.996	0.997	0.916	0.990	0.955		0.912	0.998	0.976	0.946
ROC-AUC Elastic Net	0.887	0.953	0.999	0.988	0.981	0.994	0.974		0.935	0.999	0.980	0.965
PR-AUC All Vars.	0.149	0.134	0.502	0.500	0.114	0.087	0.075		0.107	0.140	0.106	0.123
PR-AUC Market State Vars.	0.106	0.130	0.562	0.501	0.091	0.113	0.080		0.120	0.142	0.117	0.124
PR-AUC Selected Vars.	0.113	0.135	0.502	0.502	0.114	0.102	0.075		0.130	0.140	0.129	0.130
PR-AUC Elastic Net	0.103	0.140	0.509	0.500	0.092	0.104	0.079		0.138	0.144	0.117	0.127

Table V.8: Out-of-Sample ROC-AUC and PR-AUC Values for Each Model Across all Stocks

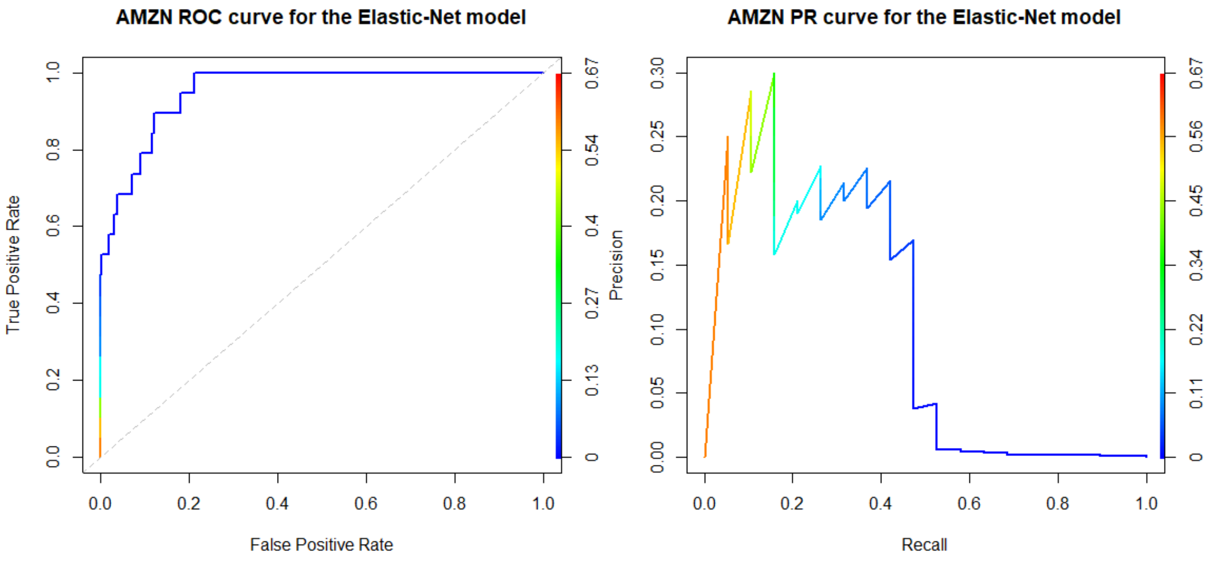
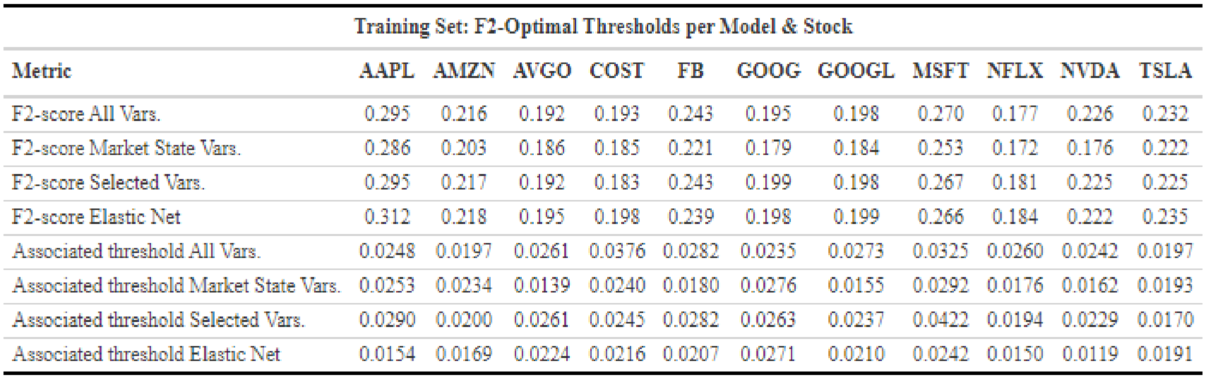


Figure V.10: Out-of-Sample ROC and PR Curves for AMZN Using the Elastic Net Model

2.3.2 Test Set Binary Classification With an F2-Optimized Threshold

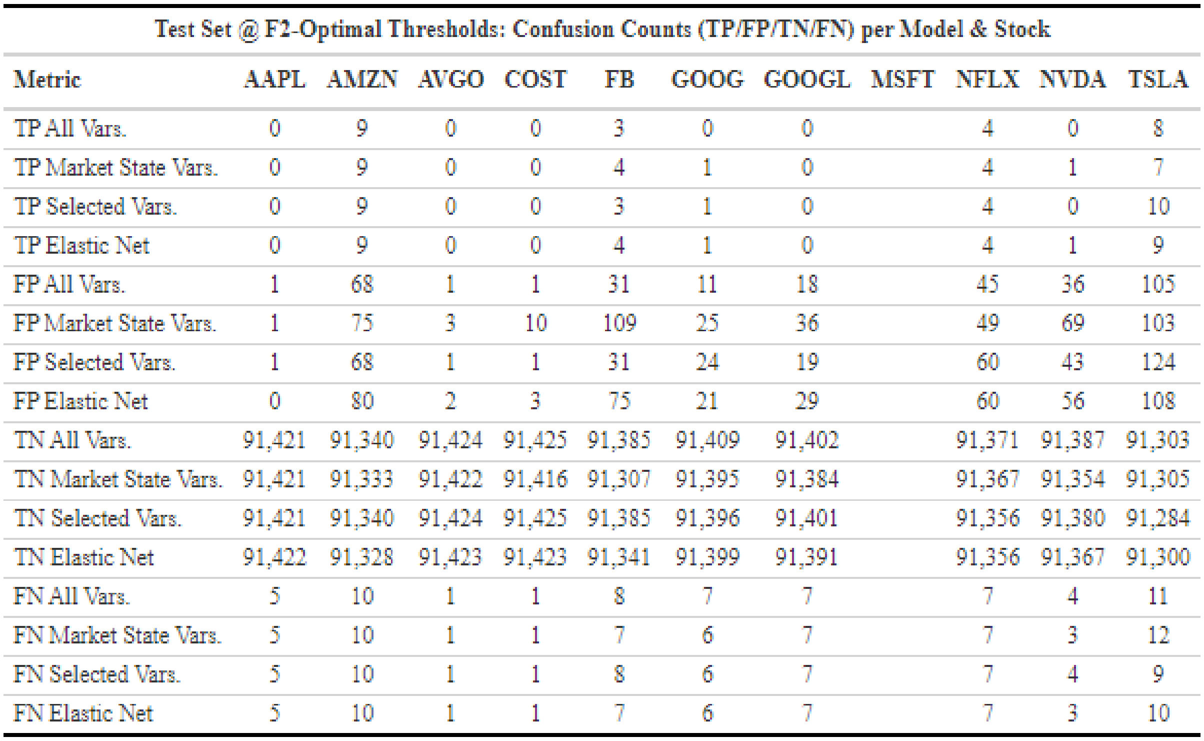
As already explained in the Methodology sub-section 2.2.3 *Regression Evaluation and Comparison on the test set*, a threshold is selected by taking the one that optimizes the F2-score on the training set. Table V.9 depicts, for each model across every stock, the maximized F2-score on the training set, as well as the optimal threshold that maximizes it. Table V.9 shows that the optimal cutoffs to maximize F2 are very small, from 1.19% to 4.22% across models and names. This is logical considering that F2 is a recall-heavy metric (i.e., it focuses on having a good recall). To be able to recover enough EPMs, the binary classification must select predicted EPMs on tiny probabilities. One interesting aspect is that thresholds of the Elastic Net regression are systematically lower: more penalized coefficients require lower thresholds to focus on recall.



Training Set: F2-Optimal Thresholds per Model & Stock											
Metric	**AAPL**	**AMZN**	**AVGO**	**COST**	**FB**	**GOOG**	**GOOGL**	**MSFT**	**NFLX**	**NVDA**	**TSLA**
F2-score All Vars.	0.295	0.216	0.192	0.193	0.243	0.195	0.198	0.270	0.177	0.226	0.232
F2-score Market State Vars.	0.286	0.203	0.186	0.185	0.221	0.179	0.184	0.253	0.172	0.176	0.222
F2-score Selected Vars.	0.295	0.217	0.192	0.183	0.243	0.199	0.198	0.267	0.181	0.225	0.225
F2-score Elastic Net	0.312	0.218	0.195	0.198	0.239	0.198	0.199	0.266	0.184	0.222	0.235
Associated threshold All Vars.	0.0248	0.0197	0.0261	0.0376	0.0282	0.0235	0.0273	0.0325	0.0260	0.0242	0.0197
Associated threshold Market State Vars.	0.0253	0.0234	0.0139	0.0240	0.0180	0.0276	0.0155	0.0292	0.0176	0.0162	0.0193
Associated threshold Selected Vars.	0.0290	0.0200	0.0261	0.0245	0.0282	0.0263	0.0237	0.0422	0.0194	0.0229	0.0170
Associated threshold Elastic Net	0.0154	0.0169	0.0224	0.0216	0.0207	0.0271	0.0210	0.0242	0.0150	0.0119	0.0191

Table V.9: Thresholds That Maximize the F2-Score Across Models and Stocks

Using the thresholds of Table V.9, the confusion counts (i.e., the principal components of confusion matrices: TP, TN, FP, and FN) from computed predictions on the test set are displayed in Table V.10. As already highlighted by Table V.1 of the sub-section 1.1 *EPM Month-by-Month Distribution*, the total number of positives is very little. At first glance, the difference between models is not flagrant. For stocks with very few or even no EPM detections (e.g., AAPL, GOOG), all models underperform, whereas for models with more EPM detections (e.g., AMZN, TSLA), recall is uniformly high without a clear model standing out.



