

Data Driven Quantitative Methods for Financial Forecasting Problems



Nikolas Michael
Kellogg College
University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2024

Acknowledgements

I would like to express my sincere gratitude to my supervisors, Mihai Cucuringu and Sam Howison, for their wisdom, patience, consistent support, and generosity throughout the last four years. It was a blessing to work with them, and their academic, professional, and personal lessons will stay with me for the rest of my life. Special thanks to Stefan Zohren and Alexander Y. Shestopaloff for being part of my viva committee; and to Rama Cont and Jan-Peter Calliess for serving as examiners on previous occasions. I want to thank EPSRC for funding this degree, and Kellogg College for all the support. I extend my appreciation to Boltzmann Research and Chariton Christou for providing data, computing resources, and practical guidance. I owe my deepest gratitude to my parents, Michalis and Aphrodite, and to my grandmother, Maroulla, for their unwavering support and sacrifices throughout the years. Lastly, I thank God for the inner strength and guidance that carried me through this journey.

Abstract

With increasing computational resources, data science and machine learning methods are becoming increasingly relevant in quantitative analysis. This thesis focuses on forecasting financial time series using data-driven methods. The contributions are divided into four parts, each based on different data sets, forecasting tasks, and methodologies.

In the first part, we develop a feature based on option volumes decomposed by distinct market participant classes to predict directional price movements in the spot market. We conduct a detailed analysis demonstrating the robustness of our methodology and providing insights into the option contracts and orders with the highest predictability.

In the second part, we introduce OFTER, a time series forecasting pipeline tailored for financial problems centered on mid-sized multivariate time series. OFTER is designed for online tasks, utilizing non-parametric models and avoiding the curse of dimensionality.

In the third part, we augment the Heterogeneous Autoregressive Regression model for forecasting realized volatility. We employ various estimators for daily, weekly, and monthly volatilities, focusing on the utilization of option price data and implied volatilities based on the Black-Scholes and Heston models.

In the final part, we formulate a network framework for forecasting problems related to E-mini S&P 500 and CBOE Volatility Index futures. By combining a multi-channel Graph Convolutional Network with a Long Short-Term Memory network, we enhance predictive performance across different forecasting problems.

Contents

1	Introduction and Background	1
1.1	Theoretical Background	2
1.1.1	Efficient Market Hypothesis	2
1.1.2	Market Excess Returns and Factor Models	5
1.2	Prediction Methods	6
1.2.1	Networks	7
1.3	Derivative Products	8
1.4	Contributions	10
1.4.1	Option Volume Imbalance as a Predictor for Equity Returns .	12
1.4.2	OFTER: An Online Pipeline for Time Series Forecasting . . .	12
1.4.3	Options-driven Volatility Forecasting	13
1.4.4	A GCN-LSTM Approach for ES-mini and VX Futures Fore- casting	13
1.5	Outline and Note on Notation:	14
2	Option Volume Imbalance as a Predictor for Equity Returns	15
2.1	Introduction	15
2.2	Motivation and the Option Volume Imbalance	17
2.2.1	Related Literature	17
2.2.2	Option Volume Imbalance	20
2.3	NASDAQ Data Sets	22
2.3.1	Data Sets Description	22
2.3.2	Descriptive Analysis of the Market	22
2.3.3	Correlation of Volume and OVI	25
2.3.4	Principal Component Analysis	26
2.4	Evaluation Strategy	27
2.4.1	Profitability and Predictability	27
2.4.2	Types of Return and Evaluation Metrics	28

