Erasmus School of Economics

Master Thesis Quantitative Finance

Unlocking Intraday Alpha Signals: A Novel Methodology for Equity Trade Execution

Author:
Stef Schrijer *(579509)*

Supervisor:
dr. Rasmus Lonn
Second assessor:
dr. Alberto Quaini

April 30, 2024

Abstract

This thesis explores the use of intraday alpha signals to enhance the execution of intraday equity trades, focusing on the cross-section of stock returns and employing ordinal classification models. Utilizing high-frequency Limit Order Book data from the MSCI US Index, we evaluate various alpha signals, to predict intraday stock returns. Our analysis introduces a regularized non-parallel ordinal probit model, comparing its performance against a standard VWAP-execution strategy. Results show that while the regularized model improves prediction accuracy and enhances trade execution during the closing session, its effectiveness varies across morning and afternoon sessions. These findings suggest that while intraday alpha signals hold potential for improving trade execution, further research into shorter trading horizons is necessary for optimal performance.

Contents

List of Abbreviations

ADV	Average Daily Traded Volume
BPPM	Basis Points Per Minute
BPS	Basis Points
ETF	Exchange-Traded Fund
GICS	Global Industry Classification Standard
LASSO	Least Absolute Shrinkage and Selection Operator
MMS	Market Micro-Structure
NYSE	New York Stock Exchange
PIR	Previous Intraday Return
VWAP	Volume-Weighted Average Price

Definitions of Variables

Listed in order of occurrence:

T_1, T_2	Timestamps to determine time intervals for predictor variables
t	Timestamp of the occurrence of a trade or change in the order book of a stock
T	Timestamp
Δ	Change in time
Price_t	Stock price of a trade taking place at time t
$V_t, V_t^{lit}, V_t^{dark}$	Size of an arbitrary/lit pool/dark pool trade taking place at time t
$\sigma(t)$	Volatility of the instantaneous log-returns process
$W(t)$	Standard Wiener process
σ_{IV}^2	Integrated variance
m	Sampling frequency for the volatility estimator
ShrOut_t	Shares outstanding at time t
$q_{t,b}, q_{t,a}$	Top-level bid/ask quote at time t
ρ_t	Order book imbalance at time t
$p_{t,b}, p_{t,a}$	Top-level bid/ask price at time t
$x_{i,t}$	Predictor variable of stock i at time t
$S_{t,j}$	Set of stocks in portfolio j at time t
$w_{i,t}$	Weights for stock i at time t to form a portfolio
$\mathbf{x}_{i,t}$	Vector of P predictor variables for stock i at time t
$\mathbf{r}_{t,\Delta}$	Vector of N returns over a forward looking interval $\text{Int}(t, t+\Delta)$
β	Vector of true exposures
$\mathcal{T}$	Number of trading days in the sample
N	Number of stocks in the cross-section
P	Number of predictor variables
X_t	$(P+1) \times N$ matrix of predictor variables including a column of ones
$\hat{\beta}_{FM}$	Vector of Fama-MacBeth estimates
ξ_k	Knot k in a regression spline
$b_k(x)$	Truncated power basis function
C	Number of classes
ψ_c	Threshold to distinguish between return class c and $c+1$
α_c	Unknown cut point in latent variable model between class c and $c+1$

Φ	CDF of the standard normal distribution
ϕ	PDF of the standard normal distribution
$\beta^*, \beta_1^*, \beta_2^*$	Exposure to predictor variables in ordinal probit models
λ	Tuning parameter for the L1-regularization
$\ell_{i,t}$	Log-likelihood function
$\upsilon_{t,m}, \upsilon_{t,a}, \upsilon_{t,c}$	Initial traded volume fraction for morning/afternoon/closing session on day t
$\upsilon_{t,m}^*, \upsilon_{t,a}^*, \upsilon_{t,c}^*$	Adjusted traded volume fraction for morning/afternoon/closing on day t
φ_f, φ_b	Front/Back-loading factor for order execution

1 Introduction

Long-term investors such as hedge funds, pension funds and insurance companies play a critical role in navigating uncertainties and optimizing portfolio performance. These investors execute orders either manually through traders or via execution algorithms. When these orders comprise considerable amounts of the daily traded volume they are not executed simultaneously, but rather executed gradually throughout the day to limit market impact (Said, 2022). The execution of these trades offer a wide range of opportunities for asset managers to benefit from market dynamics and generate additional alpha. Quantitative hedge funds have demonstrated success in this area, revealed by their substantial growth in recent decades. However, much of their research remains outside the scope of academic finance literature. These firms operate faster than traditional hedge funds, which focus on monthly horizons or longer, but slower than high-frequency traders focusing on horizons below seconds (Goldstein, Spatt, & Ye, 2021). Bridging this gap in the financial literature is a promising avenue.

Recent advancements in technology have made vast amounts of data from equity markets widely accessible. From limit order book (LOB) data at higher frequencies to alternative data sources such as news and social media analytics. Deciphering this vast amount of data to find crucial predictors for future price movements is highly challenging. Intraday price information typically exhibits a low signal-to-noise ratio, making it challenging to extract predictive information accurately. However, the development of machine learning algorithms has shown promise in leveraging feature information to predict future price behavior. One downside of these algorithms is their black box nature, which prevents the understanding of the true underlying dynamics between historical information and future price movements.

This paper builds upon the current research and addresses the following main question: *How can intraday alpha signals be used to improve the execution of intraday equity trades?* To explore this question, we investigate three subquestions. First, we evaluate the cross-section of stock returns using portfolio sorts and Fama-MacBeth regressions to identify intraday alpha signals. We assess multiple signals from existing research and combine them with new signals to evaluate their predictive power. In doing so, we answer the question: *Can cross-sectional analysis of stock returns yield reliable intraday alpha signals?* Next, we introduce a novel technique that, to our knowledge, has not been explored in existing literature. We categorize stock returns into three ordinal return classes and employ ordinal classification models to predict cross-sectional stock returns. This approach allows for the prediction of a group of observations instead of exact point predictions where

we try to minimize the error between predicted returns and true returns. To enhance interpretability, we keep our analysis to interpretable models instead of shifting to black-box modelling. We compare a standard ordinal probit model to one extended with an L1-penalty term and a non-parallel form, and evaluate both models' performance using quantitative and qualitative evaluation techniques. This analysis addresses our second subquestion: *Can classification methods effectively categorize intraday stock returns in the cross-section, and does the regularized non-parallel model outperform the standard ordinal probit model?* Finally, we assess the economic relevance of the models and perform a trade execution simulation. Comparing the models' performance against a standard execution strategy based on the volume weighted average price (VWAP) allows us to answer our last subquestion: *To what extent can ordinal probit models enhance the effectiveness of VWAP-execution strategies in intraday equity trading?* It's important to note that transaction costs are not factored into this simulation since improvements in intraday alpha are expected to enhance trade execution regardless.

Our analysis utilizes a high-frequency dataset of LOB-data sampled at a 5-minute frequency across constituents of the MSCI US Index from September 1, 2020, to September 1, 2023. While many studies focus on predicting mid-price returns, this research employs the VWAP, often considered a more accurate measure of the true market price. Furthermore, the dataset is segmented into four sessions: overnight, morning, afternoon, and closing. Our focus is on predicting returns for the last three sessions: morning, afternoon, and closing.

This paper identifies several signals in the cross-section of stock returns that are useful for predicting future stock returns on an intraday basis. Historical returns and order-flow imbalance (OFI) particularly demonstrate predictive power within this timeframe. Additionally, we find that enriching the dataset using regression splines leads to poor performance of the ordinal probit model in recognizing each return class. However, the regularized non-parallel ordinal probit model shows significant improvement in classifying return classes in-sample. Furthermore, the regularized non-parallel ordinal probit model demonstrates an improvement over a three-year timespan when compared to a standard VWAP-execution, particularly during the closing session. However, its performance is worse during the morning and afternoon sessions. Therefore, we conclude that while the non-parallel ordinal probit model is a viable option for intraday return prediction, its signals for intraday alpha should be applied to a higher frequency and smaller prediction horizon.

The remainder of this paper is structured as follows: Section 2 reviews related literature on short-term signals, the intraday cross-section of stock returns, and the predictability of intraday returns. Section 3 provides an overview of the data. The methodology, including feature creation and

evaluation, the model overview, and the simulation setup, is discussed in Section 4. Section 5 presents the empirical results of our methodology on our dataset. Finally, Section 6 concludes the research and presents a discussion of the results.

2 Literature Review

This section starts by defining the gap in finance literature between long-term and extreme short-term equity price modelling, it then describes the pre-existing literature on existing short-term alpha signals and ends with model selection for intraday stock return prediction.

2.1 Modeling stock returns

In the financial literature, two strands of literature exist for modeling equity prices. In the first place there is the conventional asset pricing literature using a factor-like structure to explain the cross-section of stock returns. This strand of literature, initiated by the introduction of the CAPM (Sharpe, 1964), seeks to understand differences in stock returns by adapting the thought of exposure to certain systematic factors. It turned out that there is a whole zoo of factors to explain differences among stock returns (J. H. Cochrane, 2011). Since these factors often consist of variables that were only published on a monthly basis, this field of research focuses on the monthly horizon or above to explain the cross-section. In the second place, there is a strand of literature that is focused on the extreme short-term behavior of equity returns and tries to explain market microstructures (MMS). Here the analysed time horizon can go down to seconds or microseconds.

In the years after Goldstein et al. (2021) defined the gap between these two strands of literature, research has slowly moved towards investigating the shorter time horizon. This shift was driven by factors such as greater availability and quality of data, as well as the transformation of alpha into beta (Jones & Mo, 2020). Additionally, the increased difficulty to identify alpha over longer periods (McLean & Pontiff, 2016) has played a role in driving the development in this field. One strand of literature tries to better explain risk premia of well-established factors. Lou et al. (2019) show that abnormal returns of momentum and short-term reversal strategies occur overnight, while abnormal returns on other strategies occur intraday. They also show that this pattern is not driven by news. Aleti (2022) goes even further and explains differences in the cross-section of Fama and French (2015) factor returns in a high-frequency setting. Another strand of literature finds new anomalies and factors within the dynamics of higher sampling frequencies. One example is that of Blitz et al. (2023), who obtain significant 'short-term alpha' by combining multiple short-term signals on a daily basis that are often discarded due to high turnover. We adopt a similar methodology, but

move even one step further and shift our focus towards intraday alpha.

2.2 Short-term alpha signals

Extracting short-term alpha signals from the cross-section has recently gained more attention. An illustrative example of research extracting high-frequency signals in the cross-section is provided by Huddleston et al. (2023). They employ a 5-minute sampling frequency for a large cross-section of stock returns to predict the market index. In the same fashion Aleti et al. (2023) use a 15-minute sampling frequency for a large set of well-established asset pricing factors to predict the market index. In this research we take a different approach and extract signals from the cross-section within the overnight, morning, afternoon and closing session instead of a constant sampling frequency.

Most of the research on short-term alpha signals makes use of historical returns to define momentum or reversal patterns. Heston et al. (2010) find that the relative performance of a stock during any half-hour of the trading day persists during the same half-hour on subsequent days. The effect lasts for at least 40 trading days and volume, order imbalance, volatility and bid-ask spreads show similar patterns but do not explain the return patterns. Murphy and Thirumalai (2017) further investigate this 'micro-momentum' effect and reveal that institutional net order submissions have a strong predictive power for returns in the same half-hour on future days. They argue that this is the result of institutions using a VWAP-strategy, generating predictable price pressure over the day.

Gao et al. (2018) are the first to find an intraday time-series momentum effect in the S&P 500 ETF where the first half-hour return as measured from the previous day's market close predicts the last half-hour return. Similarly, Zhang et al. (2019) identify a significant intraday momentum effect on the Chinese stock market where the first and/or second-to-last half-hour returns can significantly predict the last half-hour return. Also, Elaut et al. (2018) establish the presence of intraday time-series momentum in the RUB/USD currency pair as a positive relationship between the first half-hour return and the last-half hour return. Baltussen et al. (2021) find time-series momentum in the market index of equity futures, bonds, commodities and currencies sampled between 1974 and 2020. The return during the last 30 minutes before the market close is positively predicted by the return from the previous market close to the last 30 minutes.

However, Chu et al. (2019) find a reversal effect in the cross-section of the Chinese A-share market and attribute this contradiction to the difference between cross-sectional and time-series analysis. They use the first half-hour returns to negatively predict the rest of the day's returns in the cross-section. They also relate changes in large order imbalances to this reversal effect. To investigate

intraday momentum and reversal patterns we also investigate lagged returns as predictor variables in our dataset.

Kakushadze (2014) proposes a 4-factor model for overnight returns consisting of size, volatility, momentum and liquidity. These factors show to be relevant predictors for overnight returns and are suspected to also perform well for intraday returns. We therefore also add size, volatility and volume to the set of predictor variables.

As Chu et al. (2019) describes, LOB features can drive intraday prices. Chordia et al. (2002) were the first to define that returns are strongly affected by contemporaneous and lagged order book imbalance (OBI) and that returns reverse after a high-negative OBI. Cont et al. (2014) define a dynamic variant of order imbalance: OFI. Their research shows that there is a linear relation between OFI and price changes. In more recent research, Cont et al. (2023) show how OFI can be constructed over different sampling frequencies and how the predictive performance of OFI decreases over longer prediction horizons. In this research, we consider both OBI and OFI as predictor variables.

2.3 Model selection

This research aims to exploit detected signals to predict short-term stock price movements. Elaut et al. (2018), Jin et al. (2019), Wen et al. (2022), and Zhang et al. (2019) provide evidence that this intraday return predictability is present in currency and commodity exchange-traded funds (ETFs). However, there is less research available on classification methods in stock return prediction. Our prediction approach is most closely related to the work of Chen and Tsai (2020) and Cohen et al. (2020), who use Convolutional Neural Networks to classify candlestick images and identify trading patterns for buy or sell signals.

Within a large set of predictor variables, we aim to select the most important variables while leaving out irrelevant ones. A common approach in the statistical literature is to use the Least Absolute Shrinkage and Selection Operator (LASSO) (Tibshirani, 1996). Chinco et al. (2019) apply the LASSO to make rolling one-minute-ahead return forecasts and show that it achieves good out-of-sample performance by using the entire cross-section of lagged returns as candidate predictors. Additionally, Feng et al. (2020), Kozak et al. (2020), and Freyberger et al. (2020) demonstrate that penalized regression has proven effective within the finance literature. Therefore, we incorporate an L1-penalty term into the model extension of the ordinal probit model.

This research contributes to the existing strand of literature on three points. First, we evaluate

intraday return predictability throughout sessions of the trading day: the morning, afternoon, and closing sessions. Second, we investigate the predictive power of a new predictor variable related to the surprise of traded volume in dark pools. Third, we show that returns can be grouped into return classes, and an ordinal classification approach can be used to predict stock returns in the cross-section. We also demonstrate that in this setting, classification metrics are useful tools for gaining insight into the model's behavior.

3 Data

This chapter covers the data used in this research. We introduce the MSCI US Index and its components, explain the distinction between data from dark pools and lit pools, outline how we predict returns using different parts of a trading day, describe our data preprocessing, and present descriptive statistics.

3.1 MSCI US Index

In this research, we use a high-frequency dataset consisting of 5-minute price observations from constituents of the MSCI US Index, and spanning the sample period from September 1, 2020 to September 1, 2023. This cross-section consists of the largest exchange listed companies in the US, including Apple, Microsoft and NVIDIA, making them also the most liquid stocks in the world. The number of constituents in the MSCI US Index can change over time due to rebalancing. Figure 1 shows the resulting graph of the number of constituents over time, resulting in an average number of 619 constituents.

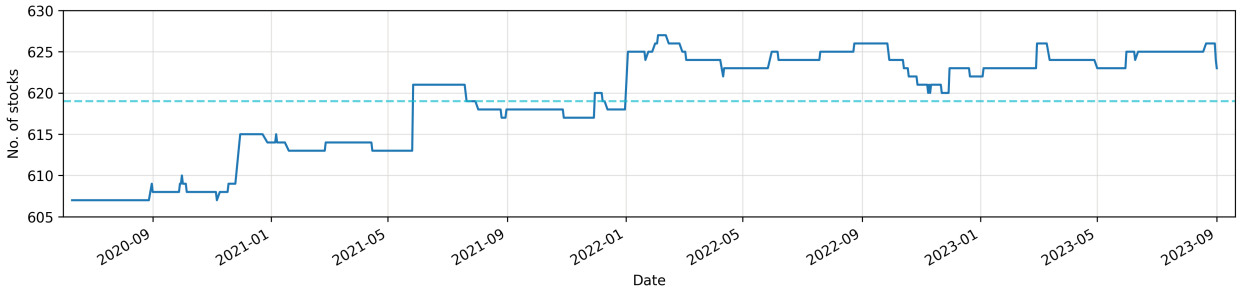

Figure 1: Count of MSCI US constituents over the sample period for each date (solid line) and the corresponding average number of MSCI US constituents (dashed line).

3.2 Dark pools vs. lit pools

All constituents of the MSCI US Index are listed on the New York Stock Exchange (NYSE) or the Nasdaq Stock Market. These public exchanges are commonly referred to as "lit pools" due to their transparent nature, serving as traditional trading venues where buyers and sellers come together to

trade securities. Within lit pools, all trading activity, including bid and ask prices, trade volumes and order sizes, is visible to market participants. In this way lit pools play a crucial role in price discovery, as they reflect the supply and demand dynamics of the market in real-time.