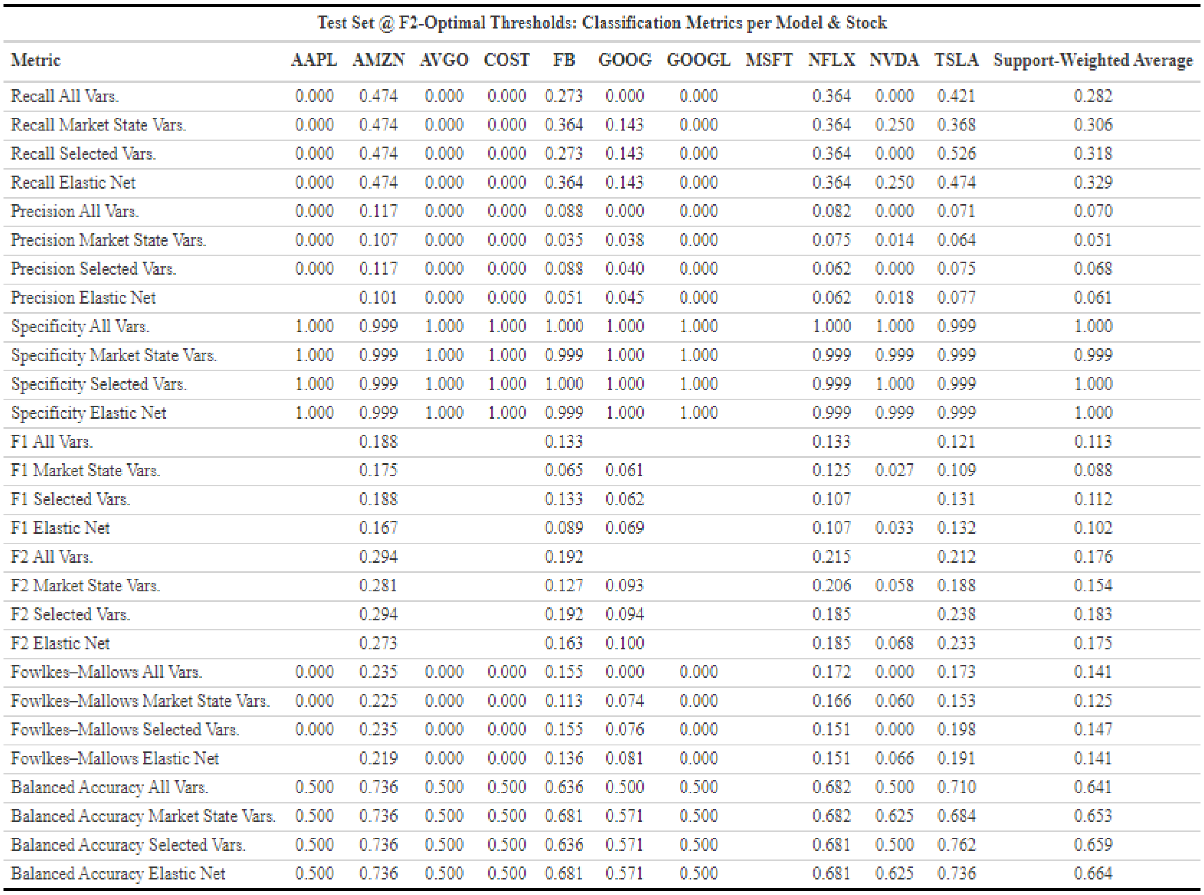
Test Set @ F2-Optimal Thresholds: Confusion Counts (TP/FP/TN/FN) per Model & Stock											
Metric	**AAPL**	**AMZN**	**AVGO**	**COST**	**FB**	**GOOG**	**GOOGL**	**MSFT**	**NFLX**	**NVDA**	**TSLA**
TP All Vars.	0	9	0	0	3	0	0		4	0	8
TP Market State Vars.	0	9	0	0	4	1	0		4	1	7
TP Selected Vars.	0	9	0	0	3	1	0		4	0	10
TP Elastic Net	0	9	0	0	4	1	0		4	1	9
FP All Vars.	1	68	1	1	31	11	18		45	36	105
FP Market State Vars.	1	75	3	10	109	25	36		49	69	103
FP Selected Vars.	1	68	1	1	31	24	19		60	43	124
FP Elastic Net	0	80	2	3	75	21	29		60	56	108
TN All Vars.	91,421	91,340	91,424	91,425	91,385	91,409	91,402		91,371	91,387	91,303
TN Market State Vars.	91,421	91,333	91,422	91,416	91,307	91,395	91,384		91,367	91,354	91,305
TN Selected Vars.	91,421	91,340	91,424	91,425	91,385	91,396	91,401		91,356	91,380	91,284
TN Elastic Net	91,422	91,328	91,423	91,423	91,341	91,399	91,391		91,356	91,367	91,300
FN All Vars.	5	10	1	1	8	7	7		7	4	11
FN Market State Vars.	5	10	1	1	7	6	7		7	3	12
FN Selected Vars.	5	10	1	1	8	6	7		7	4	9
FN Elastic Net	5	10	1	1	7	6	7		7	3	10

Table V.10: Confusion Counts at the Optimal Threshold Across Models and Stocks

Finally, Table V.11 shows different metrics computed on the confusion counts, with their support-weighted average on the right of the table as described in the *Methodology* chapter. Some metrics have no value on some stock-model pair, being the result of zero denominators.

At the F2-optimal cutoffs, our primary metric, recall, is highest for Elastic Net (0.329), followed by EN-selected (0.318), Market-only (0.306), and All-variables (0.282). As expected, higher recall comes with lower precision. This is true for all models except the Market State model which scores towards the low end on both precision and recall. The difference is light but indicates slight preference for models that have selected their variables with care: the Elastic Net model and the EN-selected unpenalized model.

As expected, specificity is extremely high across the board, as the dataset was even more extremely imbalanced for the test set than the training set. As a direct effect, balanced accuracy minus 0.5 is simply proportional to recall (exactly recall divided by 2). All three recall-precision metrics, namely F1-score, F2-score and Fowlkes–Mallows index, highlight what has already been seen by looking at recall and precision individually: the three models including liquidity variables outperform the benchmark model without liquidity variables.



Test Set @ F2-Optimal Thresholds: Classification Metrics per Model & Stock

Metric	AAPL	AMZN	AVGO	COST	FB	GOOG	GOOGL	MSFT	NFLX	NVDA	TSLA	Support-Weighted Average
Recall All Vars.	0.000	0.474	0.000	0.000	0.273	0.000	0.000		0.364	0.000	0.421	0.282
Recall Market State Vars.	0.000	0.474	0.000	0.000	0.364	0.143	0.000		0.364	0.250	0.368	0.306
Recall Selected Vars.	0.000	0.474	0.000	0.000	0.273	0.143	0.000		0.364	0.000	0.526	0.318
Recall Elastic Net	0.000	0.474	0.000	0.000	0.364	0.143	0.000		0.364	0.250	0.474	0.329
Precision All Vars.	0.000	0.117	0.000	0.000	0.088	0.000	0.000		0.082	0.000	0.071	0.070
Precision Market State Vars.	0.000	0.107	0.000	0.000	0.035	0.038	0.000		0.075	0.014	0.064	0.051
Precision Selected Vars.	0.000	0.117	0.000	0.000	0.088	0.040	0.000		0.062	0.000	0.075	0.068
Precision Elastic Net		0.101	0.000	0.000	0.051	0.045	0.000		0.062	0.018	0.077	0.061
Specificity All Vars.	1.000	0.999	1.000	1.000	1.000	1.000	1.000		1.000	1.000	0.999	1.000
Specificity Market State Vars.	1.000	0.999	1.000	1.000	0.999	1.000	1.000		0.999	0.999	0.999	0.999
Specificity Selected Vars.	1.000	0.999	1.000	1.000	1.000	1.000	1.000		0.999	1.000	0.999	1.000
Specificity Elastic Net	1.000	0.999	1.000	1.000	0.999	1.000	1.000		0.999	0.999	0.999	1.000
F1 All Vars.		0.188			0.133				0.133		0.121	0.113
F1 Market State Vars.		0.175			0.065	0.061			0.125	0.027	0.109	0.088
F1 Selected Vars.		0.188			0.133	0.062			0.107		0.131	0.112
F1 Elastic Net		0.167			0.089	0.069			0.107	0.033	0.132	0.102
F2 All Vars.		0.294			0.192				0.215		0.212	0.176
F2 Market State Vars.		0.281			0.127	0.093			0.206	0.058	0.188	0.154
F2 Selected Vars.		0.294			0.192	0.094			0.185		0.238	0.183
F2 Elastic Net		0.273			0.163	0.100			0.185	0.068	0.233	0.175
Fowlkes–Mallows All Vars.	0.000	0.235	0.000	0.000	0.155	0.000	0.000		0.172	0.000	0.173	0.141
Fowlkes–Mallows Market State Vars.	0.000	0.225	0.000	0.000	0.113	0.074	0.000		0.166	0.060	0.153	0.125
Fowlkes–Mallows Selected Vars.	0.000	0.235	0.000	0.000	0.155	0.076	0.000		0.151	0.000	0.198	0.147
Fowlkes–Mallows Elastic Net		0.219	0.000	0.000	0.136	0.081	0.000		0.151	0.066	0.191	0.141
Balanced Accuracy All Vars.	0.500	0.736	0.500	0.500	0.636	0.500	0.500		0.682	0.500	0.710	0.641
Balanced Accuracy Market State Vars.	0.500	0.736	0.500	0.500	0.681	0.571	0.500		0.682	0.625	0.684	0.653
Balanced Accuracy Selected Vars.	0.500	0.736	0.500	0.500	0.636	0.571	0.500		0.681	0.500	0.762	0.659
Balanced Accuracy Elastic Net	0.500	0.736	0.500	0.500	0.681	0.571	0.500		0.681	0.625	0.736	0.664

Table V.11: Binary Classification Metrics at the Optimal Threshold Across Models and Stocks

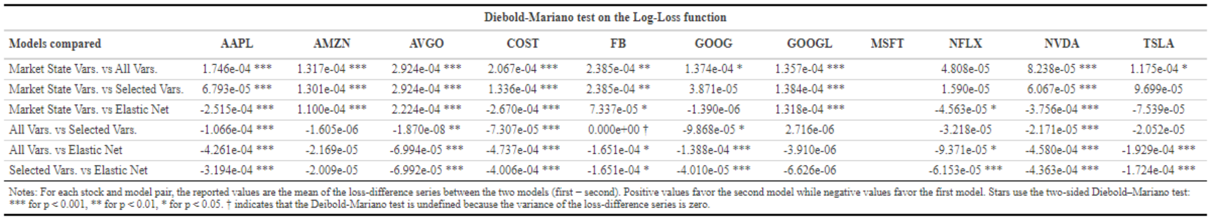
2.3.3 Forecast Comparison With the Diebold-Mariano Test

Table V.12 reports, for each stock and model pair, the mean log-loss difference $\bar{d}_s^{(m_1,m_2)}$ (i.e., mean of the log-loss of the first model minus the log-loss of the second) with the Diebold–Mariano p-values indicated by significance stars. Positive values mean the second model has lower expected log-loss, making it a better forecast model than the first one when tested on the test set.