2.4.3	P&L Regression	31
2.5	Main Empirical Results	34
2.5.1	Comparison across MPCs	35
2.5.2	Comparison across Return Types	37
2.5.3	Comparison Across OVI Types	38
2.5.4	Multi-factor Analysis and Stability Analysis across time	38
2.6	Influence of Options Features on OVI Predictability	40
2.7	Extensions	46
2.7.1	Short-Term Versus Long-Term Price Persistence	46
2.7.2	Difference Between Buys and Sells	47
2.7.3	Robustness Under Different Market Conditions	48
2.7.4	Additional Features from the NOTO Data Set	49
2.7.5	Cross-Impact Analysis	51
2.8	Conclusion and Future Research Directions	53
3	OFTER: An Online Pipeline for Time Series Forecasting	60
3.1	Introduction	60
3.2	Preliminaries and Related Works	63
3.2.1	K-nearest Neighbors and Generalized Regression Neural Network	64
3.2.2	Dimensionality Reduction	65
3.2.3	Maximal Correlation	66
3.3	Methodology	69
3.3.1	Online Update of the Embedding	69
3.3.2	Choice of Hyperparameters	72
3.3.3	Overcoming the Shortcomings of kNN and GRNN	75
3.3.4	Algorithm Outline	76
3.4	Empirical Results	76
3.4.1	Baseline Methods	78
3.4.2	Synthetic Data Sets	79
3.4.3	Data	80
3.4.4	Evaluation Metrics	82
3.4.5	Numerical Results	83
3.4.5.1	Synthetic data	84
3.4.5.2	ETF volumes	86
3.5	Applications to Feature Importance and Outlier Detection	88
3.5.1	Feature Importance	89

3.5.2	Outlier Detection	91
3.6	Conclusion and Future Work	92
4	Options-driven volatility forecasting	94
4.1	Introduction	94
4.2	Volatility Estimators	98
4.2.1	Realized Volatility	98
4.2.2	Heterogeneous Autoregressive (HAR) Model	100
4.2.3	Historical Volatility Estimators	101
4.2.4	Path Dependent Volatility	103
4.3	Extracting Volatility Estimators from Implied Volatility Surface Data	103
4.3.1	The Black–Scholes model	103
4.3.2	Implied Volatility Surface	104
4.3.3	Dimensionality reduction for implied volatility surfaces	106
4.4	Extracting Volatility Estimators under Stochastic Volatility Models .	107
4.4.1	Heston Model	107
4.4.2	Sensitivity analysis of the Heston model parameters	108
4.4.3	Full Calibration of the Heston Model	109
4.4.4	Partial Calibration of the Heston Model	110
4.4.5	Partial Calibration of the Bates Model	111
4.4.6	Choice of Weighting Scheme	112
4.5	Forecasting Setup	113
4.5.1	Data Set	113
4.5.2	Evaluation Metrics	114
4.5.3	Regression Methods	115
4.5.4	Incorporating IVS features	116
4.5.5	Description of Feature Sets	120
4.6	Results	122
4.6.1	Cross-Asset Model	127
4.7	Conclusion	129
5	A GCN-LSTM Approach for ES-mini and VX Futures Forecasting	136
5.1	Introduction and Related Literature	136
5.1.1	Futures	136
5.1.2	SPX and VIX	136
5.1.3	Relations between Returns, Volatility and Trading Volume . .	137
5.1.4	Graph Neural Networks	138

5.1.5	Main contributions of this paper	139
5.2	Model	140
5.2.1	Graph Convolutional Networks	141
5.2.2	Long-Short Term Memory Networks (LSTM)	143
5.3	Data and Graph Construction	143
5.3.1	Data	143
5.3.2	Graph Construction	145
5.3.3	Discussion	145
5.4	Forecasting Setup	149
5.4.1	Financial Terms	149
5.4.2	Loss Functions and Evaluation Metrics	150
5.4.3	Predictive Features	151
5.4.4	Baselines	152
5.4.5	Implementation Details	153
5.5	Results	153
5.5.1	Forecasting Returns	153
5.5.2	Forecasting of Volatility	154
5.5.3	Forecasting Trading Volume	156
5.5.4	Feature Importance	158
5.5.5	Different Configurations	159
5.6	Conclusion	161
6	Conclusion and Future Work	163
6.1	Conclusion	163
6.2	Future Work	165
6.2.1	Application of GCN to an Equity Network	165
6.2.2	Extensions of Methodology for Extracting volatility under the Heston Model	166
	Bibliography	168
A	Appendix for Option Volume Imbalance as a Predictor for Equity Returns	191
A.1	Financial Definitions	192
A.2	Approximation of trade flow between MPCs	193
A.3	OVI of different Market Participant Classes: Additional Results	194
A.3.1	Contemporaneous Returns	194