Next to lit pools there are also dark pools. Dark pools are private and used by large institutional investors. These pools are called "dark" because the trading activity that occurs within them is not visible to the general market participants until after the trades are executed. Unlike lit pools where order book information is public, dark pools only reveal matched trades, offering anonymity to participants. Dark pools are often used for executing block trades, which involve trading a significant number of shares at once. Since the trades are not immediately visible to the market, large orders can be executed with less impact on the stock price compared to trading on public exchanges.

The dataset used in this research consists of dark pool and lit pool data. Order book and price data is only available for public exchanges, but data on historically traded volume is available for both the public exchanges and dark pools. Section 4.1 further describes how this information is used to create predictor variables.

3.3 Sessions of the trading day

In our analysis we focus on 4 different sessions of a trading day. In this way we mimic a setting where traders execute orders throughout the day and distinguish between sessions that have their own set of characteristics. Table 1 displays the times corresponding to each of the sessions. The overnight session starts at 4:05 PM and ends before 9:30 AM on the next day. During this session the US stock markets are closed and stocks are not traded. The morning session starts at 9:30 AM indicating the opening time of the US stock markets and ends before 12:00 PM. The morning session distinguishes itself with a higher level of volatility and higher level of volume traded, as a result of overnight information that is incorporated into the stock price.

The afternoon session starts at 12:00 PM and ends before 3:50 PM. This session typically amounts to a lower level of trading activity.

The closing session starts at 3:50 PM and ends before 4:05 PM[1]. During this session the closing auction takes place. This auction is conducted by the Nasdaq and NYSE for the stocks listed at these exchanges and consists of two phases: the call phase and the price determination phase. During the call phase orders are entered, modified or cancelled. During the price determination

[1]The closing auction typically ends at 4:00 PM, but we use 4:05 PM as a safe measure due to occasional delays in the dataset.

phase no more orders can be entered and an order-matching algorithm seeks to maximize the executable volume (Kandel et al., 2012). The closing session often exhibits the highest level of trading volume throughout the day. The closing price is especially important to mutual funds and ETFs. This is because mutual funds and ETFs often create or redeem their shares based on the closing price of the assets held within the fund. Additionally, the closing auction is an attractive period to obtain liquidity and thereby limiting the amount of market impact compared to the continuous trading during the rest of the day (Jegadeesh & Wu, 2022).

Overnight	**Morning**	**Afternoon**	**Closing**
[16:05;09:30)	[09:30;12:00)	[12:00;15:50)	[15:50;16:05)

Table 1: Times corresponding to the different sessions of a trading day.

3.4 Data preprocessing

This research required extensive preprocessing steps, including validating the correctness of order book data, the alignment of stock data across multiple datasets, incorporation of stock splits and evaluation of outliers. One issue encountered was the lag in incorporated changes during rebalancing, leading to sudden spikes in the number of constituents. To overcome these spikes, constituents that leave the day after new constituents enter, are deleted when new constituents enter. Also, price observations with less than 500 constituents were deleted to keep the cross-section informative.

In Chapter 4 we describe two methods to analyze predictor variables for the cross-section of stock returns: portfolio sorts and Fama-MacBeth regressions. We use unadjusted data for these two methods, where features are truncated at the 99.5% and 0.05% levels. This ensures a straightforward interpretation of coefficients.

Additionally, we discuss two models for predicting cross-sectional stock returns: the ordinal probit regression and the regularized non-parallel ordinal probit regression. For these models we incorporate an extra data preprocessing step to normalize predictor variables. Given that we are examining the full cross-section of stock returns over a 3-year time period, stock characteristics may display significant scale differences and outliers. To address this, our normalization approach focuses on the relative size of characteristics within the cross-section, rather than solely on their absolute size. In doing so we we follow Kelly, Pruitt, and Su (2019) and Freyberger et al. (2020). The normalization is performed for each session on each date separately. Each stock characteristic is first ranked from low to high and then the ranks are mapped into the $[-1, 1]$ interval. The benefits

of this transformation are that it does not impose any assumptions due to its monotonicity, is less sensitive to outliers, and yields better out-of-sample predictions. Also, this transformation directly relates to portfolio sorting where we are typically not interested in the value of a characteristic on its own, but rather in the rank of a characteristic within the cross-section.

3.5 Descriptive statistics

The data utilized in this research comprises observations from the overnight, morning, afternoon, and closing sessions. Table 2 presents descriptive statistics for the returns per session in basis points (BPS). Analysis of the mean column reveals that the majority of returns are generated overnight, with a positive mean of 3.55 BPS and a negative skewness indicating that the distribution is skewed towards positive returns. Moreover, the relatively high kurtosis for overnight returns suggests the presence of more outliers.

The morning session, on the other hand, exhibits more negative returns on average and a moderate left skewness, indicating a longer tail on the left side of the distribution, potentially associated with occasional large losses during this period. In contrast, the afternoon session demonstrates more positive returns on average and a higher kurtosis, implying a higher frequency of outliers.

The closing session, although also showing negative average returns, exhibits less variation as indicated by the lower standard deviation. This reduced variability may be attributed to the session's smaller time interval. Additionally, both the skewness and kurtosis values for the closing session are closer to normal values, indicating a distribution that is more symmetric and has less outliers compared to the other sessions.

	N	Mean	Std. Dev.	25%	50%	75%	Skewness	Kurtosis
Overnight	496608	3.55	186.51	-44.28	6.91	53.81	-39.89	6229.36
Morning	499154	-1.76	169.85	-84.76	1.07	84.23	-0.27	15.49
Afternoon	499154	0.72	106.08	-51.00	2.50	54.54	-0.13	24.28
Closing	499154	-1.57	30.62	-18.95	-1.51	15.80	0.09	7.74

Table 2: Descriptive statistics for returns during the overnight, morning, afternoon and closing session. Returns are expressed in BPS. 25%, 50% and 75% correspond to respectively the 25th percentile, 50th percentile (median), and 75th percentile.

4 Methodology

This chapter outlines the quantitative methods used in our research. We define a response variable and introduce predictor variables to understand and forecast VWAP-returns in Section 4.1. Sections 4.2 and 4.3 discuss portfolio sorts and Fama-MacBeth regressions to identify relationships between

cross-sectional VWAP-returns and predictor variables. Section 4.4 covers using regression splines to enhance predictor variables. Section 4.5 explains our approach to transforming return prediction into a classification problem and the classification models we employ. Lastly, Section 4.7 explains how simulation assesses economic relevance, while Section 4.8 details out-of-sample predictions.

4.1 Response and predictor variables

The variables that follow are defined for each stock individually and notation is based on the work of Aït-Sahalia et al. (2022). For this notation we use a calendar interval between two timestamps $T_1, T_2 \in \mathbb{R}$ defined as an half open half closed interval:

$$\text{Int}(T_1, T_2) = \{t \in \mathbb{R} : T_1 < t \leq T_2\}. \tag{4.1}$$

4.1.1 VWAP-returns

In this research, we rely on the Volume-Weighted Average Price (VWAP) as a key metric and use it to calculate returns. VWAP's relevance lies in its ability to account for both prices and volume of actual trades that took place, providing a more accurate representation of the price that is traded on. The VWAP for a single stock at time T over a forward-looking interval $\text{Int}(T, T+\Delta)$ is calculated as

$$\text{VWAP}(T, \Delta) = \frac{\sum_{t \in \text{Int}(T, T+\Delta)} \text{Price}_t \times V_t}{\sum_{t \in \text{Int}(T, T+\Delta)} \text{Price}_t}. \tag{4.2}$$

This quantity measures the VWAP over a forward interval Δ where Price_t is the stock price of a trade that takes place during the specified interval, and V_t its corresponding volume. When Δ is fixed for an entire sample we can also interpret it as the sampling frequency. For example, VWAP(09-01-2022 09:00, 5min) amounts to the VWAP over the interval (09-01-2022 09:00, 09-01-2022 09:05] and Δ can be interpreted as the sampling frequency if the VWAP calculation is repeated for each subsequent 5 minutes. In this research we use a 5-minute sampling frequency and define

$$P_T := \text{VWAP}(T, 5\text{min}). \tag{4.3}$$

The VWAP-measure is utilized to calculate VWAP-returns, which serve as the response variable investigated in this research. For a specific calendar time at timestamp T, the return over a forward-looking interval $\text{Int}(T, T+\Delta)$ with $\Delta > 5$ minutes is calculated as

$$\text{Return}(T, \Delta) = \log\left(\frac{P_{T+\Delta-5\text{min}}}{P_T}\right) \times 10,000, \tag{4.4}$$

and returns calculated for a specific stock i are denoted as $\text{Return}(i, T, \Delta)$. For ease of notation we sometimes refer to $\text{Return}(i, T, \Delta)$ as $r_{i,t}$ leaving out the interval length for each session in case

we refer to the return in an arbitrary session. To leverage the statistical properties and address continuous compounding, we calculate log returns using the natural logarithm. The resulting log returns are then multiplied by 10,000 to express them in BPS, enhancing the interpretability of coefficients. Note that in Equation (4.4), we subtracted 5 minutes from the endpoint of the interval, taking into account that the VWAP is calculated over the last 5-minute bin. Furthermore, this return can only be calculated over intervals that are multiples of 5 minutes.

4.1.2 Predictor variables

In order to examine how the cross-sectional stock returns in different sessions can be predicted, we consider a large number of predictors that describe the intraday trading environment for a particular stock. Similar to the forward-looking interval for the response variable discussed in Section 4.1.1, we construct look-back intervals to create predictor variables.

Past returns

The first set of predictors are historical returns. Examining past returns provides valuable insights into the stock's performance and market behavior. For calendar time at timestamp T and look-back spans Δ_1 and Δ_2, for which it holds that $\Delta_1 \leq \Delta_2$ and $\Delta_2 - \Delta_1 > 5$ minutes, we calculate the previous intraday return (PIR) over $\text{Int}(T - \Delta_2, T - \Delta_1)$ as

$$\text{PIR}(T, \Delta_1, \Delta_2) = \log\left(\frac{P_{T-\Delta_1-5\text{min}}}{P_{T-\Delta_2}}\right) \times 10,000. \tag{4.5}$$

Similar to Equation (4.4), we subtract 5 minutes from the right-end point of the interval, and the PIR can only be calculated over intervals with lengths that are multiples of 5 minutes. Using returns of preceding parts of the day and comparing them to subsequent returns we can uncover short-term momentum or reversal patterns. We define an action pattern as follows:

$$\begin{cases} \text{momentum if } sign\ \text{PIR}(T, \Delta_1, \Delta_2) = sign\ \text{Return}(T, \Delta) \\ \text{reversal if } sign\ \text{PIR}(T, \Delta_1, \Delta_2) \neq sign\ \text{Return}(T, \Delta). \end{cases} \tag{4.6}$$

Realized volatility

We acknowledge the potential concern that the predictive strength of returns may be influenced by the volatility of the stock. Volatility can be described as a measure of the variability of stock prices over some period of time. When a stock experiences fluctuations in its price over time, understanding and quantifying this volatility can become informative for predicting future market movements. For instance, heightened volatility may indicate increased uncertainty or potential market excitement, influencing the likelihood of more significant price swings. In financial literature,

it is commonly assumed that the instantaneous logarithmic price of a stock follows a simple diffusion process (Bachelier, 1900):

$$d\log P(t) = \sigma(t)dW(t). \tag{4.7}$$

Here, $P(t)$ represents the unobservable instantaneous continuous time price, $\sigma(t)$ represents the volatility of the instantaneous log-returns process and $W(t)$ is the standard Wiener process. The integrated variance over a time interval $\text{Int}(T-\Delta_2, T-\Delta_1)$ is then defined as:

$$\sigma^2_{IV}(T, \Delta_1, \Delta_2) = \int_{T-\Delta_2}^{T-\Delta_1} \sigma^2(t)dt. \tag{4.8}$$

The integrated variance over the interval $\text{Int}(T-\Delta_2, T-\Delta_1)$ is known to be the actual, but unobservable variance. However, the so-called realized variance can be used as an alternative measure for the integrated variance (Andersen & Bollerslev, 1998). Under the assumptions that (i) the log-prices follow the diffusion process in Equation (4.7) and (ii) there are no microstructure frictions, it can be shown that the realized variance converges in probability to the integrated variance when the number of sub-intervals τ tends to infinity:

$$\underset{\tau\to\infty}{p\lim}\ RV(T, \Delta_1, \Delta_2, m) = \sigma^2_{IV}(T, \Delta_1, \Delta_2), \tag{4.9}$$

with the realized variance on the time interval $\text{Int}(T-\Delta_2, T-\Delta_1)$ partitioned in τ equidistant points:

$$RV(T, \Delta_1, \Delta_2, m) = \sum_{j=1}^{\tau} (\log P_{T-\Delta_2+m(j-1)} - \log P_{T-\Delta_2+mj})^2, \tag{4.10}$$

where $m = \frac{(T-\Delta_1)-(T-\Delta_2)}{\tau-1} = \frac{\Delta_2-\Delta_1}{\tau-1}$ is defined as the sampling frequency. This means that accuracy improves as the sampling frequency increases. However, in practice it turns out that assumption (ii) of no microstructure frictions is a rather crucial one. When moving to extremely high frequency or even tick-by-tick data, market friction is a source of additional noise in estimating the volatility. The sampling frequency is therefore based on a trade-off between accuracy and biases due to this market microstructure noise.

We make this trade-off using a volatility signature plot (Andersen et al., 2000), representing a graphical representation of the average realized volatility against the sampling frequency. We calculate the daily realized volatility per sampling frequency according to (4.10) and take the time series average to get the average realized volatility per stock. Figure 2 shows the signature plot for the average daily realized volatility of the cross-section including a 95% confidence interval. To make the trade-off between accuracy and bias we choose the highest frequency where the average realized volatility appears to stabilize. This comes down to a sampling frequency of 30 minutes. Hence, the

square root of the realized variance based on the 30 minute sampling frequency is used as a measure for the volatility of each stock at timestamp T over a look-back interval $\text{Int}(T-\Delta_2, T-\Delta_1)$:

$$\text{Volatility}(T, \Delta_1, \Delta_2) = \sqrt{\sum_{j=1}^{\tau} (\log P_{T-\Delta_2+(j-1)\times 30\text{min}} - \log P_{T-\Delta_2+j\times 30\text{min}})^2}. \tag{4.11}$$

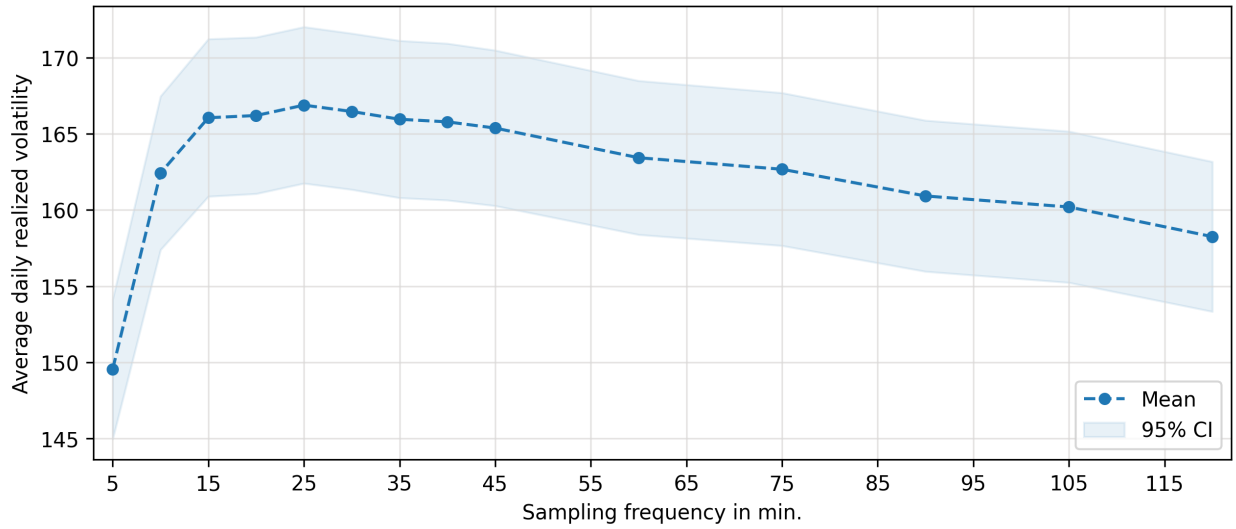

Figure 2: Volatility signature plot for average daily realized volatility per sampling frequency. For each stock the time series average of the daily realized volatility is calculated per sampling frequency. The dotted line visualizes the cross-sectional mean of the average daily realized volatility per sampling frequency, together with a 95% confidence interval.

Size

To measure the size of a stock we use the natural logarithm of its market capitalization - the current stock price multiplied by the total value of its outstanding shares:

$$\text{Size}(T, \Delta_1, \Delta_2) = \log[P_{T-\Delta_2} \times \text{ShrOut}_{\max(\mathbf{I})}], \tag{4.12}$$

with $\mathbf{I} = \text{Int}(T-\Delta_1, T-\Delta_2)$ and $\max(\mathbf{I})$ the last available observation over this time interval. In this way we can account for the size effect (Fama & French, 1992) and distinguish between small- and large-cap companies. Generally, large companies are more stable and less volatile, while smaller companies present higher growth potential but with increased risk.