The main conclusion coming from this table is the following: adding liquidity variables beats the Market State benchmark on probabilities calibration. Indeed, the All-variables and the EN-selected variables specifications beat the Market State model on every single stock and often significantly. However, the Elastic Net machine learning model is not a clear winner over the benchmark, with predicted probabilities being less strong. This makes sense when considering the very low F2-optimal thresholds of sub-section 2.3.2 *Test Set Binary Classification with an F2-optimized threshold*: with its penalized shrunk coefficients, the Elastic Net regression has a tendency to give lower probabilities of EPMs on average, leading to lower thresholds for classification but also higher losses in the Log-Loss function when an EPM happens. As a result, the unpenalized regression on EN-selected variables always beats its Elastic Net counterpart.

The remaining question is the following: which of the basic All-variables model and the meticulously EN-selected variables model has the best probabilities calibration. The answer is that the All-variables logistic regression typically attains a slightly lower log-loss than the EN-selected variables logistic regression, with small mean differences but leading to an often significant Diebold–Mariano test result.

To conclude, on strict probability calibration, liquidity-augmented models dominate the benchmark Market State model and, among them, the model including all variables is most often best, with the EN-selected variable model a close second.



Diebold-Mariano test on the Log-Loss function

Models compared	AAPL	AMZN	AVGO	COST	FB	GOOG	GOOGL	MSFT	NFLX	NVDA	TSLA
Market State Vars. vs All Vars.	1.746e-04 ***	1.317e-04 ***	2.924e-04 ***	2.067e-04 ***	2.385e-04 **	1.374e-04 *	1.357e-04 ***		4.808e-05	8.238e-05 ***	1.175e-04 *
Market State Vars. vs Selected Vars.	6.793e-05 ***	1.301e-04 ***	2.924e-04 ***	1.336e-04 ***	2.385e-04 **	3.871e-05	1.384e-04 ***		1.590e-05	6.067e-05 ***	9.699e-05
Market State Vars. vs Elastic Net	-2.515e-04 ***	1.100e-04 ***	2.224e-04 ***	-2.670e-04 ***	7.337e-05 *	-1.390e-06	1.318e-04 ***		-4.563e-05 *	-3.756e-04 ***	-7.539e-05
All Vars. vs Selected Vars.	-1.066e-04 ***	-1.605e-06	-1.870e-08 **	-7.307e-05 ***	0.000e+00 †	-9.868e-05 *	2.716e-06		-3.218e-05	-2.171e-05 ***	-2.052e-05
All Vars. vs Elastic Net	-4.261e-04 ***	-2.169e-05	-6.994e-05 ***	-4.737e-04 ***	-1.651e-04 *	-1.388e-04 ***	-3.910e-06		-9.371e-05 *	-4.580e-04 ***	-1.929e-04 ***
Selected Vars. vs Elastic Net	-3.194e-04 ***	-2.009e-05	-6.992e-05 ***	-4.006e-04 ***	-1.651e-04 *	-4.010e-05 ***	-6.626e-06		-6.153e-05 ***	-4.363e-04 ***	-1.724e-04 ***

Notes: For each stock and model pair, the reported values are the mean of the loss-difference series between the two models (first − second). Positive values favor the second model while negative values favor the first model. Stars use the two-sided Diebold–Mariano test: *** for p < 0.001, ** for p < 0.01, * for p < 0.05. † indicates that the Deibold-Mariano test is undefined because the variance of the loss-difference series is zero.

Table V.12: Diebold-Mariano Test Using the Log-Loss Function
Across all Model Pairs and Stocks

Chapter VI

Conclusion

This study examined intraday Extreme Price Movements (EPMs) through the lens of market microstructure, using high-frequency NASDAQ limit order book data to treat strong short-term price dislocations as liquidity events rather than purely volatility episodes. The first aim of the work was to determine whether liquidity information, covering supply-side LOB information and demand-side order flow aggressiveness, actually improves the short-horizon predictability of 10-second EPMs once the usual market-state conditions are already accounted for. In other words, the analysis asked whether predicted EPM probabilities meaningfully sharpen when liquidity signals are layered market-state variables such as time-of-day, short-horizon volatility, and trading intensity. Complementing that objective, the master thesis aimed to pinpoint which predictors matter most for flagging an imminent EPM. Rather than treating liquidity as a single concept, the study took a multi-dimensional view and looked at each dimension alongside the core market-state controls to identify variables that consistently add incremental information about EPM risk.

Together, these aims frame the contribution: evidence on the added value of liquidity variables for EPM prediction, and strong insights on the drivers that most strongly explain elevated EPM probabilities at the short-term horizon.

First, on incremental predictive power, liquidity-augmented logistic regression specifications generally outperform the market-state benchmark specification along three practical dimensions: parsimony-adjusted in-sample fit via the Bayesian Information Criterion, out-of-sample rare-event classification via a focus on the ability to detect as many EPMs as possible (recall focus), and out-of-sample probability forecast accuracy via Diebold–Mariano tests. From a threshold-free perspective (precision–recall under heavy class imbalance), adding liquidity improves discrimination, with the strongest classification performance achieved when liquidity is added parsimoniously rather than indiscriminately. When moving to actual decision-making with a recall-tilted cutoff, the same message holds: once liquidity signals are layered onto market-state signals, the models recover more true EPMs without precision collapsing too much, even though the operational F2-optimized thresholds are necessarily very small to enhance recall. When the predicted probabilities forecasted out-of-sample are judged, liquidity again helps, with Diebold-Mariano tests clearly showing a preference for the model including all variables, followed closely by the EN-selected unpenalized one. The trade-off between parsimony and raw fit is also clear. An appropriately selected set of liquidity measures added on top of market-state variables improves greatly model fit without being overly complex. Overall, the evidence underscores the added value of liquidity variables for EPM prediction, but it also highlights that interpretability, predictive stability, and parsimony are best preserved when these liquidity signals are incorporated selectively rather than by indiscriminately loading the model with collinear measures.

Second, on which predictors matter most, a coherent picture emerges of how EPM risk builds. A robust core emerges across stocks and specifications: recent absolute returns carry a strong positive association with imminent EPM risk; trading intensity likewise pushes risk up; quoted tightness rises ahead of price dislocations, with multi-level spread measures proving more informative than top-of-book snapshots; and execution pressure, measured by the fill rate, increases ahead of EPMs. These ingredients are selected almost always by the Elastic Net regression and remain influential in all multivariable models, which is consistent with the descriptive diagnostics showing widening spreads, elevated activity, and persistent pressure before EPMs. By contrast, limit order book depth, imbalances in the liquidity supply (with Limit Order Ratio and Limit Order Flow Imbalance as measures), and more elaborate constructs such as Kyle's λ and Probability of Informed Trading do not emerge as primary drivers at the 10-second horizon. Overall, the models place greater weight on rolling averages than on single-interval signals, underscoring the importance of persistence. Similarly, they extract more predictive value from aggregated measures spanning five levels of the order book than from a narrow, precise view of its visible extremity. One caveat is essential: in a multivariable setting with correlated predictors, coefficient signs are conditional and can flip relative to univariate intuition. The Median Realized Volatility measure is the prime example, often turning negative in regressions because its information is largely absorbed by simpler proxies like absolute returns and trading intensity. This highlights how multicollinearity makes coefficients harder to interpret and obscures and complicates the assessment of which predictors matter most.

Moreover, this study goes beyond answering the two core questions on EPM predictability, and maps how the signals co-move so they can be interpreted together, not in isolation. The descriptive analysis points to persistent regimes, with ACFs and CCFs between variable showing high significant values, meaning that windows of elevated risk are forecasted rather than precise timestamps. The correlation structure further clarifies four distinct but connected blocks of variables moving in the same direction at the same time. A prime illustration of the descriptive section added value is the joint behavior of Net Dollar Limit Order Volume (NDLOV) and Cancellation Rate (CR): when trading intensity and volatility rises, as well as strictly before EPMs, NDLOV tends to increase while CR tends to fall. Such a pattern indicates that, on average, liquidity providers and market makers increase their liquidity supply position when faced with more pressure from the liquidity demand side, which reinforces the views of Brogaard et al. (2018). Overall, this shows that considering the joint behavior of variables deepens the comprehension of the microstructure dynamics at play, and helps to confirm established literature evidence on their interrelations, such as the links between volatility, trading intensity, and the occurrence of EPMs.

The findings in this study give rise to a number of practical implications. For risk managers, the persistence of co-moving signals over short horizons implies that EPMs, or at least a heightened probability of price dislocations, can be flagged pre-emptively by monitoring a compact set of both liquidity and market-state indicators, such as short-window absolute returns, multi-level tightness, sustained trading intensity, and execution pressure. Monitoring indicators such as the ones used in this study supports a risk approach calibrated to windows of elevated risk, with temporary adjustments to risk limits, rather than point-in-time alarms. For execution desks, elevated spreads and persistent trading activity identify intervals where adverse selection and market impact are costlier; scheduling, order slicing, and aggressiveness can be modulated toward identified lower-risk windows, using indicators that show a more liquid and stable market. For exchanges and regulators, a central policy objective is to sustain the presence of market makers in order to mitigate EPM risk, even during intervals of heightened adverse-selection risk and order flow toxicity. The empirical results of this study do not place flow-toxicity metrics at the core of short-horizon prediction. Nevertheless, this study shows that they are significantly correlated with the tightness of the market, with spreads widening as toxicity rises. One practical risk-transfer tool was already proposed by Easley et al. (2011): an exchange could list a contract that pays more when adverse selection is high. Market makers could buy this contract to offset expected losses and at the same time remain in the market. Their initial idea was based on the VPIN metric as underlying, but the underlying metric could differ. The evidence of this study advocates for a more robust liquidity-market-state composite built from the signals that move reliably before EPMs.