A.3.2	Alternative Bet-size weighting schemes	195
A.3.3	Cumulative P&L for non-market maker participants	198
A.3.4	Proportion of “profitable” days	199
A.4	Broader Scope Analysis: Additional results	201
A.4.1	Decomposition based on Option Features	201
A.4.2	Difference between Opening and Closing Transactions	202
A.4.3	Influence of underlying equity’s features on OVI’s predictive power	202
A.4.4	Cross-Impact and Networks	203
A.5	Intra-day strategies.	205
B	Appendix for OFTER: An Online Pipeline for Time Series Forecast- ing	212
B.1	A convexifying approach for solving the Secular equation	213
B.2	Brief summary of baseline methods	213
B.3	Robustness analysis	216
B.4	Non-Financial Application	217
C	Appendix for Options-driven volatility forecasting	219
C.1	Heston Model Calibration	220
C.1.1	Particle Swarm Optimization	220
C.2	Financial Trading Metrics	222

List of Figures

2.1	Breakdown of NOTO and PHOTO into their Constituents	23
2.2	Approximation of the Volume Flow between Market Participants . .	24
2.3	OVI Correlation Map	25
2.4	Volume and OVI Principal Components	26
2.5	Volume Principal Components' Scree Plot	26
2.6	Quantile Rank Portfolio Illustration Diagram	30
2.7	Cumulative P&L Plot - Overnight Returns	35
2.8	Sharpe Ratio Significance Tests	36
2.9	Cumulative P&L Plot - Market Maker OVI	37
2.10	Comparison across OVI Types	38
2.11	Multiple-Factor Regression for Different Market Participants	39
2.12	Summary Statistics for Breakdown of Data set into Buckets Based on Greeks	42
2.13	Summary Statistics for Breakdown of Data set into Buckets Based on Option Features	43
2.14	Pairwise Sharpe Ratio Difference Tests for Buckets Based on Greeks and Option Features	44
2.15	Price Impact over Longer Holding Periods	46
2.16	SR Difference Tests for Buy / Sell transactions	48
2.17	Cumulative P&L Plot - Market Conditions	56
2.18	Detailed Overview of Predictability across Data Sets	57
2.19	Multiple Factors Regression for Broader Scope Analysis	58
2.20	Summary Information for Cross Impact Networks	59
3.1	Pictorial Illustration of OFTER	70
3.2	Effect of Smoothing Parameters on Forecasts	73
3.3	Barplot Showing Performance Comparison of OFTER with Baselines for the Synthetic Data Set	86
3.4	Cumulative P&L Plot for the ETF Price Movement Forecasting Task	88

3.5	Feature Importance Weights for Synthetic Data Set	89
3.6	Outlier Detection Algorithm using Minimum Distance	91
4.1	Implied Volatility Surface Example	105
4.2	Sensitivity of Call Option Prices to the Parameters of the Heston Model	131
4.3	Pictorial illustration for the Process of the Extraction of the Volatility Estimator for the Heston Model	132
4.4	Pictorial Illustration for the Investigated Features	132
4.5	Plots for Principal Components	133
4.6	Barplot of the Median Absolute Value of each Feature's Coefficient .	134
4.7	Box and Whisker plot of the Median Correlation of the Different Daily Volatility Estimators	135
4.8	Cumulative P&L Plot for VIXM	135
5.1	Diagram for the LSTM-ANN Module	141
5.2	Overview Diagram for the GCN-LSTM Module Used	142
5.3	Correlation Plot based on the Hourly Returns of ES and VX Futures	144
5.4	Visualization of Networks based on Returns, Volatility and Volume .	148
5.5	Barplot for the Out Degrees of the Contemporaneous Unweighted Graphs	149
5.6	Cumulative P&L Plot for the Hourly Return Forecasting Problem . .	156
A.1	Results for Contemporaneous and Lagged Forecasting	194
A.2	Cumulative Returns for Different Betting Scheme	197
A.3	Sharpe Ratio Difference Tests for Different Betting Schemes	198
A.4	Cumulative P&L plot - Firm	199
A.5	Cumulative P&L plot - Broker	200
A.6	Cumulative P&L plot - Customer	200
A.7	Cumulative P&L plot - Professional Customer	201
A.8	P&L accuracy Barplot	202
A.9	Boxplots of Sharpe Ratios by Sector	204
A.10	In and Out-Degrees of Networks	205
A.11	In and Out Degrees per Sector	206
A.12	Instruments with High Centralities	208
A.13	Median Intraday Volume	209
A.14	Intraday Returns for Firm open_t OVI	210
A.15	Intraday Returns for Market Maker open_t OVI	211
B.1	Air Pollution Forecasting Task Heatmap	218