Volume

To measure the amount of trading interest for a stock, we use traded volume as a feature. The traded volume is simply defined as the number of shares traded within a specified timeframe:

$$\text{Volume}(T, \Delta_1, \Delta_2) = \sum_{t \in \text{Int}(T-\Delta_2, T-\Delta_1)} V_t. \tag{4.13}$$

Volume surprise

The trading volume throughout the trading day usually exhibits a hockey stick-like pattern. Figure 3 illustrates that the day starts with higher trading volume, attributed to the incorporation of overnight information. Subsequently, the volume gradually declines to its lowest point around 13:00, before increasing again. The highest trading volume is observed in the last 10 minutes of the day, primarily during the closing auction, a period dominated by institutional investors.

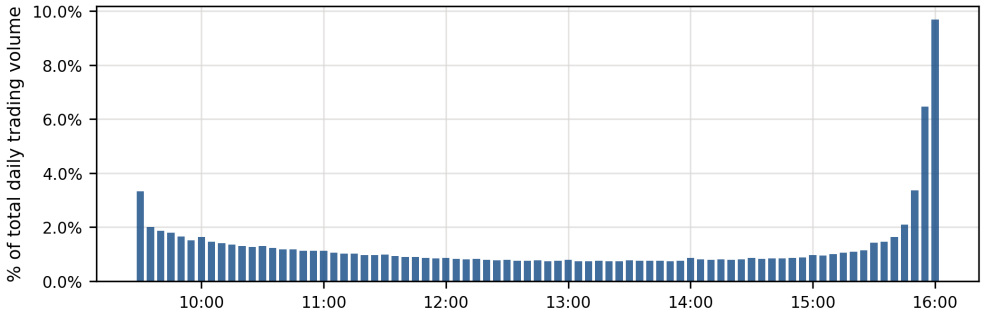

Figure 3: Percentage of total daily traded volume per 5-minute time bin. For each stock the volume is scaled by stock size and averaged out per 5-minute time bin over the full cross-section and entire time span from October 6, 2020, to January 9, 2023.

Index-fund managers typically engage in trading near the close to align with the returns of their benchmark. Similarly, mutual funds often wait to execute trades, ensuring they have a clearer understanding of how much cash they need to raise or invest (Jegadeesh & Wu, 2022). A large deviation from the usual volume pattern indicates that their is new information being incorporated into the stock prices. However, since institutional traders often want to keep their information advantage over other traders, they start by trading in dark pools (Buti et al., 2022). Within dark pools the order book is not visible and trades can be placed anonymously. Therefore, we use a measure that compares a stock's trading volume during a session to the expected trading volume. The expected trading volume is calculated using the average trading volume of the same interval over the previous 25 trading days. To account for the additional information of the volume traded on the dark pool, we separate between dark pool volume and lit pool volume:

$$\text{LitVolumeSurprise}(T, \Delta_1, \Delta_2) = \frac{\sum_{t \in \text{Int}(T-\Delta_2, T-\Delta_1)} V_t^{lit}}{\frac{1}{25}\sum_{d=1}^{25} \sum_{t \in \text{Int}(T-\Delta_2-d, T-\Delta_1-d)} V_t^{lit}} \tag{4.14}$$

$$\text{DarkVolumeSurprise}(T, \Delta_1, \Delta_2) = \frac{\sum_{t \in \text{Int}(T-\Delta_2, T-\Delta_1)} V_t^{dark}}{\frac{1}{25}\sum_{d=1}^{25} \sum_{t \in \text{Int}(T-\Delta_2-d, T-\Delta_1-d)} V_t^{dark}} \tag{4.15}$$

Order book imbalance

Following the literature (Cartea et al., 2015) we determine a measure for the amount of imbalance between the bid and ask side. The average OBI defined over the look-back interval is defined as:

$$\begin{aligned} \mathrm{OBI}(T, \Delta_1, \Delta_2) &= \mathrm{Average}\left[\rho_t : t \in \mathrm{Int}^{\mathrm{back}}(T, \Delta_1, \Delta_2)\right], \\ \rho_t &= \frac{q_{t,a} - q_{t,b}}{q_{t,a} + q_{t,b}}, \quad -1 \leq \rho_t \leq 1. \end{aligned} \tag{4.16}$$

Where q_t^b and q_t^a are the top-level bid and ask order book quotes, respectively. The key idea of this metric is to evaluate whether the market for the stock is leaning more towards selling or buying. Cartea et al. (2018) define a value of ρ_t between -1 and $-\frac{1}{3}$ as a sell-heavy signal, between $-\frac{1}{3}$ and $\frac{1}{3}$ as a neutral signal, and between $\frac{1}{3}$ and 1 as a buy-heavy signal.

Order flow imbalance

Since OBI looks at the total volume in the top-level order book, it could be that some of the volume results from older orders that contain less relevant information. Cont et al. (2014) found that over short intraday time intervals price changes are mainly driven by OFI. We follow the notation used by Cont et al. (2023)[2] and define the bid order flows ($\mathrm{OF}_{t,b}$) and ask order flows ($\mathrm{OF}_{t,a}$) at time t as:

$$\mathrm{OF}_{t,b} = \begin{cases} q_{t,b} & \text{if } p_{t,b} > p_{t-1,b}, \\ q_{t,b} - q_{t-1,b} & \text{if } p_{t,b} = p_{t-1,b}, \\ -q_{t-1,b} & \text{if } p_{t,b} < p_{t-1,b}, \end{cases} \qquad \mathrm{OF}_{t,a} = \begin{cases} -q_{t-1,a} & \text{if } p_{t,a} > p_{t-1,a}, \\ q_{t,a} - q_{t-1,a} & \text{if } p_{t,a} = p_{t-1,a}, \\ q_{t,a} & \text{if } p_{t,a} < p_{t-1,a}. \end{cases} \tag{4.17}$$

The order flows are then used to calculate the accumulative order flow imbalances during a given time interval:

$$\mathrm{OFI}(T, \Delta_1, \Delta_2) = \frac{1}{\mathrm{AvgDepth}(T, \Delta_1, \Delta_2)} \sum_{t \in \mathrm{Int}^{back}(T-\Delta_2, T-\Delta_1)} \mathrm{OF}_{t,b} - \mathrm{OF}_{t,a}, \tag{4.18}$$

$$\mathrm{AvgDepth}(T, \Delta_1, \Delta_2) = \frac{1}{2} \times \mathrm{Average}\left[q_{t,b} + q_{t,a} : t \in \mathrm{Int}^{\mathrm{back}}(T, \Delta_1, \Delta_2)\right].$$

[2]Cont et al. (2023) made an error in their notation, they incorrectly associated cases where $p_{t,b} < p_{t-1,b}$ and $p_{t,a} > p_{t-1,a}$ with the negative best bid $(-q_{t,b})$ and best ask quantity $(-q_{t,a})$ at time t. However, these cases actually refer to the negative best bid $(-q_{t-1,b})$ and best ask quantity $(-q_{t-1,a})$ at time $t-1$, not at time t as stated. This indicates situations where the entire best bid/ask quantity is traded away, resulting in a negative order flow equivalent to the size of the previous best bid/ask quantity.

The net order flow imbalance is scaled by the average order book depth over the same interval to account for the intraday pattern in limit order depth (Ahn et al., 2001; Harris & Panchapagesan, 2005).

Seasonal variables

Stock returns can display seasonal fluctuations at various intervals. To account for this, we introduce three seasonal categories into the dataset: year, quarter, and day of the week. These categories are converted into dummy variables using one-hot encoding to assess their impact. One-hot encoding transforms categorical variables into numerical format, crucial for models requiring numerical inputs. It ensures that for n categories, $n-1$ dummy variables are created. Each dummy variable represents a category, with all dummies being zero indicating the presence of the nth category. Including all n dummy variables as predictor variables in a model leads to perfect linear dependence, which is now avoided by using one class as a reference. This strategy effectively mitigates multicollinearity issues.

Sector classification

The dataset used in this research comprises over 600 stocks from the MSCI US Index, categorized into various industry sectors. Stocks within the same sector typically respond similarly to macroeconomic and market conditions, while also encountering comparable industry-specific challenges and opportunities (Huang & Zhang, 2022). To adjust for variations in stock returns across sectors, sector dummies are included. The Global Industry Classification Standard (GICS) is utilized for sector-level grouping. Similar to seasonal dummy variables, sector dummy variables are constructed as predictor variables using one-hot encoding. Table 3 presents each GICS sector alongside its corresponding GICS reference number and the count of unique stocks within the MSCI US Index during the sample period[3]. Notably, the dataset is predominantly composed of stocks from Information Technology, Financials, and Industrials sectors, each exceeding 100 stocks. Conversely, the Energy, Utilities, and Materials sectors are less represented, with 35 or fewer stocks each.

4.2 Portfolio sorts

To explore the cross-sectional relationships between historical returns and session returns, we employ a well-established method from the cross-sectional literature known as univariate portfolio sorts (Jegadeesh & Titman, 1993). Univariate portfolio sorts offer two significant advantages: they are non-parametric and allow for the filtering of returns associated with a specific sorting

[3]The total number of stocks exceeds the maximum number of constituents in the MSCI US Index because constituents can enter or leave the index over time.

	GICS number	# of stocks		GICS number	# of stocks
Energy	10	31	Financials	40	108
Materials	15	35	Information Technology	45	137
Industrials	20	109	Communication Services	50	48
Consumer Discretionary	25	74	Utilities	55	34
Consumer Staples	30	41	Real Estate	60	46
Health Care	35	91	**Total**	-	**754**

Table 3: Number of stocks and GICS reference number per sector.

characteristic through the analysis of high-minus-low portfolios. We define the sorting variable as $\text{PIR}(t, \Delta_1, \Delta_2)$ for stock i at time t across different lookback intervals specified by Δ_1 and Δ_2. Additionally, $\text{Return}(i, t, \Delta)$ represents the outcome variable, denoting the VWAP-return of stock i over a forward looking interval defined by Δ. This outcome variable is evaluated during morning, afternoon, and closing sessions. At the start of each session t, we iterate through PIRs of preceding sessions:

1. Rank stocks from losers to winners based on $\text{PIR}(t, \Delta_1, \Delta_2)$ of stock i.
2. Make 5 quintile portfolios with $S_{t,j}$ the set of stocks in the j'th portfolio at time t.
3. Consider portfolios $j = 1$ and $j = 5$, and the high-minus-low portfolio (5-1).
4. Evaluate the portfolio returns for portfolio j at time t using stock specific weights $w_{i,t}$:

$$\overline{r}_{j,t} = \frac{\sum_{i \in S_{j,t}} w_{i,t} \times \text{Return}(i, t, \Delta)}{\sum_{i \in S_{j,t}} w_{i,t}}. \tag{4.19}$$

After repeating this procedure at each timestamp t we evaluate the j'th portfolio and the high-minus-low portfolio as follows:

$$\overline{r}_j = \frac{\sum_{t=1}^{\mathcal{T}} \overline{r}_{j,t}}{\mathcal{T}}, \qquad \overline{r}_{5\text{-}1} = \frac{\sum_{t=1}^{\mathcal{T}} \overline{r}_{5,t} - \overline{r}_{1,t}}{\mathcal{T}}, \tag{4.20}$$

where $\mathcal{T}$ is the number of trading days in the sample that have a morning, afternoon or closing session. These evaluation metrics are then tested on their mean being significantly different from zero. We evaluate equally-weighted portfolios ($w_{i,t} = 1$) such that the portfolio return is the simple average of the stock returns, and we evaluate the value-weighted portfolios using $w_{i,t}$ as the market capitalization of stock i. Here we expect to see differences between equally-weighted and value-weighted portfolio returns when several small cap firms might produce extreme high returns. Since the MSCI US Index only consists of approximately 620 constituents, we choose to sort stocks in 5 quintile portfolios such that the portfolios represent a significant amount of stocks.

4.3 Fama-MacBeth regressions

When adding more explanatory variables to sort on in portfolio sorts, the analysis can become cumbersome. We therefore employ the Fama-MacBeth regression (Fama & MacBeth, 1973) to investigate possible factors driving intraday returns while correcting for other effects. We do this in a more standard setup[4], as described by (J. Cochrane, 2009) and (Bali et al., 2016):

$$\text{Return}(i,t,\Delta) = \mathbf{x}_{i,t}^{\top}\beta + \varepsilon_{i,t} \quad i = 1,...,N; t = 1,...,\mathcal{T}-1. \tag{4.21}$$

The return over a forward looking interval $\text{Return}(i,t,\Delta)$ on the left hand side is explained by a vector of true exposures β to a vector of predictors $\mathbf{x}_{i,t}$, and the residual $\varepsilon_{i,t}$. Note that the data in this regression setting has a cross-sectional element as well as a time-series element.

Our goal is to explain future stock returns using signals from the cross-section. The Fama-MacBeth regression is an alternative approach to finding these signals and can be used in a setting with multiple explanatory variables to explain cross-sectional returns. This approach can be described in two steps. First, we run a cross-sectional regression of the explanatory variables on all the stock returns at each moment in time to get beta estimates:

$$\hat{\beta}_t = (X_t^{\top}X_t)^{-1}X_t^{\top}\mathbf{r}_{t,\Delta}. \tag{4.22}$$

Here, $\mathbf{r}_{t,\Delta}$ is a $1 \times N$ vector of N stock returns over the forward looking interval $\text{Int}(t, t+\Delta)$, X_t is a $(P+1) \times N$ matrix consisting of a column of ones and P predictors for N stocks at time t, and $\hat{\beta}_t$ is a $1 \times (P+1)$ vector of estimated beta's for each predictor variable at time point t. The result is a time series of intercepts, slope coefficients and error terms. Second, the Fama-MacBeth estimates are calculated as the average of the cross-sectional regression estimates:

$$\hat{\beta}_{FM} = E_{\mathcal{T}}(\hat{\beta}_t). \tag{4.23}$$

To investigate the explanatory power of each of the explanatory variables we can perform a t-test on the Fama-MacBeth estimate to check whether it is statistically different from zero. Returns sampled at a higher frequency are likely to be autocorrelated due to for example market trends and momentum effects, and the variance of the error terms is highly likely to change over time due to for example volatility clustering. To account for these issues we perform the t-test by calculating Newey-West standard errors (Newey & West, 1987). In this way we adjust for autocorrelation and heteroskedasticity.

[4]In the original expected return-beta asset pricing model, Fama and MacBeth (1973) utilized factor loadings β_i for $x_{i,t}$, and the factor risk premium λ as β together with a rolling window to estimate risk premia.

4.4 Regression splines

The Fama-MacBeth regression setting makes use of a linear relationship between the predictor variables to explain returns in the cross-section. Often, it is convenient to represent the features by a linear model since it offers a clear interpretation, and it is able to fit the data without overfitting. However, within the complexity of financial markets it is extremely unlikely that the true function $f(X)$ is linear in X. Instead of moving directly to black-box modelling, we employ a non-linear model that can exhibit local linearity to keep a descent level of interpretability. This model is referred to as a degree-d regression spline, which is a piecewise degree-d polynomial with continuity in derivatives up to degree $d-1$ at each knot. Hence, for the linear regression spline we can place knots on several points within the domain of the predictor variable where the function is continuous. Between these knots the model fits a degree-1 polynomial, representing a linear relationship between the predictor and response variable. The linear regression spline with knots at ξ_k for $k = 1, ..., K$ can be represented as

$$\text{Return}(i,t,\Delta) = \beta_0 + \beta_1 b_1(x_{i,t}) + \beta_2 b_2(x_{i,t}) + \cdots + \beta_{K+1} b_{K+1}(x_{i,t}) + \varepsilon_i, \tag{4.24}$$

for some basis functions $b_1, b_2, ..., b_{K+1}$ and predictor variable $x_{i,t}$. To represent a linear spline we start with a simple basis function $b_1(x_{i,t}) = x_{i,t}$ and add a truncated power basis function per knot, defined as

$$b_k(x_{i,t}) = (x_{i,t} - \xi_k)_+ = \begin{cases} (x_{i,t} - \xi_k) & \text{if } x_{i,t} > \xi_k \\ 0 & \text{otherwise.} \end{cases} \tag{4.25}$$

For illustrative purposes the left panel of Figure 4 shows a linear spline with a knot at $x_{i,t} = 4$. The right panel shows the corresponding basis functions. Adding the points for both lines, would lead to the linear regression spline in the left panel. A great advantage of the model in Equation (4.24) is that it can be fit using least squares. By making use of regression splines we can enrich the set of predictor variables in a situation where we assume the relationship between the response variable and a predictor variable to be non-linear. The truncated basis functions can be added to the functional form as new predictor variables on which coefficients are estimated.