Despite these contributions, this master thesis is subject to several limitations that should be acknowledged, leading to potential improvements for future work on the subject:

- First, the analysis is restricted to the ten largest NASDAQ stocks, mostly concentrated in the technology sector, which limits the generalizability of the findings to smaller-cap firms, other sectors, or different exchanges. Future work might focus on other exchange-traded stock types.

- Second, the study period covers only the year 2020, a year marked by extraordinary volatility due to the COVID-19 crisis. While this provides a rich stress-testing environment, results may not fully extend to calmer or structurally different market phases, even though the year 2020 cannot be reduced only to a stress episode. In addition, the robustness of the results is only partially tested, as the out-of-sample test is limited to the end of the 2020 year, leaving open the question of whether the models' predictive power holds over other periods. Expanding the temporal scope to include multiple years and diverse market regimes would also help assess the stability of liquidity-based predictors across environments.

- Multicollinearity among explanatory variables was often substantial, limiting the interpretability of some coefficients. Future work could address this issue by employing dimensionality-reduction techniques to mitigate collinearity while preserving interpretability.

- On the other hand, the variable set, while broad, is not exhaustive. Notably, measures such as PIN or Limit Order Flow Imbalance could have benefited largely from being implemented with volume-based sampling. Future research could explore exclusively volume-synchronized sampling schemes, which may yield more coherent and robust predictors.

- Lastly, although logistic regression with regularization proved to be useful in variable selection, it remains a relatively simple modeling approach that may not fully capture the complex, nonlinear dynamics of liquidity and market-state variables together to predict EPMs. Looking forward, an important direction lies in developing more sophisticated yet interpretable models. Approaches such as random forests, gradient boosting methods, or neural networks with explainability techniques could provide richer insights while avoiding the opacity of pure 'black-box' approaches.

With these limitations in mind, the central finding of this master thesis remains the following: liquidity variables meaningfully enhance the prediction of intraday extreme price movements, particularly when incorporated selectively and interpreted alongside market-state conditions. The results demonstrate that EPM risk is not only random noise, but reflects coherent liquidity and market-state dynamics, underscoring the potential of microstructure-based prediction to enhance decision-making in financial markets.

Appendices

Appendix A – Cross-Correlation Functions

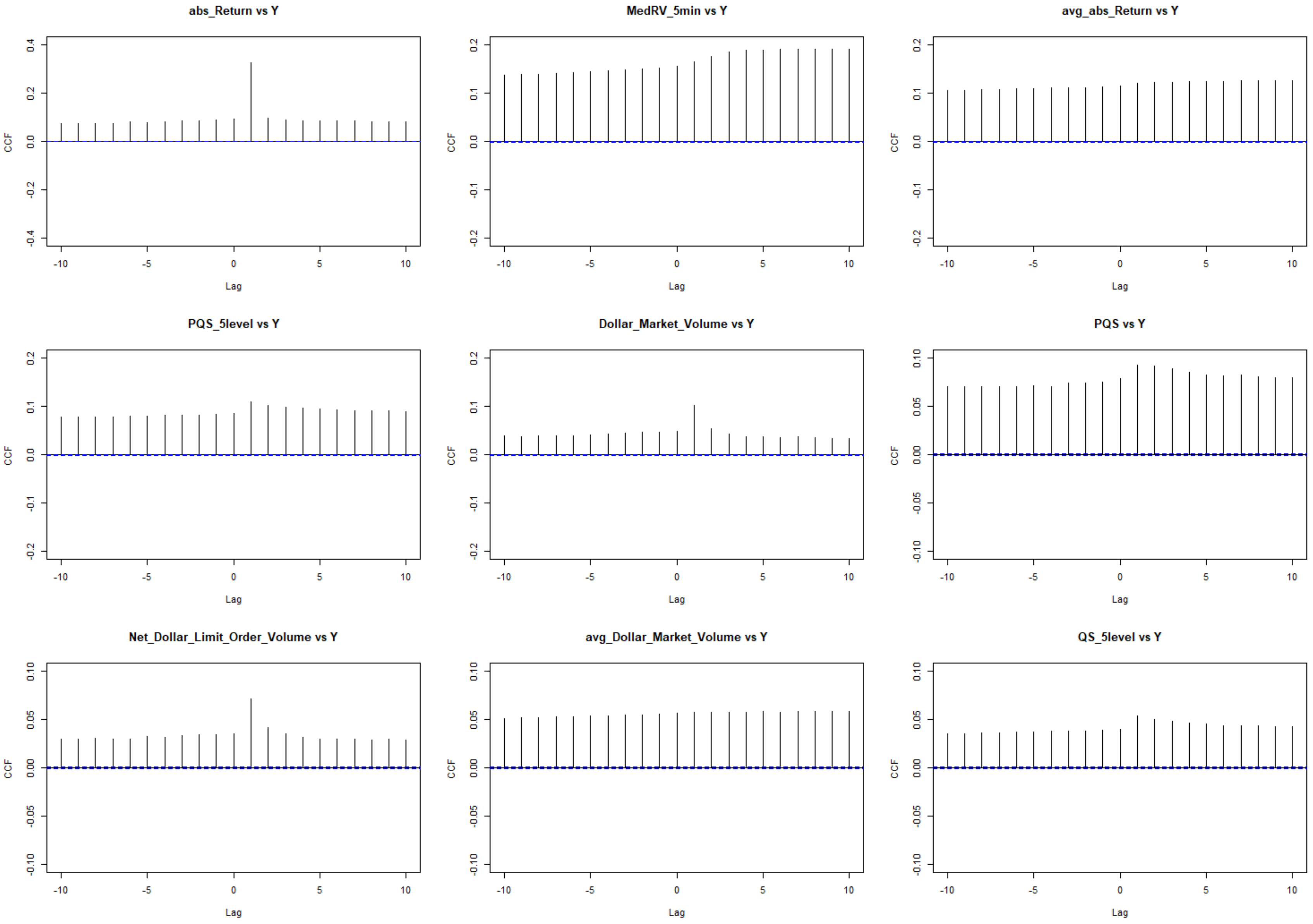


Figure A.1: CCFs Between all Independent Variables and the Dependent Variable (part 1)

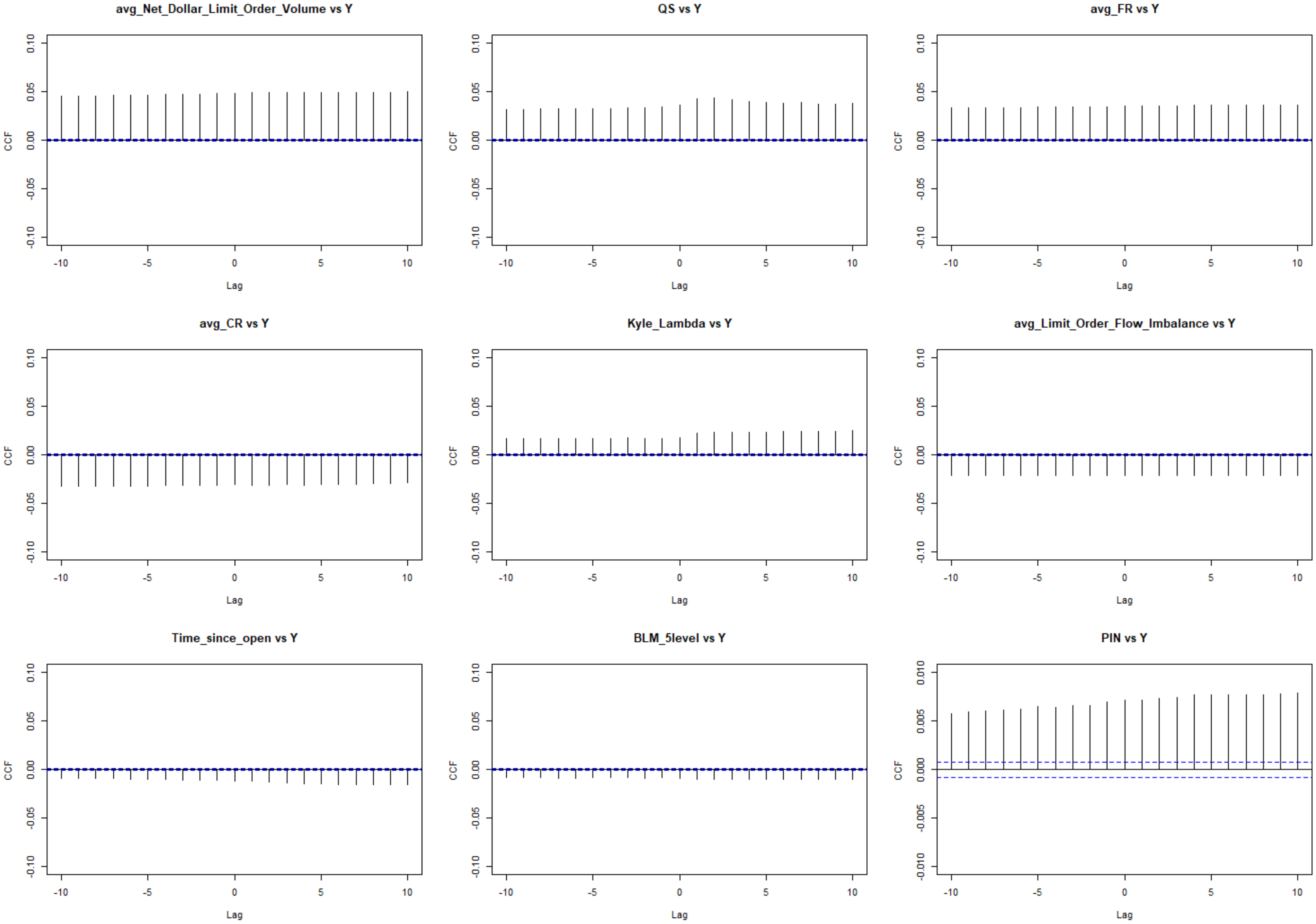


Figure A.2: CCFs Between all Independent Variables and the Dependent Variable (part 2)