Chapter 1

Introduction and Background

The financial system has made a huge contribution to the stabilization of the modern world and its rapid development. By ensuring a secure and efficient transfer of resources between borrowers and lenders, it can help funnel funds for development and innovation in an efficient manner. Moreover, it can grant individuals and companies security and an ability to reliably control the growth of their wealth. For example, trading allows individuals, firms, banks and other traders, to utilize their understanding of the market, to make investments catered for their capital, risk tolerance, and return requirements. **Price, price volatility, and trading volume** of instruments are often considered the three main financial quantities of interest in understanding the current state and forecasting the future state of the system.

In this chapter, we give an overview and brief literature review of the main components of the thesis, introduce performance metrics that will be used throughout the different chapters, and discuss preliminary theoretical background, as well as the some of the most commonly used methods in the literature for the financial time-series forecasting task.

Basic Setup Stock prices and returns are some of the most widely studied financial time series. This dissertation deals with multiple types of returns. One type of return which will play a big role are the close to close returns.

In particular, let $P_{i,d}^{\text{close}}$ denote the closing price of an instrument i on day d . Also denote by $D_{i,d}$ the dividends paid on day d to an owner of one stock of asset i . The daily close to close return is defined as

$$f_{i,d}^{\text{CL-tmCL}} = \frac{P_{i,d+1}^{\text{close}} + D_{i,d} - P_{i,d}^{\text{close}}}{P_{i,d}^{\text{close}}}. \quad (1.1)$$

Practitioners are quite interested in predicting directional movement of future prices. For example, if a trader correctly predicts that the price of stock i will rise, they can buy $u_{i,d}$ units of stock i at price $P_{i,d}^{\text{close}}$, or equivalently, spend a total of $b_{i,d} = u_{i,d} \times P_{i,d}^{\text{close}}$ on stock i . The trader can sell all units back the next day at $P_{i,d+1}^{\text{close}}$, making a profit (or loss) of $u_{i,d} \times (P_{i,d+1}^{\text{close}} - P_{i,d}^{\text{close}}) = b_{i,d} \times f_{i,d}^{\text{CL-tmCL}}$.

Assume an investor observes a signal $s_{i,d} \in \mathbb{R}$ concerning asset i , which they buy if $s_{i,d} > 0$, and sell if $s_{i,d} < 0$. Note that a trade usually results in trading costs, including fees and spreads, which we denote by $c_{i,d}$. Therefore, if the trader considers N_d assets for day d , the total profit resulting from trading based on signals $\{s_{i,d}\}_{i=1}^N$ is given by

$$\text{P\&L}_d = \sum_{i=1}^{N_d} b_{i,d} \times f_{i,d}^{\text{CL-tmCL}} \times \text{sign}(s_{i,d}) - c_{i,d}. \quad (1.2)$$

The above is known as the **Profit and Losses** metric. Note that the signals could either be some observed variable, such as the previous daily returns, the output of a predictive model, or simply a quantity based on the intuition of the trader. Due to the existence of costs ($c_{i,d}$), traders not only need to accurately predict the future returns' sign but further only act when the magnitude of the expected returns ($\mathbb{E}[|f_{i,d}^{\text{CL-tmCL}}|]$) is high enough to counteract the incurred costs. This can partly be controlled via the bet size $b_{i,d}$, which instead of being fixed, could be made to depend on the value of $s_{i,d}$ (for example, $b_{i,d}$ could be proportional to the absolute value of $s_{i,d}$).

Following such a strategy for a longer time period results in the vector $\mathbf{P\&L} = (\text{P\&L}_1, \text{P\&L}_2, \dots, \text{P\&L}_{D-1})$ of daily P\&L's. To assess the performance of a stock for a time period we can define the P\&L per dollar traded (PPD) as

$$\text{PPD} = \frac{\sum_{d=1}^{D-1} \text{P\&L}_d}{\sum_{d=1}^{D-1} B_d}, \quad (1.3)$$

where $B_d = \sum_{i=1}^{N_d} b_{i,d}$ is the total bet size for day d . For the rest of this chapter, $f_{i,d}$ will denote a generic unspecified type of return.