4.5 Return classification

4.5.1 Order execution

In this research we take a setting where a trader receives trade orders that have to be executed before the end of the day. Since the size of these orders comprise considerable amounts of the daily traded volume, they are not executed simultaneously. These orders are executed gradually throughout the

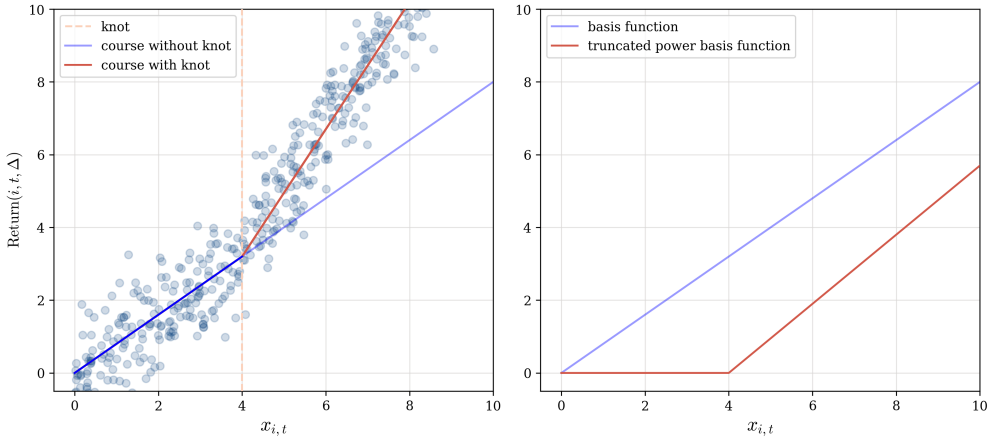

Figure 4: The left panel shows an illustration of a linear spline with a single knot and what the line looks like without the knot, which represents a simple linear model. The right panel shows the corresponding basis functions.

day to limit the impact they have on the market (Said, 2022). When a considerable amount of stocks is traded simultaneously the market tends to move in the opposite direction, leading to a loss on the performed trades. Since the trading volume exhibits seasonal patterns throughout the day, traders take this volume into account when executing their orders. For example, around the opening and closing of the trading day volume tends to be higher (see Figure 3) meaning a larger amount of trades can be executed without experiencing too much market impact.

However, there is still a trade-off to be made since the stock price during these periods can be less advantageous. If traders can receive information on the expected returns for an upcoming time interval, they can decide on the level of aggressiveness to place trades. For example, when a trader has a large buy order that needs to be executed throughout the day and he gets a signal that prices will move up in the afternoon he can start executing trades more aggressively, meaning he is prepared to accept a less favorable price to ensure faster execution.

To allow for this distinction in trade execution we classify returns in three possible categories: low, moderate and high. These classes are based on the threshold values ψ_1 and ψ_2 that split the range of return values into three classes:

$$y_{i,t} = \begin{cases} \textit{negative} & \text{if Return}(i,t,\Delta) \leq \psi_1, \\ \textit{moderate} & \text{if } \psi_1 \leq \text{Return}(i,t,\Delta) \leq \psi_2, \\ \textit{positive} & \text{if Return}(i,t,\Delta) \leq \psi_2. \end{cases} \tag{4.26}$$

One could argue that the problem of order execution can also be resolved by predicting numerical returns. However, by reformulating the prediction of a numerical variable to a classification task has several advantages. In the first place it allows to focus on predicting a group of observations

instead of exact point predictions where we try to minimize the error between each of the predicted return observations and true return observations. In the second place, it allows to evaluate how well models are able to distinguish between classes. This allows for a direct connection between a specific class and the way of executing the order.

The next step is to use classification methods to accurately classify the response variables $\mathbf{y}_t = (y_{1,t}, ..., y_{N,t})$ with $y_{i,t} \in \{low, moderate, high\}$ using the set of proposed predictor variables. Since the response variables have an ordered nature by construction, as reflected by Equation (4.26), the ordering of the variables contains additional information and our model should reflect this. Following the growing literature on machine learning techniques for predicting stock returns (e.g. Gu et al. (2020), Huddleston et al. (2023)) we use several machine learning techniques to classify the response variables and incorporate their natural ordering.

4.5.2 Ordinal probit regression

We start with the most common regression model for this type of data, the ordinal probit regression model (McCullagh, 1980), which is also known as the ordered probit model. This model accounts for both the classification nature and the ordered multi-class response variables. Following Wooldridge (2010) ordinal data $y_{i,t}$ can be modeled by assuming an underlying latent variable model:

$$y^*_{i,t} = \mathbf{x}^\top_{i,t}\beta^* + \varepsilon_{i,t}, \quad \varepsilon_{i,t} \sim N(0,1), \tag{4.27}$$

where $y^*_{i,t}$ is continuous but latent data, β^* is the exposure to the predictor vector $\mathbf{x}_{i,t}$, and this vector does not contain a constant. When estimating this model each of the unknown $C-1$ cut points α_c between two of the C classes is defined as an intercept. The random variable $y_{i,t}$ can then be viewed as a class variable for the latent variable falling in one of the classes:

$$y_{i,t} = \begin{cases} 0 & \text{if } y^*_{i,t} \leq \alpha_1, \\ 1 & \text{if } \alpha_1 \leq y^*_{i,t} \leq \alpha_1, \\ \quad \vdots & \\ C & \text{if } y^*_{i,t} \geq \alpha_{C-1}. \end{cases} \tag{4.28}$$

However, in our setting the response variable is created upfront based on the continuous variable $\text{Return}(i,t,\Delta)$ and we can therefore treat the latent variable as an observed variable (i.e. $y^*_{i,t} = \text{Return}(i,t,\Delta)$). The intercepts correspond to the threshold values ψ_1 and ψ_2 in Equation (4.26),

and cumulative class probabilities can be derived:

$$\begin{aligned} \mathrm{P}(y_{i,t} \leq c \mid \mathbf{x}_{i,t}) &= \mathrm{P}(y^*_{i,t} \leq \psi_c \mid \mathbf{x}_{i,t}) \\ &= \mathrm{P}(\varepsilon_{i,t} \leq \psi_c - \mathbf{x}^\top_{i,t}\beta^* \mid \mathbf{x}_{i,t}) \\ &= \Phi(\psi_c - \mathbf{x}^\top_{i,t}\beta^*). \end{aligned} \tag{4.29}$$

Now the ordinal probit regression model can be written as:

$$\begin{aligned} &\mathrm{P}(y_{i,t} = \textit{negative} \mid \mathbf{x}_{i,t}) = \mathrm{P}(y^*_{i,t} \leq \psi_1 \mid \mathbf{x}_{i,t}) = \Phi(\psi_1 - \mathbf{x}^\top_{i,t}\beta^*), \\ &\mathrm{P}(y_{i,t} = \textit{moderate} \mid \mathbf{x}_{i,t}) = \mathrm{P}(\psi_1 \leq y^*_{i,t} \leq \psi_2 \mid \mathbf{x}_{i,t}) = \Phi(\psi_2 - \mathbf{x}^\top_{i,t}\beta^*) - \Phi(\psi_1 - \mathbf{x}^\top_{i,t}\beta^*), \\ &\mathrm{P}(y_{i,t} = \textit{positive} \mid \mathbf{x}_{i,t}) = \mathrm{P}(y^*_{i,t} > \psi_2 \mid \mathbf{x}_{i,t}) = 1 - \Phi(\psi_2 - \mathbf{x}^\top_{i,t}\beta^*). \end{aligned} \tag{4.30}$$

In these equations, $\mathbf{x}_{i,t}$ is a $P \times N\mathcal{T}$ matrix of predictors, where time series observations and cross-sectional observations are stacked over the rows, and each column corresponds to a predictor. β^* is a $1 \times P$ column vector of unknown coefficients, Φ represents the cumulative distribution function of the standard normal distribution, and the 'unknown' parameters ψ_1, ψ_2 and β^* are estimated by maximum likelihood. Additionally, it is important to note that the three probabilities sum to one. While the thresholds ψ_1 and ψ_2 are in fact observed, as they were defined upfront to group returns into classes, these values could be passed to the maximum likelihood estimation as given. However, in practice, it is often not feasible to pass threshold values directly in regression packages so we leave ψ_1 and ψ_2 to be estimated. This approach allows the thresholds to be estimated in such a way that it optimally classifies the return classes. For each observation $y_{i,t}$ the log-likelihood function is then given by:

$$\begin{aligned} \ell_{i,t}(\psi_1, \psi_2, \beta^*) = &\ \mathbb{I}_{[y_{i,t}=\textit{negative}]} \log[\Phi(\psi_1 - \mathbf{x}^\top_{i,t}\beta^*)] \\ + &\ \mathbb{I}_{[y_{i,t}=\textit{moderate}]} \log[\Phi(\psi_2 - \mathbf{x}^\top_{i,t}\beta^*) - \Phi(\psi_1 - \mathbf{x}^\top_{i,t}\beta^*)] \\ + &\ \mathbb{I}_{[y_{i,t}=\textit{positive}]} \log[1 - \Phi(\psi_2 - \mathbf{x}^\top_{i,t}\beta^*)]. \end{aligned} \tag{4.31}$$

Note that we now use the observed predictors and classes instead of the random variables. The objective for the ordered probit model is then to estimate the unknown parameters by maximizing the log-likelihood function for the entire sample of $N \times \mathcal{T}$ observations. This is equal to minimizing the objective function:

$$\underset{\psi_1, \psi_2, \beta^*}{\operatorname{argmin}} \left\{ -\frac{1}{N \times \mathcal{T}} \sum_{i=1}^{N} \sum_{t=1}^{\mathcal{T}} \ell_{i,t}(\psi_1, \psi_2, \beta^*) \right\}, \tag{4.32}$$

to obtain the estimates $\hat{\psi}_1$, $\hat{\psi}_2$ and $\hat{\beta}^*$. Here we follow the common convention to scale the negative log-likelihood by the number of observations (Friedman et al., 2010) such that the negative

log-likelihood can be compared across datasets of different sizes. After estimating the model, probabilities for each of the three classes are calculated using Equation 4.30. The predicted class is then determined by selecting the class with the highest probability.

4.5.3 Regularized non-parallel ordinal probit regression

Given the large number of predictor variables relative to the sample size and the low signal-to-noise ratio in stock return data, we aim to design a model capable of selecting relevant features. Additionally, our objective is for the model to extract complex relationships between the data and the response variable. Lastly, we want to maintain the interpretability of the ordinal probit model. To achieve this, we add an L1-penalty to the objective function and allow the cumulative class probabilities to vary across cumulative probabilities.

Non-parallel model form

The model in Section 4.5.2 has a functional form where the predictor variables have the same set of coefficients across all three categories. Yee (2010) refers to this restriction as the parallelism assumption. This assumption implies the presence of parallel regression lines for the same given value across all cumulative probabilities $\mathrm{P}(y_{i,t}^* \leq c \mid \mathbf{x}_{i,t})$. The intuition behind this is that if the distribution of the latent variable is shifted for each of the different thresholds, the way it shifts through each of these thresholds is always the same. This means that the probability of going up and down for a certain class follows exactly the same pattern for each of the thresholds. The slope at a given value $0.5 = \Phi(\psi_c - \mathbf{x}_{i,t}^\top \beta^*)$ with respect to the linear predictors $\mathbf{x}_{i,t}^\top$ is

$$\frac{\partial \mathrm{P}(y_{i,t}^* \leq c \mid \mathbf{x}_{i,t})}{\partial \mathbf{x}_{i,t}^\top \beta^*} = \frac{\partial \Phi(\psi_c - \mathbf{x}_{i,t}^\top \beta^*)}{\partial \mathbf{x}_{i,t}^\top \beta^*} = -\phi(\psi_c - \mathbf{x}_{i,t}^\top \beta^*) = -\phi(0), \tag{4.33}$$

where ϕ is the standard normal PDF and we use the fact that $\psi_c - \mathbf{x}_{i,t}^\top \beta^* = \Phi^{-1}(0.5)$. Equation (4.33) implies that the slope has a constant value irrespective of the corresponding threshold.

The left pane of Figure 5 shows a visual example of an ordinal probit model with three classes, resulting in two thresholds for evaluating the cumulative distribution function (CDF). The first CDF represents the probability of $y_{i,t} \leq 1$, indicating the probability of ending up in the negative returns class, which decreases as the regressor $x_{i,t}$ increases. Similarly, the second CDF, corresponding to $y_{i,t} \leq 2$, indicating the probability of ending up in the negative or moderate return class, also decreases as $x_{i,t}$ decreases. Notably, the slopes where the curves intersect 50% probability, indicating an equal probability of ending up above and below a category, remain consistent across both thresholds.

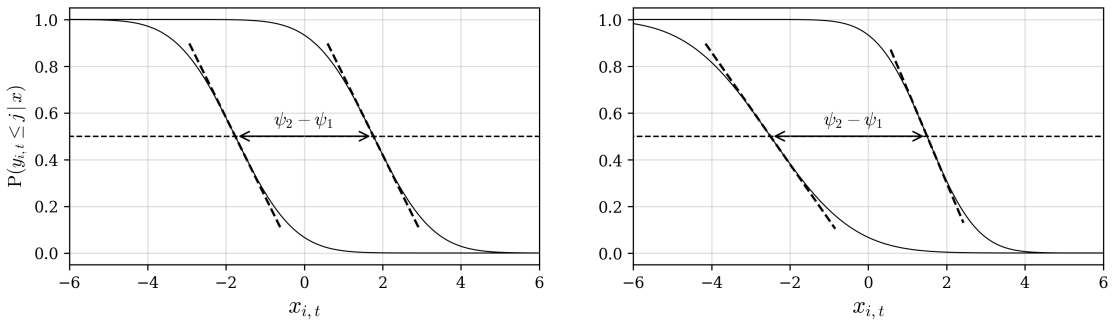

Figure 5: Visualization of the parallel ordinal model form. The left pane shows a parallel regression form, and the right pane shows a non-parallel regression form.

The parallel model form can be relaxed by allowing the beta-coefficients to vary over the cumulative class probabilities. This relaxation leads to the non-parallel model form. The right pane of Figure 5 shows this model form where the parallelism assumption does not hold. In this case the probability of being in the negative return class goes down slower than the probability of being in the negative or moderate return class. Here we have different beta-coefficients for the different thresholds, allowing for a more flexible model form. Specifically, for higher values of $x_{i,t}$, the probability of ending up at most in the moderate return class decreases faster than the probability of ending up at most in the negative return class.

We now have two different column vectors β_1^* and β_2^* that need to be estimated. In this way, the cumulative class probabilities are not forced to shift together, resulting in a more flexible model for ordinal data. The corresponding log-likelihood of the non-parallel model form can be written as:

$$\begin{aligned} \ell_{i,t}(\psi_1, \psi_2, \beta_1^*, \beta_2^*) = &\ \mathbb{I}_{[y_{i,t}=negative]} \log[\Phi(\psi_1 - \mathbf{x}_{i,t}^\top \beta_1^*)] \\ + &\ \mathbb{I}_{[y_{i,t}=moderate]} \log[\Phi(\psi_2 - \mathbf{x}_{i,t}^\top \beta_2^*) - \Phi(\psi_1 - \mathbf{x}_{i,t}^\top \beta_1^*)] \\ + &\ \mathbb{I}_{[y_{i,t}=positive]} \log[1 - \Phi(\psi_2 - \mathbf{x}_{i,t}^\top \beta_2^*)]. \end{aligned} \tag{4.34}$$

L1-regularization

Next to the non-parallel model form we also incorporate the L_1-penalty term into the objective function. In this way sparsity is enforced in the coefficient matrix by effectively shrinking the coefficients of unimportant features towards zero. The $L1$-penalty makes the model particularly useful for evaluating the relevance of predictor variables when dealing with a large set of potential predictors.

To estimate the regularized ordinal probit regression we need to minimize the scaled negative

log-likelihood of the sample with respect to the parameters ψ_1, ψ_2, β_1^* and β_2^*:

$$\underset{\psi_1,\psi_2,\beta_1^*,\beta_2^*}{\operatorname{argmin}} \left\{ -\frac{1}{N \times \mathcal{T}} \sum_{i=1}^{N} \sum_{t=1}^{\mathcal{T}} \ell_{i,t}(\psi_1, \psi_2, \beta_1^*, \beta_2^*) + \lambda \sum_{j=1}^{P} \left(\mid \beta_{1,j}^* \mid + \mid \beta_{2,j}^* \mid \right) \right\}, \tag{4.35}$$

with penalty parameter $\lambda \geq 0$, and $\beta_{1,j}^*$ and $\beta_{2,j}^*$ the j'th coefficient of the coefficient vectors β_1^* and β_2^*. When λ is zero the model falls back to the non-parallel ordinal regression model and a higher value of λ indicates a larger penalty on the beta-coefficients. Since there is no analytical solution for the coefficients in this model, the model is optimized using a coordinate descent algorithm and the optimal value for the hyperparameter λ is obtained by cross-validation. In order to implement the regularized ordinal regression model we use the ordinalNet R-package from Wurm et al. (2021). The hyperparameter λ is determined by generating a sequence of twenty lambda values where the one with the largest log-likelihood for the sample data is selected. The sequence is determined by λ_{max} that is equal to the smallest value that sets every coefficient to zero and $\lambda_{min} = \lambda_{max} \cdot 0.01$. The sequence runs from λ_{min} to λ_{max} uniformly on the logarithmic scale, meaning that the spacing between different values of lambda increase exponentially[5].

4.6 In-sample evaluation

We start by performing an in-sample evaluation of the three previously mentioned models. We evaluate the models quantitatively using the confusion matrix, accuracy, precision, recall and $F1$-score, and evaluate the models qualitatively using calibration plots and probability histograms for the return classes.

To calculate the quantitative evaluation metrics we make use of four basic classification metrics: True Positives, True Negatives, False Positives, and False Negatives. Since the classification setting described in Section 4.5 makes use of three classes (negative, moderate, positive), the interpretation of these metrics becomes more nuanced. True Positives (TP): In a multi-class setting, TP refers to the instances correctly predicted for the specific class of interest. For example, the number of correctly predicted 'positive' returns. True Negatives (TN): Number of instances correctly predicted as not belonging to the class of interest. False Positives (FP): Number of instances incorrectly predicted as belonging to the class of interest. False Negatives (FN): Number of instances incorrectly predicted as not belonging to the class of interest in a multi-class scenario. Using these four classification metrics we can now define accuracy, precision, recall, $F1$-score and weighted $F1$-score.