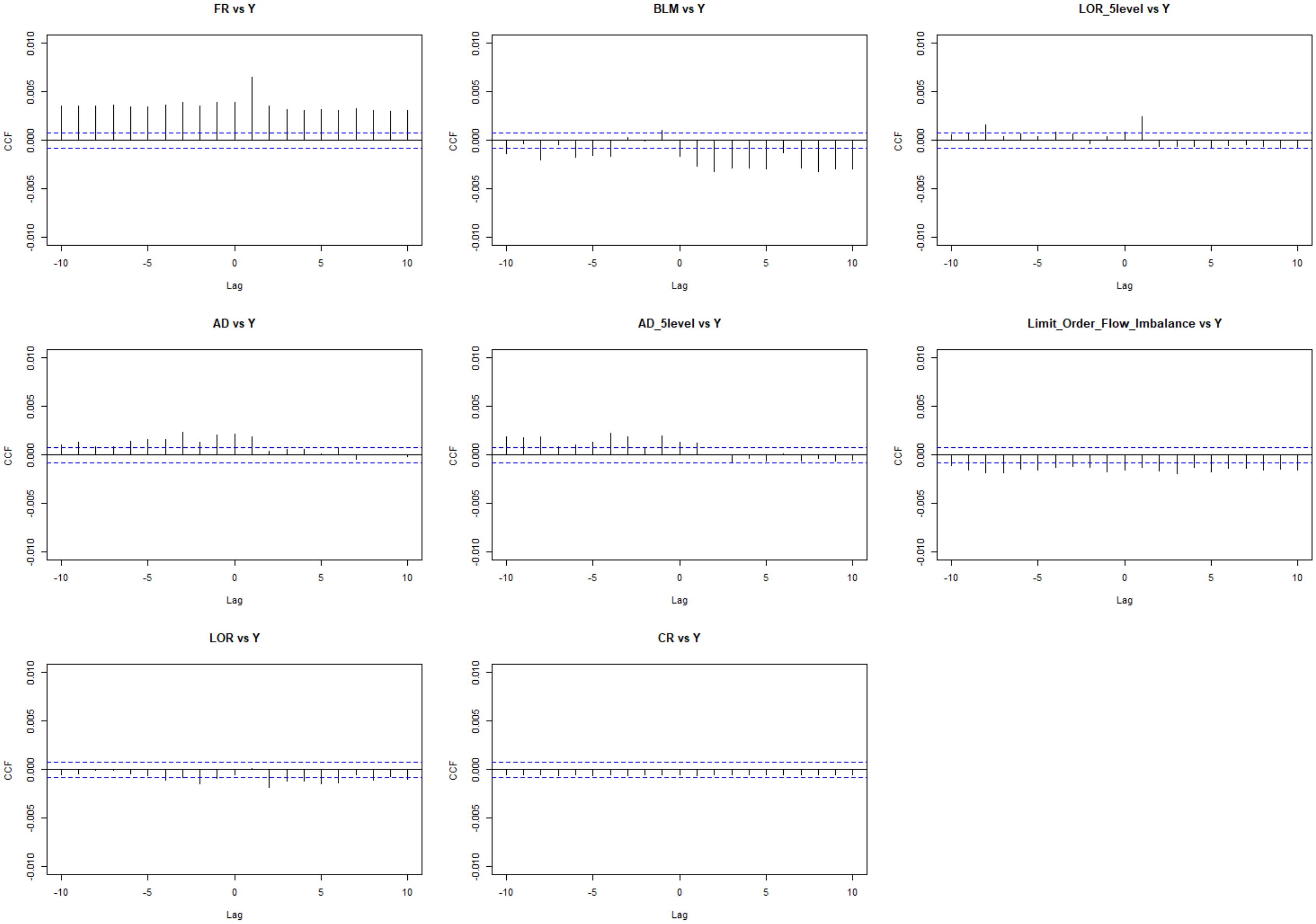
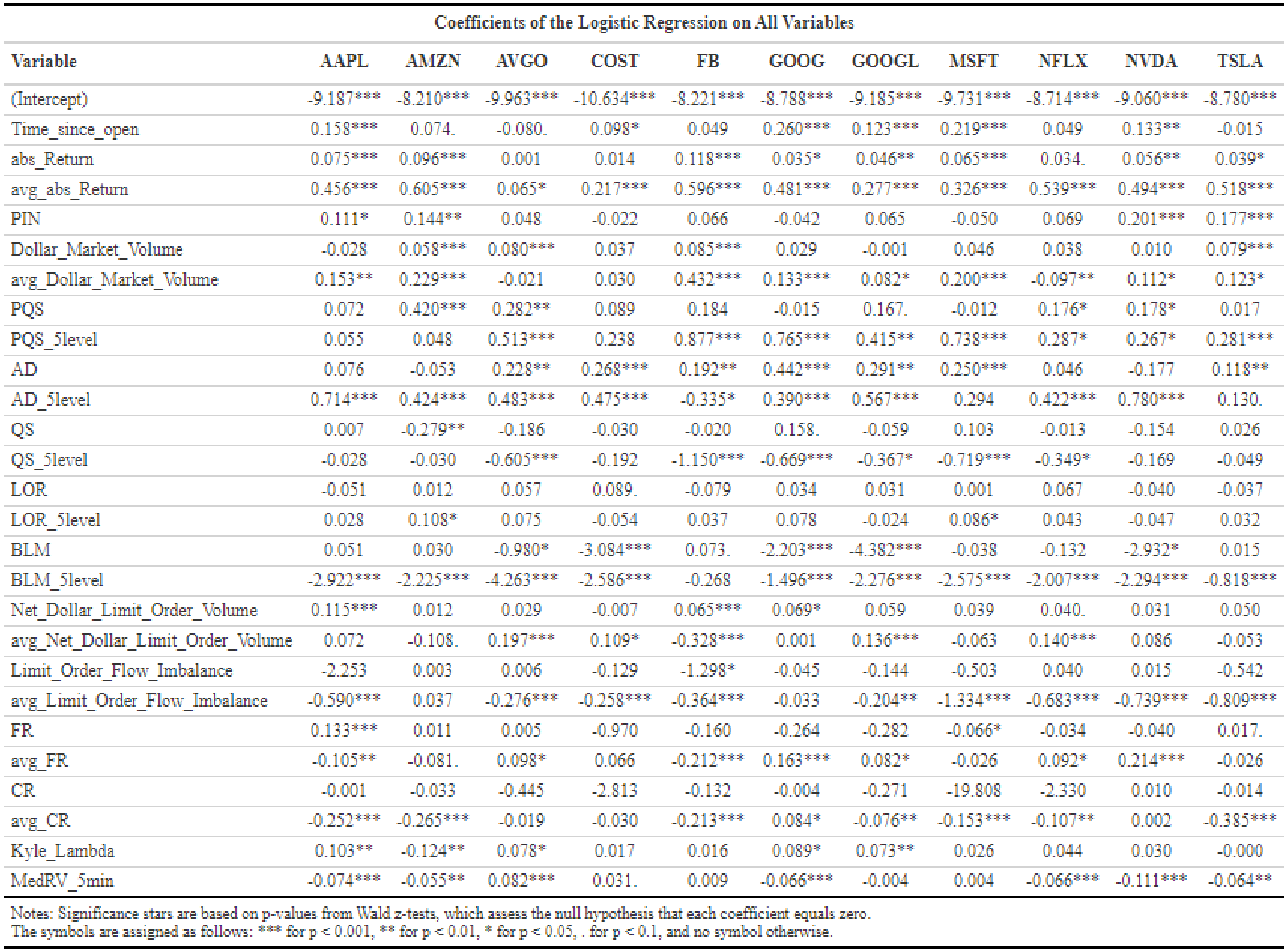


Figure A.3: CCFs Between all Independent Variables and the Dependent Variable (part 3)