1.1 Theoretical Background

1.1.1 Efficient Market Hypothesis

The Efficient Market Hypothesis (EMH) is a central theorem of financial economics, and states that current prices accurately reflect all available information that could

indicate the future of its movements. Mathematically, if $\mathcal{F}_d$ denotes all information available at the end of day d , then EMH implies that

$$P_{i,d}^{close} = \mathbb{E} [M_{d+1} (P_{i,d+1}^{close} + D_{i,d+1}) | \mathcal{F}_d], \quad (1.4)$$

for some M_d (which is called the stochastic discount factor for day d).

Whether the EMH holds in practice, has been debated for decades with evidence given for both directions. Tests on the validity of the EMH could generally be divided into three categories, which include testing the weak-form efficiency, semi-strong form efficiency, and strong-form efficiency.

The **weak-form efficiency** asserts that all historical pricing information is fully reflected in current asset prices, implying the absence of any correlation effects in lagged returns. However, such correlations have been noted, e.g by [122], who shows the existence of negative auto-correlation for monthly returns, but positive auto-correlations for yearly returns. In a different study regarding the returns of six European countries, [38] showed positive auto-correlation in minor markets, such as the daily returns of Portugal and Greece, showing how the weak form of EMH can fail in non-major markets.

The **semi-strong form efficiency** assert that no publicly available information can be used to predict future returns, as all such information is already reflected in current pricing. Examples of exploitable information has been found, for example [25] empirically showed that stocks with low Price to Earning Ratios (P/E) tend to have higher returns on average, [3] showed the under-performance of stocks post-acquisition, and [140] showed that dumped stocks (stocks without recent analyst coverage) tend to outperform covered stocks.

According to the **strong-form efficiency**, no information could be used to predict future returns, whether that is public or private information. However, [210] used data from the Securities and Exchange Commission (SEC), to show that some trades used insider information to make profits.

Note that arguments have been given against any of these findings being contradictions of the EMH. Furthermore, any finding on exploitable ways to make a profit, will lead to investors adopting to take advantage of it, forcing the market prices to adjust by taking that information into account.

Modern Portfolio Theory Modern Portfolio Theory (MPT) is a prominent mathematical finance framework. Consider a trader who has to choose between portfolio A and B, with future returns $f_{A,d}$ and $f_{B,d}$ respectively. Under this framework,

$\mathbb{E}[f_{A,d}] \leq \mathbb{E}[f_{B,d}]$, is not sufficient reasoning for a trader to choose investing in portfolio A as it might be a riskier portfolio than portfolio B. Instead, the risk of an investment is accounted for, and is modeled by the variance of the future returns. If $\mathbb{E}[f_{A,d}] \leq \mathbb{E}[f_{B,d}]$, $\text{Var}(f_{A,d}) \geq \text{Var}(f_{B,d})$, and one of the two inequalities is strict, the investor should never choose portfolio A, as it is a strictly worst investment than portfolio B.

A useful concept in understanding the trade-off between expected returns and risk is the utility as a function of wealth. The utility of investors is often modeled with the quadratic utility function defined as

$$U(W) = W - \frac{1}{2}\lambda W^2, \quad (1.5)$$

where W represents wealth and λ represents the risk aversion coefficient, which captures the declining marginal utility of wealth and the increasing marginal disutility of risk.

Sharpe Ratio Based on the MPT, [212] used the reward-to-variability ratio equal to the expected return divided by its standard deviation, as a measure which both gives preference to high expected returns, and low variance. This metric has since been called the Sharpe Ratio and can be defined both as an ex-ante metric (with expected returns) and an ex-post metric (using historic sample mean and variance). We will be using the annualized version of its ex-post form defined as

$$\text{SR} = \frac{\text{mean}(\mathbf{P\&L}) \times \sqrt{252}}{\text{sd}(\mathbf{P\&L})}, \quad (1.6)$$

to assess the performance of hypothetical strategies. Note that the Sharpe-Ratio can be loosely associated with a t-statistic, when testing the hypothesis that the true mean of a normally distributed random variable is equal to 0. In this case, for a sample of length n following a normal distribution with sample mean $\bar{x}$ and sample standard deviation $\bar{\sigma}$, the t-statistic would be equal to

$$\frac{\bar{x}}{\left(\frac{\bar{\sigma}}{\sqrt{n}}\right)}. \quad (1.7)$$

The constant 252 refers to the usual number of trading days in a financial year.