[5] A logarithmic scale search ensures that the entire hyperparameter space is explored in a more balanced way. For example, when tuning a hyperparameter on a linear scale search between 0.0001 and 1.0 might unevenly allocate training resources, spending 90% on the range 0.1 to 1.0 and only 10% on 0.0001 to 0.1. Hence, the logarithmic scale gives a better sample of the entire range.

Accuracy is a measure of the overall correctness of the model. It is the ratio of correctly predicted instances to the total number of instances:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}. \tag{4.36}$$

Precision is the accuracy of positive predictions. It is the ratio of correctly predicted positive observations to the total predicted positives:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{4.37}$$

Recall is a measure of the ability of the model to capture all the relevant instances. It is the ratio of correctly predicted positive observations to all observations in the actual class:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{4.38}$$

The F1-score is the harmonic mean of precision and recall. It provides a balanced measure that considers both false positives and false negatives:

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4.39}$$

The weighted F1-score calculates the average F1-score, considering the number of samples in each class. It provides a more representative measure when classes are imbalanced.

$$\text{Weighted F1-score} = \frac{\sum_{i=1}^{N}(\text{F1-score}_i \times \text{no. of samples in class}_i)}{\text{Total no. of samples}}$$

In addition to quantitative evaluation metrics, we also assess the model's performance qualitatively using calibration plots and probability histograms. Calibration plots illustrate how well a model's predicted probabilities align with actual outcomes. On these plots, the x-axis represents the average predicted probabilities grouped into bins, while the y-axis indicates the actual proportions in those bins. A diagonal line on the plot signifies accurate predictions. Probability histograms, on the other hand, visually compare predicted probabilities for observations within the true class against those outside the true class. For instance, when observations outside the class have lower predicted probabilities and those inside the class have higher predicted probabilities, the two histograms are further further located from each other. This indicates that the model effectively distinguishes between observations falling inside and outside the class. By examining both calibration plots and probability histograms for each class, we gain a comprehensive understanding of the model's reliability and accuracy in predicting probabilities for the three return classes.

4.7 Order simulation

To evaluate the economic significance of the predictor variables, we employ an order simulation. This simulation mimics a trading environment where buy or sell orders for different stocks arrive before each session and must be executed within one day of trading. For example, an order received before the morning session needs to be executed across the morning, afternoon, and closing sessions, while an order received before the afternoon session should span the afternoon, closing, and the following morning session. Within this simulated trading setting, each session involves executing a fraction of the total order volume. To minimize market impact, the executed fraction during a session is relative to the volume traded in the market during that specific session. This session's initial volume execution (υ) is expressed as a fraction of the average daily traded volume (ADV), which we calculate using a 25-day rolling average. Next, we apply the regularized non-parallel ordinal probit regression model to predict returns in the first session. Based on this prediction (whether returns are expected to be negative, moderate, or positive), we adjust the initially calculated volume fraction for that session (υ^*). This adjustment involves back or front loading or no adjustment at all, resulting in a new volume fraction in that session that is higher, lower or equal to the initial volume fraction.

Figure 6 visually illustrates this back and front loading process after the model's prediction. Once the first part of the trade is executed in the initial session, we gain new information during the first session that informs the prediction for returns in the next session. Using this updated information, we adjust the volume fraction for the subsequent session accordingly (υ^{**}). The remaining portion of the volume fraction is then allocated for the final session.

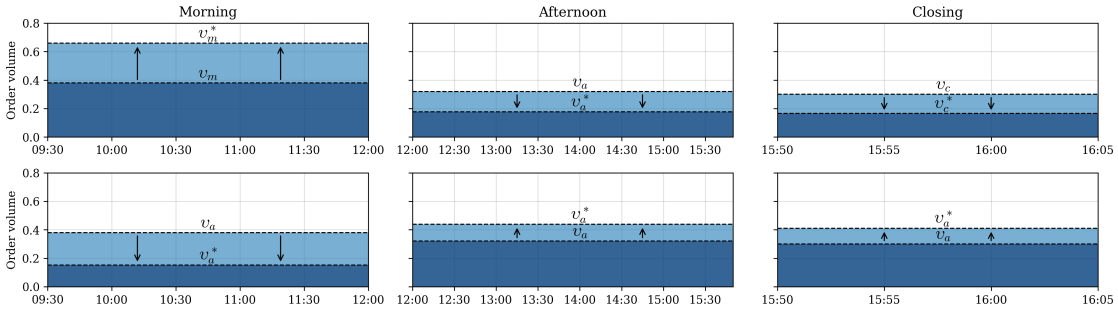

Figure 6: The graph compares two strategies for an order before the morning session: front loading (upper row) and back loading (lower row). It shows how the initial volume fractions $\upsilon_m, \upsilon_a, \upsilon_c$ are adjusted to $\upsilon_m^*, \upsilon_a^*, \upsilon_c^*$ after predicting the morning session. The x-axis represents trading sessions, and the y-axis indicates volume fractions. The difference in executed volume is shown in light blue.

Algorithm 1 outlines the pseudocode for setting up the simulation, specifically for buy orders

arriving before the morning session. During this simulation we must ensure that the executed volume per session υ^{**} remains within the range of zero to one. This constraint ensures that we execute the order efficiently without taking opposing positions or executing more than necessary. We achieve this by front loading using a fraction (φ_f) of the initial execution volume of subsequent session(s):

$$\upsilon^*_{t,m} = \upsilon_{t,m} + \varphi_f \times (\upsilon_{t,a} + \upsilon_{t,c}), \tag{4.40}$$

and back loading using a fraction (φ_b) of the initial execution volume of the current session:

$$\upsilon^*_{t,m} = \upsilon_{t,m} - \varphi_b \times \upsilon_{t,m}. \tag{4.41}$$

To evaluate the model's performance, we compare a standard VWAP-execution, based on a 25-day rolling average for each session, against VWAP-execution using predictions from the regularized ordinal probit model. This comparison is based on the cumulative BPS difference between the two execution strategies. For buy orders aiming for the lowest execution price:

$$\frac{VWAP_t - VWAP^{**}_t}{VWAP_t} \times 10,000, \tag{4.42}$$

and for sell orders targeting the highest execution price:

$$\frac{VWAP^{**}_t - VWAP_t}{VWAP_t} \times 10,000. \tag{4.43}$$

This simulation setup is performed for both buy and sell orders and orders arriving before the morning, afternoon, or closing session.

4.8 Out-of-sample evaluation

Lastly, we perform an out-of-sample evaluation. To perform an out-of-sample prediction the data must consist of at least one training set to train the model and its parameters on, and one test set to evaluate the model on. This resembles the real-life situation where the model is trained on a provided dataset and evaluated on unseen data. The standard approach in machine learning literature is to perform K-fold cross-validation by K times randomly picking a new train and test set. However, due to the time series nature of our model it is not possible to randomly choose samples to assign to a test or training set. This is due to the temporal dependency of time series observations and applying K-fold cross validation would lead to data leakage. To overcome this, we need to ensure that the test set always represents a subsequent moment in time compared to the training set. After the first train/test-split the test observations are included in the train set and the subsequent data points are predicted. This form of cross-validation is called time-series cross-validation and is shown in Figure 7.

Algorithm 1 Order simulation algorithm for buy orders arrived in the morning

for $t = 1$ **to** T **do**

  for each stock s in the set of available stocks S_t **do**

    Calculate 25-day rolling average volume for the morning, afternoon and closing:
    $\overline{V}^{25}_{t,m} = \frac{1}{25}\sum_{d=t-25}^{t-1} V_{d,m}$, $\overline{V}^{25}_{t,a} = \frac{1}{25}\sum_{d=t-25}^{t-1} V_{d,a}$, $\overline{V}^{25}_{t,c} = \frac{1}{25}\sum_{d=t-25}^{t-1} V_{d,c}$

    Calculate the fraction of volume to be executed for the morning, afternoon and closing:
    $\upsilon_{t,m} = \frac{\overline{V}^{25}_{t,m}}{\overline{V}^{25}_{t,m}+\overline{V}^{25}_{t,a}+\overline{V}^{25}_{t,c}}$, $\upsilon_{t,a} = \frac{\overline{V}^{25}_{t,a}}{\overline{V}^{25}_{t,m}+\overline{V}^{25}_{t,a}+\overline{V}^{25}_{t,c}}$, $\upsilon_{t,c} = \frac{\overline{V}^{25}_{t,c}}{\overline{V}^{25}_{t,m}+\overline{V}^{25}_{t,a}+\overline{V}^{25}_{t,c}}$

    Classify the return for the morning session: $\hat{r}_{t,m}$

    if $\hat{r}_{t,m}$ = positive **then**

      Back load the volume: $\varphi_f = 0$, $\varphi_b = 0.25$

    else if $\hat{r}_{t,m}$ = moderate **then**

      Do not alter the volume: $\varphi_f = 0$, $\varphi_b = 0$

    else if $\hat{r}_{t,m}$ = negative **then**

      Front load the volume: $\varphi_f = 0.25$, $\varphi_b = 0$

    end if

    Recalculate the volume per session:

      $\upsilon^*_{t,m} = \upsilon_{t,m} + \varphi_f \times (\upsilon_{t,a} + \upsilon_{t,c}) - \varphi_b \times \upsilon_{t,m}$

      $\upsilon^*_{t,a} = \upsilon_{t,a} - \varphi_f \times \upsilon_{t,a} + \varphi_b \times \frac{\upsilon_{t,a}}{\upsilon_{t,a}+\upsilon_{t,c}} \times \upsilon_{t,m}$

      $\upsilon^*_{t,c} = \upsilon_{t,c} - \varphi_f \times \upsilon_{t,c} + \varphi_b \times \frac{\upsilon_{t,c}}{\upsilon_{t,a}+\upsilon_{t,c}} \times \upsilon_{t,m}$

    Classify the return for the afternoon session: $\hat{r}_{t,a}$

    if $\hat{r}_{t,a}$ = positive **then**

      Back load the volume: $\varphi_f = 0$, $\varphi_b = 0.25$

    else if $\hat{r}_{t,a}$ = moderate **then**

      Do not alter the volume: $\varphi_f = 0$, $\varphi_b = 0$

    else if $\hat{r}_{t,a}$ = negative **then**

      Back load the volume: $\varphi_f = 0.25$, $\varphi_b = 0$

    end if

    Recalculate the volume per session:

      $\upsilon^{**}_{t,m} = \upsilon^*_{t,m}$

      $\upsilon^{**}_{t,a} = \upsilon^*_{t,a} + \varphi_f \times \upsilon^*_{t,c} - \varphi_b \times \upsilon^*_{t,a}$

      $\upsilon^{**}_{t,c} = \upsilon^*_{t,c} - \varphi_f \times \upsilon^*_{t,c} + \varphi_b \times \upsilon^*_{t,a}$

    Calculate VWAP for baseline execution: $VWAP_t = P_{t,m} \times \upsilon_{t,m} + P_{t,a} \times \upsilon_{t,a} + P_{t,c} \times \upsilon_{t,c}$

    Calculate VWAP for model-based execution: $VWAP^{**}_t = P_{t,m} \times \upsilon^{**}_{t,m} + P_{t,a} \times \upsilon^{**}_{t,a} + P_{t,c} \times \upsilon^{**}_{t,c}$

    Calculate improvement upon baseline execution in bps: $\frac{VWAP_t - VWAP^{**}_t}{VWAP_t} \times 10,000$

  end for

end for

Evaluate performance over time for all stocks: $\sum_{t=1}^{T}\sum_{s\in S_t} \frac{VWAP_t - VWAP^{**}_t}{VWAP_t} \times 10,000$

To evaluate the classification models we perform 5-fold time series cross-validation. The first fold starts with a 50/10 train/test split, after which the test set is added to the train set resulting in a 60/10 split for the second fold. This process is repeated 5 times such that the full sample data is used. Within each training set 5-fold cross-validation is used to estimate the hyperparameter λ.

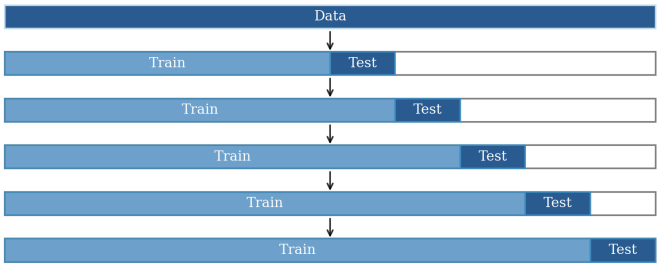

Figure 7: Visual representation of a 5-fold time series cross-validation.

5 Empirical results

In this chapter, we present and discuss the results obtained from the methods explained in the previous chapter. First, we discuss in Section 5.1 an analysis of the cross-section of stock returns to extract predictor variables that have explanatory power for intraday returns. The analysis starts with univariate portfolio sorts, followed by univariate and multivariate Fama-MacBeth regressions. Second, in Section 5.2 we extend the set of predictor variables as input for ordinal regression models and compare the models' predictive accuracy. Third, in Section 5.3 we evaluate the economic relevance of these predictor variables. We end this chapter with an out-of-sample evaluation of the best performing model in Section 5.4.

5.1 In-sample predictor analysis

Portfolio sorts

To find potential intraday momentum and reversal signals in the cross-section of stock returns we begin by performing portfolio sorts. Table 4 shows time-series average returns of 5 quintile portfolios that are sorted and evaluated on each day. Stocks are sorted on the PIR of the overnight, morning and afternoon session in respectively panel A, B and C, and grouped into quintile portfolios from low to high PIR, and the market-neutral high-minus-low portfolio.

In panel A we see a clear reversal pattern between the overnight PIR and the morning returns for all decile portfolios. We also see that in the equally-weighted portfolios the pattern is mainly driven by the long leg and in the value-weighted portfolios in the short leg. The overnight PIR portfolios still show a small reversal effect in the afternoon returns, and a momentum effect in

	PIR	$\overline{r}_{morning}$				$\overline{r}_{afternoon}$				$\overline{r}_{closing}$			
	EW	EW	t(EW)	VW	t(VW)	EW	t(EW)	VW	t(VW)	EW	t(EW)	VW	t(VW)
Panel A: $PIR_{overnight}$													
Low	-0.100	0.024	(1.16)	0.030	(1.61)	-0.002	(-0.19)	0.003	(0.28)	-0.036	(-0.65)	-0.035	(-0.57)
2	-0.024	0.010	(0.69)	0.010	(0.70)	0.000	(0.05)	0.002	(0.27)	-0.047	(-0.91)	-0.018	(-0.32)
3	0.004	-0.000	(-0.01)	0.003	(0.21)	-0.002	(-0.24)	-0.001	(-0.09)	-0.059	(-1.12)	-0.031	(-0.54)
4	0.031	-0.008	(-0.57)	-0.008	(-0.58)	-0.001	(-0.17)	-0.001	(-0.14)	-0.062	(-1.18)	-0.037	(-0.64)
High	0.102	-0.041	(-2.01)	-0.011	(-0.58)	-0.005	(-0.55)	-0.004	(-0.47)	-0.030	(-0.56)	-0.022	(-0.36)
High-Low	0.202	-0.064	(-3.17)	-0.041	(-2.06)	-0.003	(-0.36)	-0.007	(-0.80)	0.006	(0.24)	0.014	(0.44)
Panel B: $PIR_{morning}$													
Low	-1.156					-0.003	(-0.28)	-0.001	(-0.14)	-0.022	(-0.39)	0.001	(0.01)
2	-0.362					0.000	(0.04)	0.005	(0.62)	-0.040	(-0.75)	-0.017	(-0.29)
3	0.004					-0.001	(-0.14)	0.003	(0.43)	-0.048	(-0.92)	-0.033	(-0.58)
4	0.364					-0.002	(-0.27)	-0.002	(-0.22)	-0.068	(-1.33)	-0.046	(-0.79)
High	1.136					-0.004	(-0.46)	-0.002	(-0.23)	-0.056	(-1.05)	-0.033	(-0.55)
High-Low	2.291					-0.002	(-0.23)	-0.001	(-0.11)	-0.034	(-1.45)	-0.033	(-1.20)
Panel C: $PIR_{afternoon}$													
Low	-0.468									0.025	(0.44)	0.004	(0.07)
2	-0.152									-0.043	(-0.80)	-0.023	(-0.38)
3	-0.002									-0.065	(-1.22)	-0.043	(-0.74)
4	0.148									-0.073	(-1.40)	-0.037	(-0.64)
High	0.464									-0.079	(-1.52)	-0.037	(-0.69)
High-Low	0.932									-0.104	(-4.15)	-0.041	(-1.35)

Table 4: This table presents the time-series average portfolio returns for quintile portfolios obtained by sorting daily by equal-weighted (EW) PIR. The returns are evaluated for equal-weighted and value-weighted (VW) portfolios including the high minus low portfolio. Panel A, B, and C correspond to the overnight (16:00-09:30), morning (09:30-12:00), and afternoon (12:00-15:50) sessions to calculate the PIR. The first column shows PIRs for the quintile portfolios, followed by columns displaying equally and value-weighted portfolio returns for the morning, afternoon and closing (15:50-16:00) session. PIRs and portfolio returns are reported as log returns in BPS per minute, with t-statistics in parentheses based on Newey-West standard errors with 6 lags.

the closing session. This effect is mainly driven by the negative momentum of the short leg and becomes stronger for the value-weighted high-minus-low portfolios. It is clear that all overnight PIR portfolios show negative returns in the closing session. However, none of the portfolio returns are significantly different from zero.