Appendix B – Coefficients of the Four Logistic Regression Models

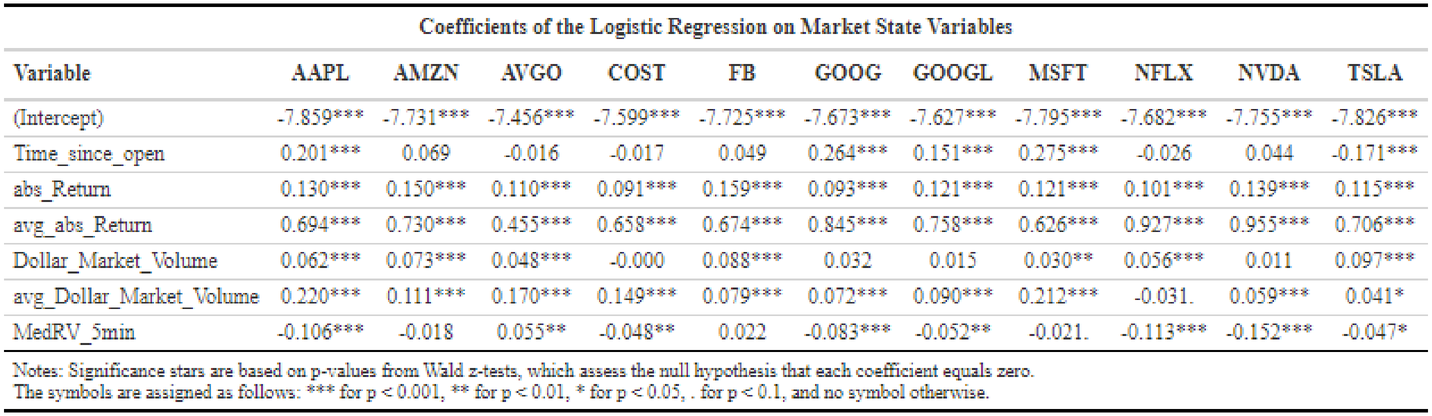


Coefficients of the Logistic Regression on All Variables

Variable	AAPL	AMZN	AVGO	COST	FB	GOOG	GOOGL	MSFT	NFLX	NVDA	TSLA
(Intercept)	-9.187***	-8.210***	-9.963***	-10.634***	-8.221***	-8.788***	-9.185***	-9.731***	-8.714***	-9.060***	-8.780***
Time_since_open	0.158***	0.074.	-0.080.	0.098*	0.049	0.260***	0.123***	0.219***	0.049	0.133**	-0.015
abs_Return	0.075***	0.096***	0.001	0.014	0.118***	0.035*	0.046**	0.065***	0.034.	0.056**	0.039*
avg_abs_Return	0.456***	0.605***	0.065*	0.217***	0.596***	0.481***	0.277***	0.326***	0.539***	0.494***	0.518***
PIN	0.111*	0.144**	0.048	-0.022	0.066	-0.042	0.065	-0.050	0.069	0.201***	0.177***
Dollar_Market_Volume	-0.028	0.058***	0.080***	0.037	0.085***	0.029	-0.001	0.046	0.038	0.010	0.079***
avg_Dollar_Market_Volume	0.153**	0.229***	-0.021	0.030	0.432***	0.133***	0.082*	0.200***	-0.097**	0.112*	0.123*
PQS	0.072	0.420***	0.282**	0.089	0.184	-0.015	0.167.	-0.012	0.176*	0.178*	0.017
PQS_5level	0.055	0.048	0.513***	0.238	0.877***	0.765***	0.415**	0.738***	0.287*	0.267*	0.281***
AD	0.076	-0.053	0.228**	0.268***	0.192**	0.442***	0.291**	0.250***	0.046	-0.177	0.118**
AD_5level	0.714***	0.424***	0.483***	0.475***	-0.335*	0.390***	0.567***	0.294	0.422***	0.780***	0.130.
QS	0.007	-0.279**	-0.186	-0.030	-0.020	0.158.	-0.059	0.103	-0.013	-0.154	0.026
QS_5level	-0.028	-0.030	-0.605***	-0.192	-1.150***	-0.669***	-0.367*	-0.719***	-0.349*	-0.169	-0.049
LOR	-0.051	0.012	0.057	0.089.	-0.079	0.034	0.031	0.001	0.067	-0.040	-0.037
LOR_5level	0.028	0.108*	0.075	-0.054	0.037	0.078	-0.024	0.086*	0.043	-0.047	0.032
BLM	0.051	0.030	-0.980*	-3.084***	0.073.	-2.203***	-4.382***	-0.038	-0.132	-2.932*	0.015
BLM_5level	-2.922***	-2.225***	-4.263***	-2.586***	-0.268	-1.496***	-2.276***	-2.575***	-2.007***	-2.294***	-0.818***
Net_Dollar_Limit_Order_Volume	0.115***	0.012	0.029	-0.007	0.065***	0.069*	0.059	0.039	0.040.	0.031	0.050
avg_Net_Dollar_Limit_Order_Volume	0.072	-0.108.	0.197***	0.109*	-0.328***	0.001	0.136***	-0.063	0.140***	0.086	-0.053
Limit_Order_Flow_Imbalance	-2.253	0.003	0.006	-0.129	-1.298*	-0.045	-0.144	-0.503	0.040	0.015	-0.542
avg_Limit_Order_Flow_Imbalance	-0.590***	0.037	-0.276***	-0.258***	-0.364***	-0.033	-0.204**	-1.334***	-0.683***	-0.739***	-0.809***
FR	0.133***	0.011	0.005	-0.970	-0.160	-0.264	-0.282	-0.066*	-0.034	-0.040	0.017.
avg_FR	-0.105**	-0.081.	0.098*	0.066	-0.212***	0.163***	0.082*	-0.026	0.092*	0.214***	-0.026
CR	-0.001	-0.033	-0.445	-2.813	-0.132	-0.004	-0.271	-19.808	-2.330	0.010	-0.014
avg_CR	-0.252***	-0.265***	-0.019	-0.030	-0.213***	0.084*	-0.076**	-0.153***	-0.107**	0.002	-0.385***
Kyle_Lambda	0.103**	-0.124**	0.078*	0.017	0.016	0.089*	0.073**	0.026	0.044	0.030	-0.000
MedRV_5min	-0.074***	-0.055**	0.082***	0.031.	0.009	-0.066***	-0.004	0.004	-0.066***	-0.111***	-0.064**

Notes: Significance stars are based on p-values from Wald z-tests, which assess the null hypothesis that each coefficient equals zero.
The symbols are assigned as follows: *** for p < 0.001, ** for p < 0.01, * for p < 0.05, . for p < 0.1, and no symbol otherwise.

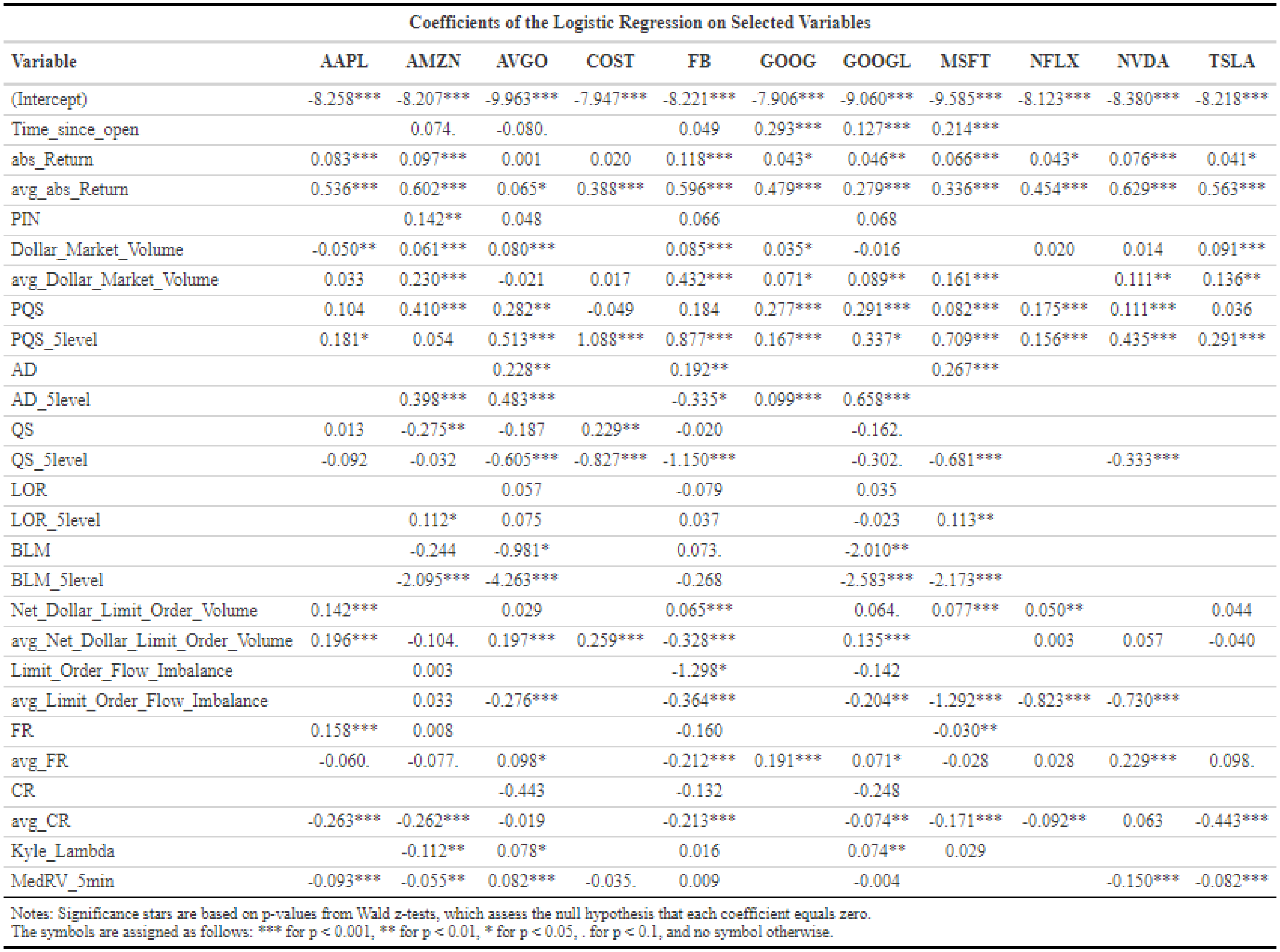
Table B.1: Unpenalized Logistic Regression Coefficients on the Model Comprising All Variables



Coefficients of the Logistic Regression on Market State Variables											
Variable	AAPL	AMZN	AVGO	COST	FB	GOOG	GOOGL	MSFT	NFLX	NVDA	TSLA
(Intercept)	-7.859***	-7.731***	-7.456***	-7.599***	-7.725***	-7.673***	-7.627***	-7.795***	-7.682***	-7.755***	-7.826***
Time_since_open	0.201***	0.069	-0.016	-0.017	0.049	0.264***	0.151***	0.275***	-0.026	0.044	-0.171***
abs_Return	0.130***	0.150***	0.110***	0.091***	0.159***	0.093***	0.121***	0.121***	0.101***	0.139***	0.115***
avg_abs_Return	0.694***	0.730***	0.455***	0.658***	0.674***	0.845***	0.758***	0.626***	0.927***	0.955***	0.706***
Dollar_Market_Volume	0.062***	0.073***	0.048***	-0.000	0.088***	0.032	0.015	0.030**	0.056***	0.011	0.097***
avg_Dollar_Market_Volume	0.220***	0.111***	0.170***	0.149***	0.079***	0.072***	0.090***	0.212***	-0.031.	0.059***	0.041*
MedRV_5min	-0.106***	-0.018	0.055**	-0.048**	0.022	-0.083***	-0.052**	-0.021.	-0.113***	-0.152***	-0.047*

Notes: Significance stars are based on p-values from Wald z-tests, which assess the null hypothesis that each coefficient equals zero.
The symbols are assigned as follows: *** for p < 0.001, ** for p < 0.01, * for p < 0.05, . for p < 0.1, and no symbol otherwise.

Table B.2: Unpenalized Logistic Regression Coefficients
on the Model Comprising Market State Variables

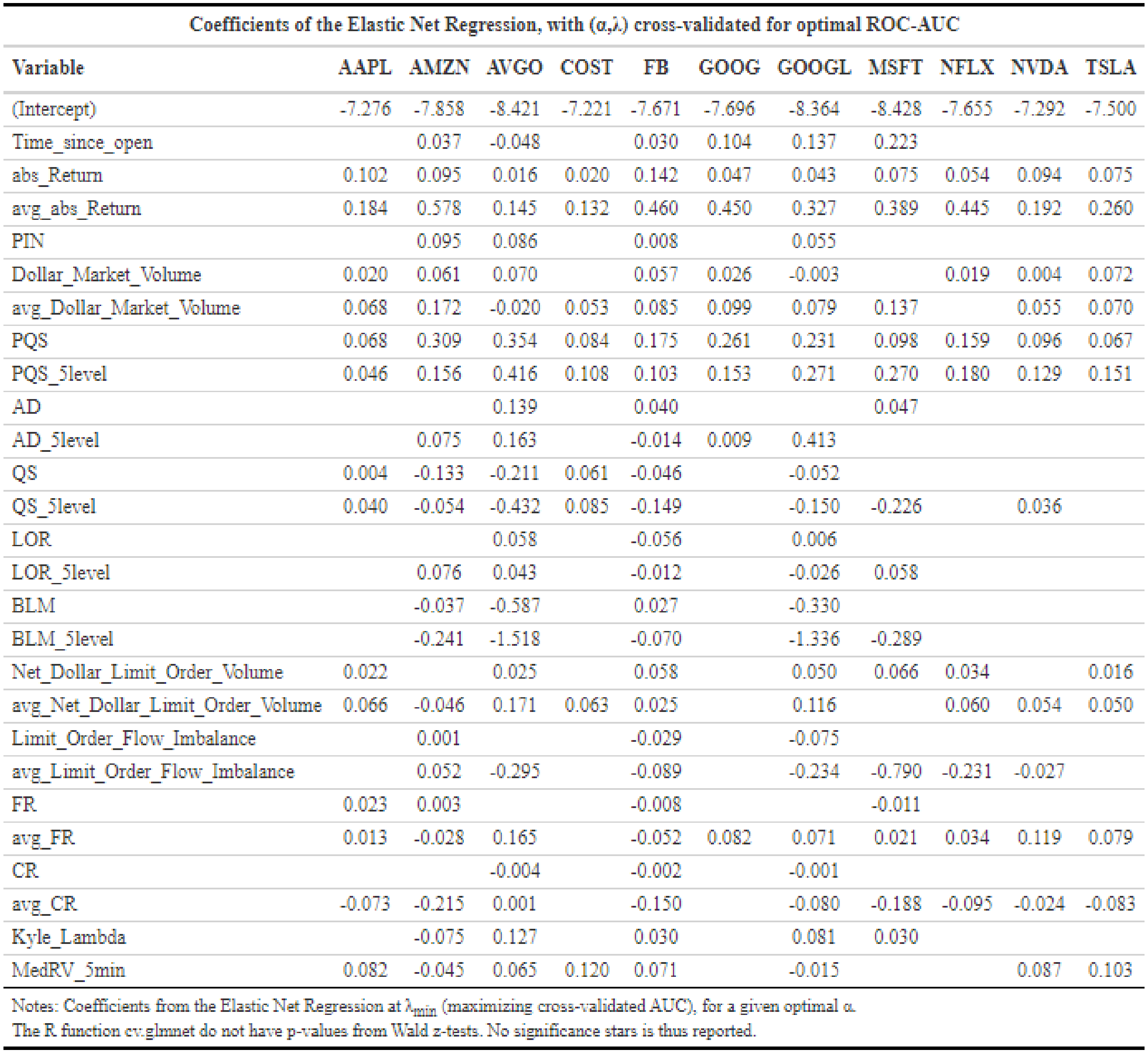


Coefficients of the Logistic Regression on Selected Variables

Variable	AAPL	AMZN	AVGO	COST	FB	GOOG	GOOGL	MSFT	NFLX	NVDA	TSLA
(Intercept)	-8.258***	-8.207***	-9.963***	-7.947***	-8.221***	-7.906***	-9.060***	-9.585***	-8.123***	-8.380***	-8.218***
Time_since_open		0.074.	-0.080.		0.049	0.293***	0.127***	0.214***			
abs_Return	0.083***	0.097***	0.001	0.020	0.118***	0.043*	0.046**	0.066***	0.043*	0.076***	0.041*
avg_abs_Return	0.536***	0.602***	0.065*	0.388***	0.596***	0.479***	0.279***	0.336***	0.454***	0.629***	0.563***
PIN		0.142**	0.048		0.066		0.068				
Dollar_Market_Volume	-0.050**	0.061***	0.080***		0.085***	0.035*	-0.016		0.020	0.014	0.091***
avg_Dollar_Market_Volume	0.033	0.230***	-0.021	0.017	0.432***	0.071*	0.089**	0.161***		0.111**	0.136**
PQS	0.104	0.410***	0.282**	-0.049	0.184	0.277***	0.291***	0.082***	0.175***	0.111***	0.036
PQS_5level	0.181*	0.054	0.513***	1.088***	0.877***	0.167***	0.337*	0.709***	0.156***	0.435***	0.291***
AD			0.228**		0.192**			0.267***			
AD_5level		0.398***	0.483***		-0.335*	0.099***	0.658***				
QS	0.013	-0.275**	-0.187	0.229**	-0.020		-0.162.				
QS_5level	-0.092	-0.032	-0.605***	-0.827***	-1.150***		-0.302.	-0.681***		-0.333***	
LOR			0.057		-0.079		0.035				
LOR_5level		0.112*	0.075		0.037		-0.023	0.113**			
BLM		-0.244	-0.981*		0.073.		-2.010**				
BLM_5level		-2.095***	-4.263***		-0.268		-2.583***	-2.173***			
Net_Dollar_Limit_Order_Volume	0.142***		0.029		0.065***		0.064.	0.077***	0.050**		0.044
avg_Net_Dollar_Limit_Order_Volume	0.196***	-0.104.	0.197***	0.259***	-0.328***		0.135***		0.003	0.057	-0.040
Limit_Order_Flow_Imbalance		0.003			-1.298*		-0.142				
avg_Limit_Order_Flow_Imbalance		0.033	-0.276***		-0.364***		-0.204**	-1.292***	-0.823***	-0.730***	
FR	0.158***	0.008			-0.160			-0.030**			
avg_FR	-0.060.	-0.077.	0.098*		-0.212***	0.191***	0.071*	-0.028	0.028	0.229***	0.098.
CR			-0.443		-0.132		-0.248				
avg_CR	-0.263***	-0.262***	-0.019		-0.213***		-0.074**	-0.171***	-0.092**	0.063	-0.443***
Kyle_Lambda		-0.112**	0.078*		0.016		0.074**	0.029			
MedRV_5min	-0.093***	-0.055**	0.082***	-0.035.	0.009		-0.004			-0.150***	-0.082***

Notes: Significance stars are based on p-values from Wald z-tests, which assess the null hypothesis that each coefficient equals zero.
The symbols are assigned as follows: *** for p < 0.001, ** for p < 0.01, * for p < 0.05, . for p < 0.1, and no symbol otherwise.

Table B.3: Unpenalized Logistic Regression Coefficients on the Model Comprising Elastic-Net-Selected Variables



Coefficients of the Elastic Net Regression, with (α,λ) cross-validated for optimal ROC-AUC

Variable	AAPL	AMZN	AVGO	COST	FB	GOOG	GOOGL	MSFT	NFLX	NVDA	TSLA
(Intercept)	-7.276	-7.858	-8.421	-7.221	-7.671	-7.696	-8.364	-8.428	-7.655	-7.292	-7.500
Time_since_open		0.037	-0.048		0.030	0.104	0.137	0.223			
abs_Return	0.102	0.095	0.016	0.020	0.142	0.047	0.043	0.075	0.054	0.094	0.075
avg_abs_Return	0.184	0.578	0.145	0.132	0.460	0.450	0.327	0.389	0.445	0.192	0.260
PIN		0.095	0.086		0.008		0.055				
Dollar_Market_Volume	0.020	0.061	0.070		0.057	0.026	-0.003		0.019	0.004	0.072
avg_Dollar_Market_Volume	0.068	0.172	-0.020	0.053	0.085	0.099	0.079	0.137		0.055	0.070
PQS	0.068	0.309	0.354	0.084	0.175	0.261	0.231	0.098	0.159	0.096	0.067
PQS_5level	0.046	0.156	0.416	0.108	0.103	0.153	0.271	0.270	0.180	0.129	0.151
AD			0.139		0.040			0.047			
AD_5level		0.075	0.163		-0.014	0.009	0.413				
QS	0.004	-0.133	-0.211	0.061	-0.046		-0.052				
QS_5level	0.040	-0.054	-0.432	0.085	-0.149		-0.150	-0.226		0.036	
LOR			0.058		-0.056		0.006				
LOR_5level		0.076	0.043		-0.012		-0.026	0.058			
BLM		-0.037	-0.587		0.027		-0.330				
BLM_5level		-0.241	-1.518		-0.070		-1.336	-0.289			
Net_Dollar_Limit_Order_Volume	0.022		0.025		0.058		0.050	0.066	0.034		0.016
avg_Net_Dollar_Limit_Order_Volume	0.066	-0.046	0.171	0.063	0.025		0.116		0.060	0.054	0.050
Limit_Order_Flow_Imbalance		0.001			-0.029		-0.075				
avg_Limit_Order_Flow_Imbalance		0.052	-0.295		-0.089		-0.234	-0.790	-0.231	-0.027	
FR	0.023	0.003			-0.008			-0.011			
avg_FR	0.013	-0.028	0.165		-0.052	0.082	0.071	0.021	0.034	0.119	0.079
CR			-0.004		-0.002		-0.001				
avg_CR	-0.073	-0.215	0.001		-0.150		-0.080	-0.188	-0.095	-0.024	-0.083
Kyle_Lambda		-0.075	0.127		0.030		0.081	0.030			
MedRV_5min	0.082	-0.045	0.065	0.120	0.071		-0.015			0.087	0.103

Notes: Coefficients from the Elastic Net Regression at λ_{min} (maximizing cross-validated AUC), for a given optimal α.
The R function cv.glmnet do not have p-values from Wald z-tests. No significance stars is thus reported.