In panel B we see that the portfolios sorted by the morning session have a spread (2.291) of more than 11 times the overnight PIR spread (0.202). We can conclude that the returns generated during the morning session are more dispersed. This could be related to overnight news that is incorporated by traders after the market opening resulting in more differences between returns during the morning session. We then see that for these portfolios less return is generated during the afternoon session. The SIR during the closing session show more of a reversal pattern but are still not significant.

In panel C we see that the spread in portfolio returns sorted on PIR in the afternoon session are more than 4 times the overnight PIR spread. Also, we see a clear reversal pattern with the closing

session SIR for the equally-weighted long-minus-short portfolio. This becomes non-significant when evaluating the value-weighted portfolios. It is worth noting that all of the PIR-sorted portfolios show negative returns in the closing session, except for the low portfolio showing a clearly positive return.

From the univariate portfolio sorts we can conclude that we mostly see reversal patterns between sessions throughout the day, but only the reversal between overnight returns and morning returns and between afternoon and closing returns (only for equally-weighted portfolios) are significant. Due to the fact that we use the MSCI US Index which already encompasses the largest stocks in the US, there is not much of a difference between equally weighted and size weighted portfolios.

Fama-MacBeth regressions

In order to evaluate the effect of several predictor variables simultaneously, we employ the Fama-MacBeth regression approach to explain returns for the morning, afternoon and closing session by using PIRs. In doing so we add three control variables: log volatility, log volume and log size. Log volatility is based on a 15-day Parkinson volatility estimator, log volume on a 15-day rolling average and log size based on the stock price and market capitalization at the previous day's closing. However, Bouchaud et al. (2018) states that the daily traded volume for a typical stock usually amounts to 0.1%-1% of its total market capitalization. Highly correlated predictors often lead to multicollinearity issues; making it difficult to isolate individual effects between correlated predictors and unstable coefficient estimates. We therefore evaluate log size and log volume in separate regressions.

The first three columns in table 5 show regressions on the morning returns. The univariate regression in column 1 shows that the overnight PIR is significant and negatively related to the morning returns, indicating a reversal pattern between the overnight and morning returns. The regression in column 2 controls for log volatility and log volume, and in column 3 for log volatility and log size. With these last two models we are able to explain around 13% of the total variance in the morning returns. We see that a 1% increase in returns during the overnight session is associated with a 0.21% decrease in returns in the morning session when correcting for volatility and volume, and associated with a 0.22% decrease in returns in the morning session when correcting for volatility and size.

The four subsequent columns in table 5 show regressions on the afternoon returns. None of the regressions show significant predictors except for log size. Therefore previous intraday returns cannot be used to explain afternoon returns.

Variable	$r_{t,morning}$			$r_{t,afternoon}$				$r_{t,closing}$				
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**	**(9)**	**(10)**	**(11)**	**(12)**
Intercept	0.0017	0.1339	-0.0313	-0.0020	-0.0035	0.0010	-0.0487	-0.0585	-0.0702	-0.6141***	-0.8431***	-0.6466***
	(0.12)	(1.45)	(-0.29)	(-0.26)	(-0.46)	(0.03)	(-1.33)	(-1.12)	(-1.37)	(-6.45)	(-5.68)	(-6.57)
$PIR_t^{afternoon}$								-0.1152***	-0.1395***	-0.1798***	-0.1790***	-0.1733***
								(-5.12)	(-6.56)	(-8.95)	(-8.89)	(-8.62)
$PIR_t^{morning}$				0.0001	-0.0013	-0.0039	-0.0040		-0.0192*	-0.0294***	-0.0292***	-0.0268**
				(0.05)	(-0.52)	(-1.93)	(-1.95)		(-2.18)	(-3.39)	(-3.36)	(-3.06)
$PIR_t^{overnight}$	-0.2581***	-0.2137***	-0.2238***		-0.0065	-0.0077	-0.0091		0.0095	-0.0007	0.0003	
	(-3.44)	(-3.50)	(-3.63)		(-0.25)	(-0.39)	(-0.46)		(0.12)	(-0.01)	(0.01)	
$\ln Volatility_t$		-0.0385	-0.0322			-0.0009	0.0012			0.1289***	0.1498***	0.1392***
		(-1.38)	(-1.13)			(-0.09)	(0.13)			(4.75)	(5.33)	(4.97)
$\ln Volume_t$		-0.0007				0.0001				0.0240***		0.0245***
		(-0.22)				(0.04)				(3.62)		(3.69)
$\ln Size_t$			0.0082**				0.0025*				0.0166*	
			(2.91)				(2.24)				(2.41)	
R^2	0.0400	0.1386	0.1373	0.0341	0.0611	0.1276	0.1259	0.0140	0.0380	0.0639	0.0647	0.0560
R^2_{adj}	0.0368	0.1329	0.1316	0.0309	0.0564	0.1203	0.1186	0.0107	0.0316	0.0545	0.0553	0.0482
T	808	808	808	808	808	808	808	808	808	808	808	808

Table 5: Regression coefficients from Fama-MacBeth regressions, obtained after regressing the variables from the first column on the log-returns in BPPM of the morning (09:30-12:00), afternoon (12:00-15:50) and closing session (15:50-16:00) for all stocks in the MSCI US Index at each day t. The coefficients are computed by averaging over the entire time span from October 6, 2020, to January 9, 2023. The last column for each session is obtained by consecutively deleting insignificant predictors based on the lowest t-statistic until only significant predictors are left. Here we accept a maximum probability of 5% to incorrectly reject a true null hypothesis. The intercept is always kept in the regression models. T-statistics are presented in parentheses, calculated using Newey-West standard errors with 6 lags. ***, **, and * denote 0.1%, 1%, and 5% significance level, respectively.

The five last columns show regressions on the closing returns. The univariate regression in column 8 shows that there is a significant reversal pattern present between the afternoon and closing session. Column 9 shows that returns during the morning session also have explanatory power for the closing session but returns during the overnight session not. Column 10 and 11 show that in addition log volatility, log volume, and log size also have significant coefficients. The resulting model is shown in column 14 consisting of the returns during the morning and afternoon session, log volatility and log volume. Returns are significantly reverting between the morning and closing and afternoon and closing. These reversions are smaller than between the overnight and morning session but still amount to a decrease of 0.17% in returns during the closing session when the afternoon returns increase by 1%, and a decrease of 0.03% returns during the closing session when the morning returns increase by 1%. Also, more volatile stocks and stocks with higher trading volume tend to have higher returns during the closing session.

5.2 In-sample model evaluation

Having studied the intraday dynamics of PIRs on VWAP-returns, we continue to study the in-sample performance of the classification models proposed in Section 4.5 using a larger set of predictor variables. We start by showing that the ordinal probit model fails to classify the moderate

return class and continue by showing that extending the model with an L1-penalty and non-parallel form helps in solving this issue and results in a good model fit. We then evaluate the effect of the L1-penalty on the set of predictor variables to see which variables are selected.

Ordinal probit regression

The ordinal probit regression model as described in Equation 4.30 serves as our baseline model, and the model is estimated using maximum likelihood (eq. 4.31).

Panel A in table 6 shows the evaluation metrics for each of the three sessions. The first notable thing is that the ordinal probit model fails to classify the moderate return class for each of the sessions, leaving the precision, recall and F1-score on zero for this class. The accuracy for each of the sessions is above 33%[6] implying that the ordinal probit model still has some explanatory power. However, since the model does not classify any of the observations in the moderate return class, the accuracy can give a skewed image of the true model performance. The weighted F1-score provides a more complete view, taking both precision and recall into account and weight each of the F1-scores by the number of observations per class. We now see that the weighted F1-score is highest for the morning session, followed by the closing and afternoon session. With an F1-score of around 29% we can conclude that the model does a poor job in identifying return classes for the different sessions.

	Morning			Afternoon			Closing		
	Negative	Moderate	Positive	Negative	Moderate	Positive	Negative	Moderate	Positive
Panel A: Ordinal Probit									
Precision	37.51	00.00	37.46	35.43	00.00	36.03	37.16	00.00	37.55
Recall	53.12	00.00	58.70	57.00	00.00	49.74	44.85	00.00	66.45
F1-score	43.97	00.00	45.74	43.69	00.00	41.79	40.65	00.00	47.99
Accuracy			37.48			35.71			37.39
Weighted F1			30.06			28.62			29.72
Panel B: Regularized Non-Parallel Ordinal Probit									
Precision	39.81	41.07	39.78	37.59	40.17	38.84	40.74	39.85	39.92
Recall	39.91	36.83	43.67	36.09	50.31	30.83	29.49	41.10	49.39
F1-score	39.86	38.83	41.63	36.82	44.68	34.37	34.21	40.47	44.15
Accuracy			40.17			39.00			40.09
Weighted F1			40.12			38.58			39.66

Table 6: Evaluation metrics in percentages for in-sample return classification. Precision, recall, and F1-score are assessed for each negative, moderate, and positive class. Accuracy and weighted F1-score are evaluated across all three classes. Panel A displays metrics for the Ordinal Probit Model. Panel B displays metrics for the extended model with L1-penalty and varying β-coefficients for cumulative class probabilities. The models are assessed for each trading day session.

[6] Assuming that the classes are balanced, an accuracy lower than $\frac{1}{3}$ implies that the model performs worse than randomly predicting one of the three classes, meaning that the model has no explanatory power.

To get more insight into the underlying model dynamics we also evaluate the model qualitatively. Figure 8 shows the calibration plots and probability histograms for the morning session. From the upper left and upper right calibration plot we see that the model underclassifies in the higher probability quantiles but is rather good calibrated for the other quantiles. However, the calibration plot for the moderate return class is way off the striped calibration line. When looking to the corresponding probability histogram we see that it is highly skewed around 33% indicating that the model fails to use the model features to correctly predict the moderate return class. This could be due to the parallelism assumption that the model imposes. This means that negative and positive returns can be approximately explained by the same coefficients, but that moderate returns relate in a different way to the predictors than negative and positive returns do.

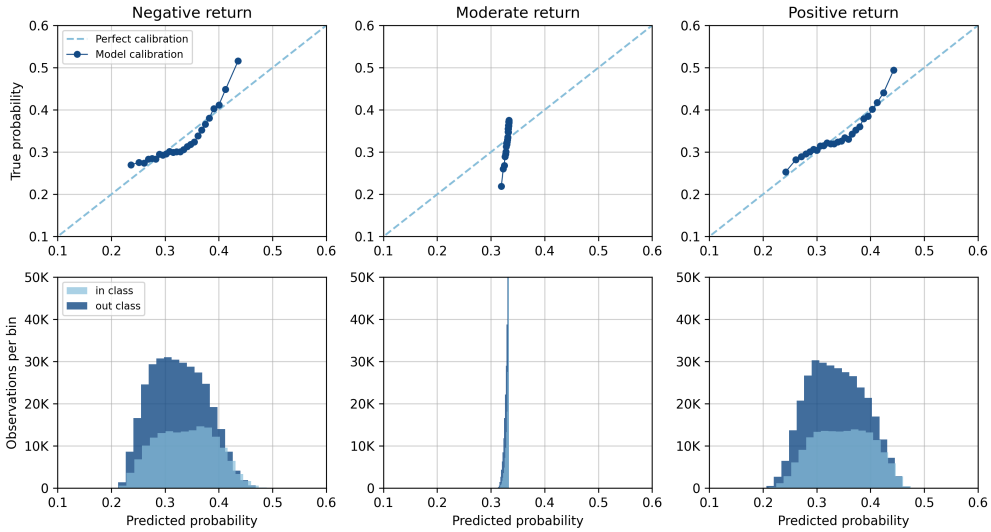

Figure 8: Morning session (09:30-12:00) calibration plots (upper row) display the in-sample ordinal probit model's calibration for negative, moderate, and positive return classes. Each dot represents the ratio of ordered predicted probabilities to true probabilities within 25 quantiles (± 6600 observations per quantile). The striped line denotes perfect calibration. The histograms of predicted probabilities (lower row) for the same classes, show observations inside and outside each return class.

Regularized non-parallel ordinal probit regression

Having established knowledge on the ordinal probit model, we examine how the L1-regularization and non-parallel form influence the model outcomes. For this, we use the objective function in Equation 4.35 to estimate model parameters $\psi_1, \psi_2, \boldsymbol{\beta_1}$ and $\boldsymbol{\beta_2}$, and select the hyperparameter λ that provides the highest in-sample log-likelihood.

From panel B in table 6 we see that the model extension succeeds in classifying each of the return classes. The accuracy for the morning and closing session is 40%, and for the afternoon 39%,

showing that the model has explanatory power. The weighted F1-scores show a similar performance, but lowers somewhat for the afternoon and closing session due to the differences between precision and recall. Based on these two metrics we can conclude that the model extension clearly improves on the baseline model for each of the sessions.

When we zoom in on the precision and recall we see that the model shows some large differences. For the morning session the model shows an equal performance on precision and recall for the negative return class of around 40%. The model shows rather good performance in accurately classifying moderate returns (41% precision), but shows less performance in detecting moderate returns (37% recall). The model shows also shows a rather good performance in detecting positive returns (44% recall) and shows descent performance in accurately classifying positive returns (40% precision). The model for the afternoon shows different performance. The negative and positive return classes show higher precision (38% and 39% respectively) than recall (36% and 31% respectively), which is low for the positive class. Hence, the model has some difficulty in detecting positive return classes in the afternoon. However, the model shows very good performance in detecting moderate return classes (50% recall) and also a descent performance in accurately classifying moderate return classes (40% precision). For the final session, we see that the model performs poorly in detecting negative returns (29% recall), but shows good performance in accurately classifying this class (41%). Hence, the model does not often predict negative returns in the closing session but when it does, it predicts them rather accurately. The model shows similar precision and recall score for the moderate returns (40% and 41% respectively). Finally, we see that the model is very good in detecting positive return classes (49% recall) and does a descent job in accurately classifying them (40%).

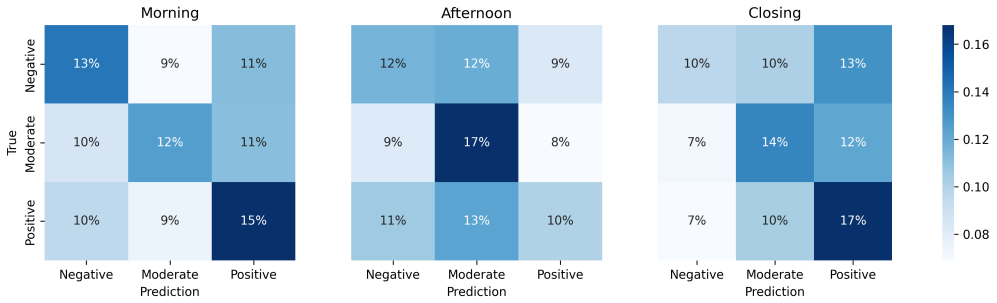

Figure 9: Confusion matrices for each session of the in-sample regularized non-parallel ordinal probit model. Entries are shown in percentages of the total number of observations.

Next, we evaluate the confusion matrices for each of the sessions. For the morning session the majority of the classes is predicted in the diagonals, which confirms the model's informative power. For the afternoon, the model succeeds in correctly predicting the moderate return classes (17% of

the observations fall in that class), but shows less performance for the positive (12%) and negative (10%) return classes. For the closing session, the model succeeds in classifying the moderate (14%) and positive (17%) return classes, but shows less performance on the negative return class (10%). Next to the correctly predicted classes, we pay special attention to two critical classes in the confusion matrix. The observations where the model predicts the negative return class while the returns were positive (NP), in the lower left corner of the confusion matrix, and the observations where the model predicts the positive return class but the returns were negative (PN), in the upper right corner. In order to evaluate the performance we need to evaluate these percentages with their corresponding diagonals (NN & PP). We see that this is especially an issue for NP and PN in the morning, NP in the afternoon and PN in the closing. We can conclude that the model performs less in making a distinction between positive and negative return classes.

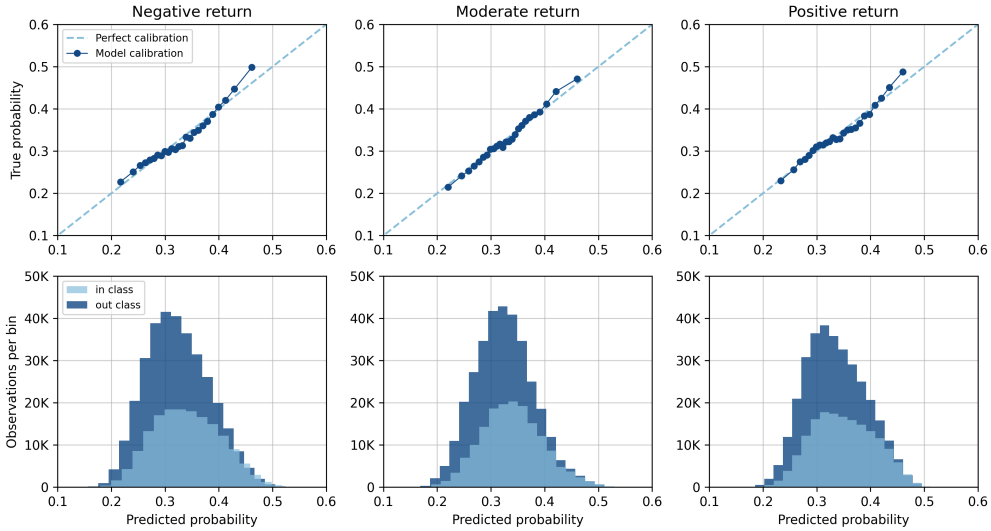

Figure 10: Morning session (09:30-12:00) calibration plots (upper row) display the in-sample regularized non-parallel ordinal probit model's calibration for negative, moderate, and positive return classes. Each dot represents the ratio of ordered predicted probabilities to true probabilities within 25 quantiles (± 6600 observations per quantile). The striped line denotes perfect calibration. The histograms of predicted probabilities (lower row) for the same classes, show observations inside and outside each return class.

To evaluate the model qualitatively we take a look at the calibration plots and predicted probability histograms in Figure 10. The three upper calibration plots show that the model extension is really well calibrated, since the model calibration neatly follows the perfect calibration line. Only the upper tale of the negative and positive return classes is somewhat overestimated. overall we can say that in each of the quantiles the ratio between the predicted probabilities and the true probabilities is correct. However, this does not mean that all the observations are correctly classified. The lower row of histograms shows that the model has difficulty to distinguish between observations

that fall into the class and outside the class. It gives many of the observations falling outside the class the same probabilities as observations falling inside the class. In an ideal situation the observations falling outside the class should obtain on average lower predicted probabilities and the ones falling inside the class higher predicted probabilities. Also, we would like to see that there are less observations predicted around 33%. Predicting observations around 33% would mean less informative power for those observations.

For a model to work accurately in executing trade orders, it is important to have higher precision for the negative and positive returns. If the model does not always detect each positive or negative return, but when it detects them, it does this more accurately, this works favourably for order execution. When working with a baseline VWAP-algorithm and differ from this execution when we have high probabilities for a negative or positive return class can enhance optimal execution.

Predictor analysis

To gain insights into the behavior of the proposed predictor variables, we assess the in-sample coefficients of the regularized non-parallel ordinal probit model. Figure 11 presents bar plots showing the estimated coefficients for the morning, afternoon, and closing sessions. Due to the non-parallel model form, the model calculates two sets of coefficients for three return classes, assigning one coefficient to each threshold. These coefficient sets can be interpreted as follows: the first set relates to the probability of having at most negative returns, while the second set relates to the probability of having at most moderate returns. A positive coefficient indicates that an increase in the predictor variable is linked to higher probability of moving up the ordinal scale, whereas a negative coefficient suggests lower probability of such movement. Additionally, the coefficient magnitude directly corresponds to the strength of the relationship between the predictor variable and the probability of transitioning between categories; a larger magnitude implies a stronger impact on the probability. When looking at Figure 11, we observe significant impacts of dummy variables on return categories for each session. Analyzing previous returns reveals that positive overnight returns lead to increased probability of transitioning from negative to moderate returns, as well as from moderate to positive returns. However, these returns have minimal impact on afternoon and closing session returns. Positive returns in the morning positively influence the probability of transitioning from negative to moderate returns and from moderate to positive returns.

Regarding volatility, higher overnight volatility raises the probability of transitioning from negative to moderate returns but reduces the probability of transitioning from moderate to positive returns. This suggests a diminishing volatility effect after reaching the moderate return class. A similar

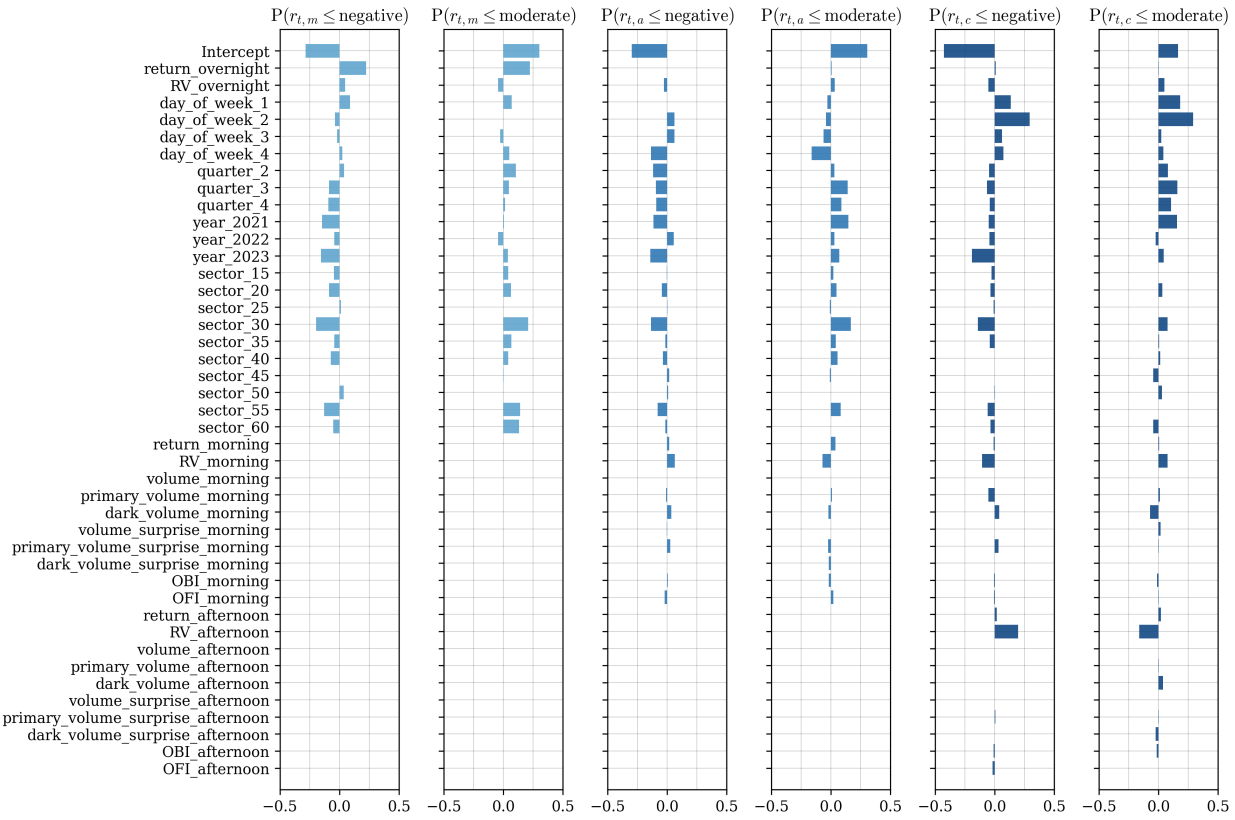

Figure 11: Coefficient plots for returns in the morning ($r_{t,m}$), afternoon ($r_{t,a}$), and closing ($r_{t,c}$) sessions using the regularized non-parallel ordinal probit model. This model categorizes intraday returns into three classes, leading to the estimation of coefficients for two thresholds.

pattern is observed with morning volatility's effect on afternoon returns. Moreover, increased morning volatility decreases the probability of transitioning from negative to moderate returns in the closing session but enhances the probability of transitioning from moderate to positive returns in the closing session. This could signify that higher price volatility tends to result in more extreme returns during the closing session, either positive or negative. Notably, afternoon volatility significantly impacts closing session returns; higher afternoon volatility increases the probability of transitioning from negative returns to moderate returns while decreasing the probability of transitioning from moderate returns to positive returns. This implies that after a volatile afternoon session, prices tend to stabilize during the closing session.

Additionally, a slight impact of dark volume surprise is noted in the afternoon. When there is a different volume traded in the dark pools than expected based on the last 25 trading days, this decreases the probability of transitioning from the moderate return class to the positive return class, whereas higher volume in dark pools in the afternoon increases the probability of transitioning from moderate returns to positive returns. This suggests that the timing of heightened dark volume

in the afternoon is important and sudden spikes in this volume create a counterbalancing effect, decreasing the probability of transitioning from moderate to positive returns. Lastly, a higher OFI in the morning lowers the probability of transitioning from negative to moderate returns but increases the probability of transitioning from moderate to positive return classes. This indicates that afternoon returns tend to lean towards extremes, whether positive or negative. For afternoon OFI, a small negative effect on transitioning from negative to moderate returns in the closing session is observed.

5.3 Economic relevance

In this section, we evaluate the in-sample economic relevance of the short-term alpha signals and the regularized non-parallel ordinal probit model. We utilize the simulation setup described in Section 4.7 and evaluate buy and sell orders for the full cross-section of MSCI US constituents during the morning, afternoon, and closing sessions. We compare their performance in BPS difference from a standard VWAP execution strategy. Figure 12 displays the resulting plots of the simulations in cumulative BPS.

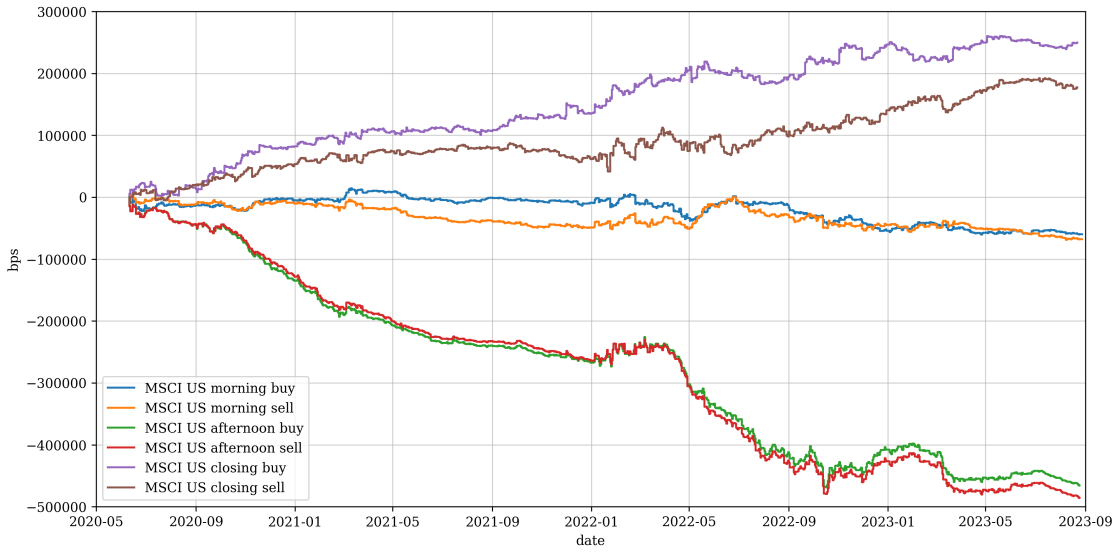

Figure 12: In-sample improvement over time of the regularized non-parallel ordinal regression model compared to standard VWAP execution (measured in basis points) for buy and sell orders across all stocks in the MSCI US Index. Simulations are performed for orders arriving before the morning, afternoon, and closing sessions.

We observe that buy and sell orders for each session perform similarly overall, but show occasional deterioration. The best-performing strategy is for buy and sell orders arrived before the closing session, improving upon the standard VWAP strategy by around 200,000 BPS. Both strategies perform similarly but diverge around the 4th quarter of 2020 and 2021.

Conversely, the worst performance is for buy and sell orders that arrive before the afternoon session. The strategy for these orders consistently underperform compared to the standard VWAP strategy. Orders arriving before the morning session show slightly more stability but still execute worse than standard VWAP, ending around 75,000 BPS lower than this benchmark.

Overall, orders arriving before the shortest session, namely the closing, can be executed more effectively. This could be attributed to better precision for negative and positive returns compared to other trading sessions. We can partly explain the poor performance by noting that although the model achieved high recall and precision for the moderate return classes, it did not fully leverage opportunities to improve upon VWAP, particularly when correctly classifying negative and positive returns and act upon these classifications by back or front loading.

5.4 Out-of-sample model evaluation

Having established understanding of the in-sample performance of the regularized non-parallel ordinal probit model, we now shift our focus to the out-of-sample evaluation. As detailed in Section 4.8, we employ 5-fold time series cross-validation to repeatedly estimate and predict out-of-sample returns. Table 7 presents the quantitative evaluation metrics for this model.

	Morning			Afternoon			Closing		
	Negative	Moderate	Positive	Negative	Moderate	Positive	Negative	Moderate	Positive
Precision	39.03	40.70	40.03	35.95	38.44	37.69	35.58	39.14	38.80
Recall	43.72	36.10	39.55	44.98	37.42	29.33	23.61	18.78	67.86
F1-score	41.24	38.26	39.79	39.96	37.93	32.99	28.39	25.38	49.37
Accuracy			39.84			37.16			38.17
Weighted F1			39.79			36.88			35.06

Table 7: Evaluation metrics in percentages for out-of-sample return classification using the regularized non-parallel ordinal probit model.

We note that the morning session exhibits the highest accuracy and weighted F1-score, approximately 40%. This marks a slightly inferior performance compared to the in-sample classification. In the afternoon session, we observe an accuracy of around 37%. Here, the model struggles to identify the positive return class, shown by the recall of 29%. Nonetheless, the performance is still better than randomly assigning one of three return classes. The closing session clearly demonstrates lower out-of-sample performance than in-sample results. The model clearly has difficulty in recognizing negative and moderate return classes, as shown by the low recall scores for these categories.

In addition to quantitative metrics, we conduct a simulation exercise similar to that discussed in Section 5.3, but now using out-of-sample return classifications to adjust order execution. Figure 13 displays the cumulative outperformance in BPS upon the standard VWAP-execution.

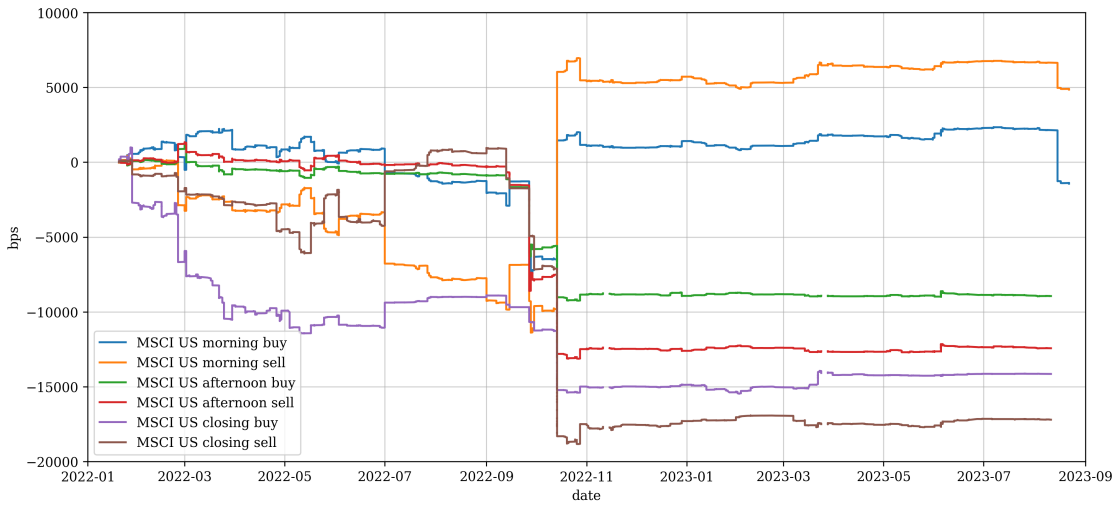

Figure 13: Out-of-sample improvement over time of the regularized non-parallel ordinal regression model compared to standard VWAP execution (measured in BPS).

We observe that sell orders arriving before the morning session are the only simulation group that outperforms standard VWAP-execution. A spike around November 2022 largely drives these differences. During this period, the morning session predictors correctly identified return classes, whereas the afternoon and closing sessions experienced a large number of misclassifications. Taking into account these findings along with the earlier in-sample results, it seems likely that the model is overfitting in the afternoon and closing sessions.

6 Conclusion

Executing large trades throughout a trading day poses a significant challenge due to market impact. Therefore, both professional traders and academics have sought intraday alpha signals to make more informed decisions on trade execution. In this thesis, we contribute to this body of research by addressing the question: *How can intraday alpha signals improve the execution of intraday equity trades?* We demonstrate that this can be achieved using a novel methodology for ordinal return classes and a large set of predictors, among which historical returns prove to be the most important ones.

Our methodology involves portfolio sorts and Fama-MacBeth regressions to identify premia related to historical returns. In doing so we identified significant reversal signals between overnight and morning returns, as well as between afternoon and closing returns. Specifically, stocks with negative returns during the overnight period carry a premium over the morning period, and similarly, stocks with negative returns in the afternoon carry a premium over the closing period. This pattern remains consistent even when adding size, volatility, and volume as control variables in the Fama-MacBeth regression setting. Thus, we demonstrate that historical returns can be used for intraday alpha signals in the cross-section.

In our proposed model setting, we show how predicting returns can be successfully changed in a classification task using existing ordinal probit models, and how this opens a wide range of interpretable evaluation metrics. For the proposed models, we found that the ordinal probit model struggles in identifying the moderate return class. However, incorporating the non-parallel form and L1-penalty appears to help in correctly identifying each class. Therefore, the flexibility of the non-parallel model form is crucial, as it appears that the relationships between predictor variables and return classes cannot be adequately captured by a parallel model form. In-sample we observe that next to historical returns also newly proposed predictors were selected by this model extension such as dark volume surprise, OBI and OFI.