Table B.4: Elastic Net Logistic Regression Coefficients

References

Aitken, M., & Comerton-Forde, C. (2003). How should liquidity be measured? *Pacific-Basin Finance Journal, 11*(1), 45–59. https://doi.org/10.1016/S0927-538X(02)00093-8

Amihud, Y. (2002). Illiquidity and stock returns: Cross-section and time-series effects. *Journal of Financial Markets, 5*(1), 31–56. https://doi.org/10.1016/S1386-4181(01)00024-6

Andersen, T. G., Bollerslev, T., Diebold, F. X., & Ebens, H. (2001). The distribution of realized stock return volatility. *Journal of Financial Economics, 61*(1), 43–76. https://doi.org/10.1016/S0304-405X(01)00055-1

Andersen, T. G., & Bondarenko, O. (2014). VPIN and the flash crash. *Journal of Financial Markets, 17*, 1–46. https://doi.org/10.1016/j.finmar.2013.05.005

Andersen, T. G., Dobrev D., & Schaumburg, E. (2012). Jump-robust volatility estimation using nearest neighbor truncation. *Journal of Econometrics, 169*(1), 75–93.https://doi.org/10.1016/j.jeconom.2012.01.011

Back, K., & Baruch, S. (2004). Information in securities markets: Kyle meets Glosten and Milgrom. *Econometrica, 72*(2), 433–465. https://doi.org/10.1111/j.1468-0262.2004.00497.x

Bandi, F. M., Russell, J. R., & Yang, C. (2013). Realized volatility forecasting in the presence of time-varying noise. *Journal of Business & Economic Statistics, 31*(3), 331–345. https://doi.org/10.1080/07350015.2013.803866

Brennan, M. J., Huh, S.-W., & Subrahmanyam, A. (2018). High-frequency measures of informed trading and corporate announcements. *The Review of Financial Studies, 31*(6), 2326–2376. `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2578168`

Brogaard, J., Carrion, A., Moyaert, T., Riordan, R., Shkilko, A., & Sokolov, K. (2018). High-frequency trading and extreme price movements. *Journal of Financial Economics, 128*(2), 253–265. `https://doi.org/10.1016/j.jfineco.2018.02.002`

Brogaard, J., Hendershott, T., & Riordan, R. (2014). High-frequency trading and price discovery. *The Review of Financial Studies, 27*(8), 2267–2306. `https://doi.org/10.1093/rfs/hhu032`

Brogaard, J., Hendershott, T., & Riordan, R. (2019). Price discovery without trading: Evidence from limit orders. *The Journal of Finance, 74*(4), 1621–1658. `https://DOI:10.1111/jofi.12769`

Brolley, M., & Cimon, D. A. (2020). Order-flow segmentation, liquidity, and price discovery: The role of latency delays. *Journal of Financial & Quantitative Analysis, 55*(8), 2555–2587. `https://doi.org/10.1017/S002210901900067X`

Brownlees, C. T., & Gallo, G. M. (2006). Financial econometric analysis at ultra-high frequency: Data handling concerns. *Computational Statistics & Data Analysis, 51*(4), 2232–2245. `https://doi.org/10.1016/j.csda.2006.09.030`

Cobandag Guloglu, Z., & Ekinci, C. (2022). Liquidity measurement: A comparative review of the literature with a focus on high frequency. *Journal of Economic Surveys, 36*(1), 41–74. `https://doi.org/10.1111/joes.12440`

Deuskar, P., & Johnson, T. C. (2011). Market liquidity and flow-driven risk. *Review of Financial Studies, 24*(3), 721–753. `https://doi.org/10.1093/rfs/hhq132`

Diebold, F. X., & Mariano, R. S. (2002). Comparing Predictive Accuracy. *Journal of Business & Economic Statistics, 20*(1), 134--144. `https://doi.org/10.1198/073500102753410444`

Easley, D., López de Prado, M. M., & O'Hara, M. (2011). The microstructure of the "Flash Crash": Flow toxicity, liquidity crashes, and the probability of informed trading. *Journal of*

Portfolio Management, 37(2), 118–128. https://doi.org/10.3905/jpm.2011.37.2.118

Easley, D., López de Prado, M. M., & O'Hara, M. (2012). Flow toxicity and liquidity in a high-frequency world. *The Review of Financial Studies, 25*(5), 1457–1493. https://doi.org/10.1093/rfs/hhs053

Easley, D., López de Prado, M. M., & O'Hara, M. (2014). VPIN and the flash crash: A rejoinder. *Journal of Financial Markets, 17*, 47–52. https://doi.org/10.1016/j.finmar.2013.06.007

Easley, D., López de Prado, M. M., O'Hara, M., & Zhang, Z. (2021). Microstructure in the machine age. *The Review of Financial Studies, 34*(7), 3316–3363. https://doi.org/10.1093/rfs/hhaa078

Easley, D. & O'Hara, M. (1992). Time and the Process of Security Price Adjustment. *The Journal of Finance, 47*(2), 577–605. https://doi.org/10.1111/j.1540-6261.1992.tb04402.x

Engle, R. F. (2000). The econometrics of ultra-high-frequency data. *Econometrica, 68*(1), 1–22. https://doi.org/10.1111/1468-0262.00091

Engle, R. F., & Lange, J. (2001). Predicting VNET: A model of the dynamics of market depth. *Journal of Financial Markets, 4*(2), 113–142. https://doi.org/10.1016/S1386-4181(00)00019-7

Engle, R. F., & Russell, J. R. (1998). Autoregressive conditional duration: A new model for irregularly spaced transaction data. *Econometrica, 66*(5), 1127–1162. https://doi.org/10.2307/2999632

Fernandes, N., & Ferreira, M. A. (2009). Insider trading laws and stock price informativeness. *Review of Financial Studies, 22*(5), 1845–1887. https://doi.org/10.1093/rfs/hhn066

Glosten, L. R., & Milgrom, P. R. (1985). Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. *Journal of Financial Economics, 14*(1), 71–100.

https://doi.org/10.1016/0304-405X(85)90044-3

Hasbrouck, J., & Seppi, D. J. (2001). Common factors in prices, order flows, and liquidity. *Journal of Financial Economics, 59*(3), 383–411. https://doi.org/10.1016/S0304-405X(00)00091-X

Hautsch, N. (2012). *Econometrics of financial high-frequency data.* Springer. https://doi.org/10.1007/978-3-642-21925-2

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*(509), 107–122. https://doi.org/10.1080/01621459.2014.982278

Jacod, J., Li, Y., Mykland, P. A., Podolskij, M., & Vetter, M. (2009). Microstructure noise in the continuous case: The pre-averaging approach. *Stochastic Processes and Their Applications, 119*(7), 2249–2276. https://doi.org/10.1016/j.spa.2008.11.004

Jurkatis, S. (2022). Inferring trade directions in fast markets. *Journal of Financial Markets, 58*, 100635. https://doi.org/10.1016/j.finmar.2021.100635

Kyle, A. S. (1985). Continuous auctions and insider trading. *Econometrica, 53*(6), 1315–1335. https://doi.org/10.2307/1913210

Lee, S. S., & Mykland, P. A. (2012). Jumps in equilibrium prices and market microstructure noise. *Journal of Econometrics, 168*(2), 396–406. https://doi.org/10.1016/j.jeconom.2012.03.001

Liu, Y., & Yang, X. (2017). Asymmetric synchronicity in extreme stock price movements: Evidence from China's stock market. *Procedia Computer Science, 122*, 1156–1161. https://doi.org/10.1016/j.procs.2017.11.486

Love, R., & Payne, R. (2008). Macroeconomic news, order flows, and exchange rates. *Journal of Financial and Quantitative Analysis, 43*(2), 467–488. https://doi.org/10.1017/S0022109000003598

Marshall, B. R. (2006). Liquidity and stock returns: Evidence from a pure order-driven market using a new liquidity proxy. *International Review of Financial Analysis, 15*(1), 21–38. https://doi.org/10.1016/j.irfa.2004.09.001

McInish, T. H., & Wood, R. A. (1992). An analysis of intraday patterns in bid/ask spreads for NYSE stocks. *The Journal of Finance, 47*(2), 753–764. https://doi.org/10.2307/2329122

O'Hara, M. (2003). Presidential address: Liquidity and price discovery. *The Journal of Finance, 58*(4), 1335–1354. https://doi.org/10.1111/1540-6261.00569

Ranaldo, A. (2000). *Intraday Trading Activity on Financial Markets: The Swiss evidence* [Doctoral dissertation]. University of Fribourg. https://folia.unifr.ch/unifr/documents/299756

Rif, A., & Utz, S. (2021). Short-term stock price reversals after extreme downward price movements. *The Quarterly Review of Economics and Finance, 81*, 123–133. https://doi.org/10.1016/j.qref.2021.05.004

Riordan, R., Storkenmaier, A., Wagener, M., & Zhang, S. (2013). Public information arrival: Price discovery and liquidity in electronic limit order markets. *Journal of Banking & Finance, 37*(4), 1148–1159. https://doi.org/10.1016/j.jbankfin.2012.11.008

Tookes, H. E. (2008). Information, trading, and product market interactions: Cross-sectional implications of informed trading. *The Journal of Finance, 63*(1), 379–413. https://doi.org/10.1111/j.1540-6261.2008.01319.x

Wu, L., Liu, H., Liu, C., & Long, Y. (2020). Determining the information share of liquidity and order flows in extreme price movements. *Economic Modelling, 93*, 559–575. https://doi.org/10.1016/j.econmod.2020.09.014

Zhang, L., Mykland, P. A., & Aït-Sahalia, Y. (2005). A tale of two time scales: Determining integrated volatility with noisy high-frequency data. *Journal of the American Statistical Association, 100*(472), 1394–1411. https://doi.org/10.1198/016214505000000169

Zhou, B. (1996). High-frequency data and volatility in foreign-exchange rates. *Journal of Business & Economic Statistics, 14*(1), 45–52. https://doi.org/10.2307/1392098

Executive Summary

This master thesis examines intraday Extreme Price Movements (EPMs) as liquidity-driven phenomena and asks two questions: which signals most reliably flag an imminent 10-second EPM, and whether liquidity-based information adds incremental predictive power beyond standard market-state controls. Using LOBSTER limit order book data for the ten largest NASDAQ-listed equities, we reconstruct prices on a 10-second grid, label EPMs at the 99.9$^{\text{th}}$ percentile of the distribution of absolute mid-quote returns, and estimate parsimonious logistic models that layer liquidity signals on top of market-state predictors. Elastic Net and unpenalized logistic regression specifications are evaluated both in-sample and out-of-sample, with January to October 2020 used for model estimation and November to December 2020 for model evaluation.

We find that selectively incorporating a compact set of liquidity signals consistently improves short-horizon EPM detection and probability calibration relative to the market-state benchmark. Across assets and specifications, the most reliable precursors to EPMs are: recent absolute returns as a volatility measure, sustained trading intensity, multi-level spread deterioration (loss of market tightness), and execution pressure captured by fill rates. Together, these results indicate that EPM risk is not mere noise but the joint outcome of market-state conditions and liquidity dynamics. These indicators enable better, real-time detection of windows of heightened EPM probability, supporting better-informed risk management for practitioners, regulators, and market participants.

Keywords: Financial Markets, Extreme Price Movements (EPMs), Market Microstructure, Liquidity, High-Frequency Data, Limit Order Book.

Word count: 27,797