We evaluate economic relevance of the regularized non-parallel ordinal probit model by conducting order simulations for both buy and sell orders and evaluating its profitability compared to a standard VWAP-strategy. From the order simulation study we may conclude that higher precision for negative and positive return classes can be associated with better economic performance. For instance, in the closing session, where the model exhibits higher precision for positive and negative return classes, the benchmark VWAP-strategy is outperformed. On the other hand, the model's high precision and recall for the moderate return class in the afternoon suggest it can better dis-

tinguish when to use the benchmark VWAP execution during that period. However, we see that this performance does not give an economical edge, resulting in a significant underperformance of the VWAP-benchmark.

Finally, we evaluate the out-of-sample performance of this model. Here, we see that the regularized non-parallel ordinal probit regression model is still able to classify morning returns, but has more difficulty in correctly classifying afternoon and closing returns. It also has difficulty outperforming a standard VWAP-benchmark in an order simulation. A possible reason for this could be overfitting on the set of predictor variables.

In this thesis, we aimed to contribute to the research gap defined by Goldstein et al. (2021) between long-term and extreme short-term behavior of equity returns. We simultaneously propose a novel method for return classification and show how this method could be beneficial for traders in practice.

The research has several limitations. Firstly, the analysis is conducted over a relatively short sample period, using data from September 2020 to September 2023, a period largely influenced by the COVID-19 pandemic. We would recommend using a longer time span to provide a more comprehensive understanding of intraday market movements across various economic cycles, including periods of volatility, stability, growth, and recession. Secondly, for future research it would be advisable to assess prediction and execution horizons at a higher and consistent frequency. This research predicts returns for morning, afternoon, and closing sessions, requiring separate analyses due to the various lengths of these sessions. Transitioning to a uniform time interval, such as 15 minutes, would facilitate clearly defined feature lags and simplify the evaluation of predictor variables. Additionally, certain features are likely to exhibit stronger predictive power over shorter time horizons. For instance, order-flow imbalance demonstrates high predictive power over a one-minute horizon but this power decreases as the prediction horizon extends, as shown by Cont et al. (2023). Thirdly, extending the set of predictors would be valuable. Incorporating economical relevant news events as predictors, as demonstrated by Chinco et al. (2019), could provide valuable insights. Moreover, analyzing the co-movement among different stocks within a cross-section could provide more areas to explore. Lastly, a comprehensive study is needed to evaluate the model's relevance in actual order execution scenarios. This entails considering buy and sell orders of varying sizes throughout the day and evaluating performance across diverse market conditions or portfolio characteristics like market capitalization, trading volume, and volatility.

References

Ahn, H., Bae, K., & Chan, K. (2001, April). Limit orders, depth, and volatility: Evidence from the stock exchange of hong kong. *The Journal of Finance*, *56*(2), 767–788. doi: 10.1111/0022-1082.00345

Aleti, S. (2022). The high-frequency factor zoo. *SSRN Electronic Journal*. doi: 10.2139/ssrn.4021620

Aleti, S., Bollerslev, T., & Siggaard, M. (2023). Intraday market return predictability culled from the factor zoo. *SSRN Electronic Journal*. doi: 10.2139/ssrn.4388560

Andersen, T. G., & Bollerslev, T. (1998, November). Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. *International Economic Review*, *39*(4), 885. doi: 10.2307/2527343

Andersen, T. G., Bollerslev, T., Diebold, F. X., & Labys, P. (2000). Great realizations. *RISK*, *13*, 105–108.

Aït-Sahalia, Y., Fan, J., Xue, L., & Zhou, Y. (2022). *How and when are high-frequency stock returns predictable?* (Working paper). National Bureau of Economic Research.

Bachelier, L. (1900). Théorie de la spéculation. *Annales scientifiques de l'École normale supérieure*, *17*, 21–86. doi: 10.24033/asens.476

Bali, T. G., Engle, R. F., & Murray, S. (2016). *Empirical asset pricing.* Hoboken, New Jersey: Wiley.

Baltussen, G., Da, Z., Lammers, S., & Martens, M. (2021). Hedging demand and market intraday momentum. *SSRN Electronic Journal*. doi: 10.2139/ssrn.3760365

Blitz, D., Hanauer, M. X., Honarvar, I., Huisman, R., & van Vliet, P. (2023, apr). Beyond fama-french factors: Alpha from short-term signals. *Financial Analysts Journal*, 1–22. doi: 10.1080/0015198x.2023.2173492

Bouchaud, J.-P., Bonart, J., Donier, J., & Gould, M. (2018). *Trades, quotes and prices: financial markets under the microscope.* Cambridge University Press.

Buti, S., Rindi, B., & Werner, I. M. (2022, April). Diving into dark pools. *Financial Management*, *51*(4), 961–994. doi: 10.1111/fima.12395

Cartea, Á., Donnelly, R., & Jaimungal, S. (2018, January). Enhancing trading strategies with order book signals. *Applied Mathematical Finance*, *25*(1), 1–35. doi: 10.1080/1350486x.2018.1434009

Cartea, Á., Jaimungal, S., & Penalva, J. (2015). *Algorithmic and high-frequency trading.* Cambridge

University Press.

Chen, J.-H., & Tsai, Y.-C. (2020, June). Encoding candlesticks as images for pattern classification using convolutional neural networks. *Financial Innovation*, *6*(1). doi: 10.1186/s40854-020-00187-0

Chinco, A., Clark-Joseph, A. D., & Ye, M. (2019, November). Sparse signals in the cross-section of returns. *The Journal of Finance*, *74*(1), 449–492. doi: 10.1111/jofi.12733

Chordia, T., Roll, R., & Subrahmanyam, A. (2002, jul). Order imbalance, liquidity, and market returns. *Journal of Financial Economics*, *65*(1), 111–130. doi: 10.1016/s0304-405x(02)00136-8

Chu, X., Gu, Z., & Zhou, H. (2019, sep). Intraday momentum and reversal in chinese stock market. *Finance Research Letters*, *30*, 83–88. doi: 10.1016/j.frl.2019.04.002

Cochrane, J. (2009). Regression-based tests of linear factor models. In *Asset pricing: Revised edition.* Princeton university press.

Cochrane, J. H. (2011, jul). Presidential address: Discount rates. *The Journal of Finance*, *66*(4), 1047–1108. doi: 10.1111/j.1540-6261.2011.01671.x

Cohen, N., Balch, T., & Veloso, M. (2020, October). Trading via image classification. In *Proceedings of the first acm international conference on ai in finance.* ACM. doi: 10.1145/3383455.3422544

Cont, R., Cucuringu, M., & Zhang, C. (2023, aug). Cross-impact of order flow imbalance in equity markets. *Quantitative Finance*, *23*(10), 1373–1393. doi: 10.1080/14697688.2023.2236159

Cont, R., Kukanov, A., & Stoikov, S. (2014). The price impact of order book events. *Journal of financial econometrics*, *12*(1), 47–88.

Elaut, G., Frömmel, M., & Lampaert, K. (2018, jan). Intraday momentum in FX markets: Disentangling informed trading from liquidity provision. *Journal of Financial Markets*, *37*, 35–51. doi: 10.1016/j.finmar.2016.09.002

Fama, E. F., & French, K. R. (1992, June). The cross-section of expected stock returns. *The Journal of Finance*, *47*(2), 427–465. doi: 10.1111/j.1540-6261.1992.tb04398.x

Fama, E. F., & French, K. R. (2015, apr). A five-factor asset pricing model. *Journal of Financial Economics*, *116*(1), 1–22. doi: 10.1016/j.jfineco.2014.10.010

Fama, E. F., & MacBeth, J. D. (1973, may). Risk, return, and equilibrium: Empirical tests. *Journal of Political Economy*, *81*(3), 607–636. doi: 10.1086/260061

Feng, G., Giglio, S., & Xiu, D. (2020, February). Taming the factor zoo: A test of new factors. *The Journal of Finance*, *75*(3), 1327–1370. doi: 10.1111/jofi.12883

Freyberger, J., Neuhierl, A., & Weber, M. (2020). Dissecting characteristics nonparametrically. *The Review of Financial Studies*, *33*(5), 2326–2377.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, *33*(1), 1.

Gao, L., Han, Y., Li, S. Z., & Zhou, G. (2018, aug). Market intraday momentum. *Journal of Financial Economics*, *129*(2), 394–414. doi: 10.1016/j.jfineco.2018.05.009

Goldstein, I., Spatt, C. S., & Ye, M. (2021, apr). Big data in finance. *The Review of Financial Studies*, *34*(7), 3213–3225. doi: 10.1093/rfs/hhab038

Gu, S., Kelly, B., & Xiu, D. (2020, feb). Empirical asset pricing via machine learning. *The Review of Financial Studies*, *33*(5), 2223–2273. doi: 10.1093/rfs/hhaa009

Harris, L. E., & Panchapagesan, V. (2005, February). The information content of the limit order book: evidence from nyse specialist trading decisions. *Journal of Financial Markets*, *8*(1), 25–67. doi: 10.1016/j.finmar.2004.07.001

Heston, S. L., Korajczyk, R. A., & Sadka, R. (2010, jul). Intraday patterns in the cross-section of stock returns. *The Journal of Finance*, *65*(4), 1369–1407. doi: 10.1111/j.1540-6261.2010.01573.x

Huang, T., & Zhang, X. (2022, January). Industry-level media tone and the cross-section of stock returns. *International Review of Economics amp; Finance*, *77*, 59–77. doi: 10.1016/j.iref.2021.09.002

Huddleston, D., Liu, F., & Stentoft, L. (2023). Intraday market predictability: A machine learning approach. *Journal of Financial Econometrics*, *21*(2), 485–527. doi: 10.1093/jjfinec/nbab007

Jegadeesh, N., & Titman, S. (1993, mar). Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of Finance*, *48*(1), 65–91. doi: 10.1111/j.1540-6261.1993.tb04702.x

Jegadeesh, N., & Wu, Y. (2022, mar). Closing auctions: Nasdaq versus NYSE. *Journal of Financial Economics*, *143*(3), 1120–1139. doi: 10.1016/j.jfineco.2021.12.003

Jin, M., Kearney, F., Li, Y., & Yang, Y. C. (2019, dec). Intraday time-series momentum: Evidence from china. *Journal of Futures Markets*, *40*(4), 632–650. doi: 10.1002/fut.22084

Jones, C. S., & Mo, H. (2020, mar). Out-of-sample performance of mutual fund predictors. *The Review of Financial Studies*, *34*(1), 149–193. doi: 10.1093/rfs/hhaa026

Kakushadze, Z. (2014). 4-factor model for overnight returns. *SSRN Electronic Journal*. doi: 10.2139/ssrn.2511874

Kandel, E., Rindi, B., & Bosetti, L. (2012). The effect of a closing call auction on market quality

and trading strategies. *Journal of Financial Intermediation*, *21*(1), 23–49.

Kelly, B. T., Pruitt, S., & Su, Y. (2019). Characteristics are covariances: A unified model of risk and return. *Journal of Financial Economics*, *134*(3), 501–524.

Kozak, S., Nagel, S., & Santosh, S. (2020, February). Shrinking the cross-section. *Journal of Financial Economics*, *135*(2), 271–292. doi: 10.1016/j.jfineco.2019.06.008

Lou, D., Polk, C., & Skouras, S. (2019, oct). A tug of war: Overnight versus intraday expected returns. *Journal of Financial Economics*, *134*(1), 192–213. doi: 10.1016/j.jfineco.2019.03.011

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B (Methodological)*, *42*(2), 109–127.

McLean, R. D., & Pontiff, J. (2016, jan). Does academic research destroy stock return predictability? *The Journal of Finance*, *71*(1), 5–32. doi: 10.1111/jofi.12365

Murphy, D. P., & Thirumalai, R. S. (2017, dec). Short-term return predictability and repetitive institutional net order activity. *Journal of Financial Research*, *40*(4), 455–477. doi: 10.1111/jfir.12131

Newey, W. K., & West, K. D. (1987, may). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, *55*(3), 703. doi: 10.2307/1913610

Said, E. (2022). *Market impact: Empirical evidence, theory and practice.* arXiv. doi: 10.48550/ARXIV.2205.07385

Sharpe, W. F. (1964, sep). Capital asset prices a theory of market equilibrium under conditions of risk. *The Journal of Finance*, *19*(3), 425–442. doi: 10.1111/j.1540-6261.1964.tb02865.x

Tibshirani, R. (1996, January). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *58*(1), 267–288. doi: 10.1111/j.2517-6161.1996.tb02080.x

Wen, Z., Bouri, E., Xu, Y., & Zhao, Y. (2022). Intraday return predictability in the cryptocurrency markets: Momentum, reversal, or both. *SSRN Electronic Journal*. doi: 10.2139/ssrn.4080253

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data.* MIT press.

Wurm, M. J., Rathouz, P. J., & Hanlon, B. M. (2021). Regularized ordinal regression and the ordinalnet r package. *Journal of Statistical Software*, *99*(6). doi: 10.18637/jss.v099.i06

Yee, T. W. (2010). Thevgampackage for categorical data analysis. *Journal of Statistical Software*, *32*(10). doi: 10.18637/jss.v032.i10

Zhang, Y., Ma, F., & Zhu, B. (2019, jan). Intraday momentum and stock return predictability:

Evidence from china. *Economic Modelling*, *76*, 319–329. doi: 10.1016/j.econmod.2018.08.009

Tables

A Appendix

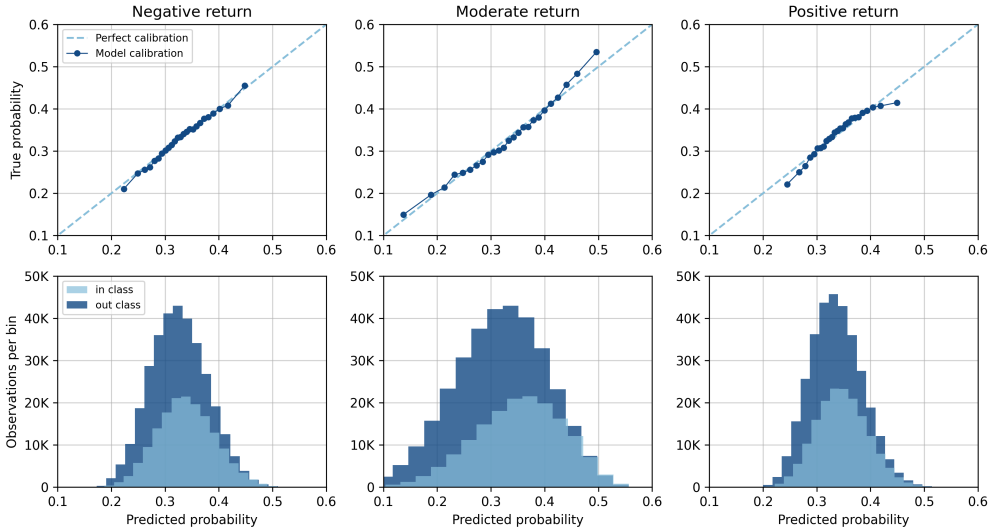

Figure 14: Afternoon session (12:00-15:50) calibration plots (upper row) display the in-sample regularized non-parallel ordinal probit model's calibration for negative, moderate, and positive return classes. Each dot represents the ratio of ordered predicted probabilities to true probabilities within 25 quantiles (± 6600 observations per quantile).

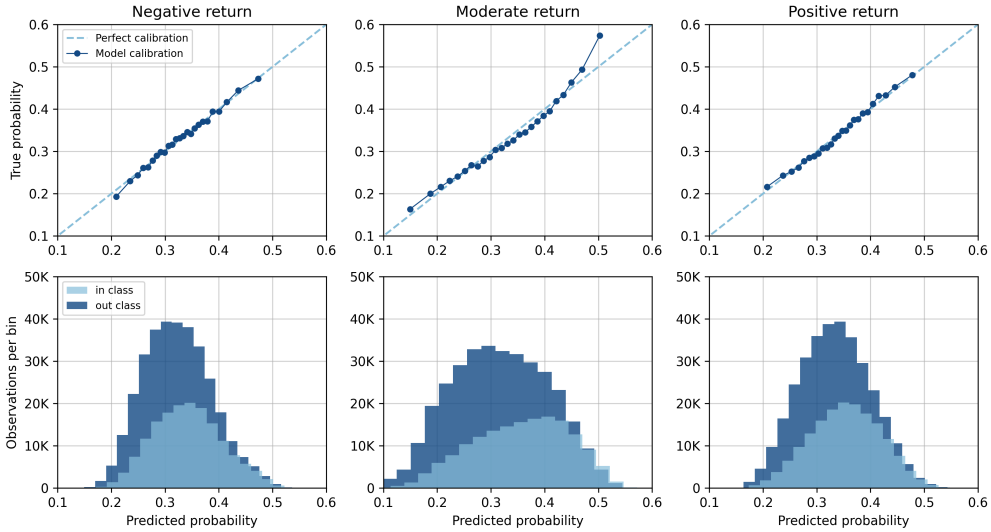

Figure 15: Closing session (15:50-16:05) calibration plots (upper row) display the in-sample regularized non-parallel ordinal probit model's calibration for negative, moderate, and positive return classes. Each dot represents the ratio of ordered predicted probabilities to true probabilities within 25 quantiles (± 6600 observations per quantile).