1.1.2 Market Excess Returns and Factor Models

An important concept is that of market return, which acts as an indicator of the overall price movement of the market. Let $E_{i,d}$ denote the market cap of asset i for day d (which is defined as the product of shares outstanding times the current price per share). We then define the market return for the d^{th} day as

$$f_{\text{Market},d} = \sum_{i=1}^n f_{i,d} \left(\frac{E_{i,d}}{\sum_{k=1}^n E_{k,d}} \right). \quad (1.8)$$

It is usual practice to evaluate strategies by considering stock's price movements relative to the market price movement. We approximate market returns using the S&P-500 index, which is a widely-used proxy for market returns, due to its wide range of large-cap firms and quarterly portfolio updates. Therefore, we consider the excess market returns for asset i at day d , defined as

$$\tilde{f}_{i,d} = f_{i,d} - f_{\text{Market},d}.$$

The **Capital Asset Pricing Model (CAPM)** is a prominent model, which (under a number of economic assumptions) suggests that the expected excess return of an instrument is proportional to the expected excess return of the market. The constant of proportionality is known as the β of the asset, which is in turn proportional to the covariance of the asset return and the market return.

$$\mathbb{E}[f_{i,d}] = f_{\text{Risk-free},d} + \beta_i (\mathbb{E}[f_{\text{Market},d}] - f_{\text{Risk-free},d}). \quad (1.9)$$

The CAPM has a number of shortcomings. Firstly, it is a theoretical result requiring various questionable assumptions, that are not necessarily true. Secondly, the CAPM is only valid for one-period models. Lastly, estimating the covariance matrix, is not a trivial task. The latter, led [212] to formulate what is known as the **Single Index Model**, which is the statistical counterpart of CAPM. Unlike the CAPM, the single index model is a regression model, which does not require the exact covariance matrix and does not require all the assumptions of the CAPM. Furthermore, it allows for a non-zero intercept term, which can be interpreted as the “abnormal returns” parameter

$$f_{i,d} - f_{\text{Risk-free},d} = \alpha_i + \beta_i (f_{\text{Market},d} - f_{\text{Risk-free},d}) + \epsilon_{i,d}, \quad (1.10)$$

where

$$\epsilon_{i,d} \sim N(0, \sigma^2) \text{ for some } \sigma^2 \geq 0 \text{ and } \text{Cov}(\epsilon_{i,d}, \epsilon_{k,d}) = 0 \text{ for } i \neq k.$$

Note that the existence of high abnormal returns (i.e α_i significantly different from 0), shows a violation of the CAPM.

The Single Index Model was extended by [88] to what is known today as the **Fama and French three factor model**, by introducing a factor capturing the difference between returns of assets with high and low market caps, as well as the difference between stocks with high and low book-to-market equity. The Fama and French model was further extended by [51], who argued that *Momentum* held additional explanatory power, thus leading to the *Cahart four factor model*. On the other hand, [89] extended their model into a five-factor model by including profitability and investment factors.

1.2 Prediction Methods

If the Efficient Market Hypothesis is true, then any attempt to predict future stock movements would be futile, as all available information is perfectly captured in the available prices. While the extent of EMH’s applicability is debatable, virtually everyone would concede that it does not perfectly capture reality. Therefore, it is still worth attempting to predict future stock prices, and it is in fact a very widely studied problem.

Traditionally, the task of modelling stock returns was performed via time series models, which in most cases are parametric linear models. Most notable are Autoregressive Moving Average (ARMA) models, Autoregressive Integrated Moving Average (ARIMA) [172], Simple and Damped Exponential Smoothing Model [42] and [87] respectively, and the additive and multiplicative Holt-Winters Model [235]. All of the above are parametric methods assuming a linear trend, with some of the methods also including seasonal effects. Non-linear extensions of these methods have also been proposed, such as the Autoregressive Conditional Heteroscedastic (ARCH) model [85], the Functional-Coefficient Autoregressive Models [53] and the Smoothed Transition Autoregressive (STAR) [158].

However, with the rapid rise in computing power in recent decades, machine learning methods have become an attractive alternative to traditional time series methods. An early comparison between different machine learning methods on the task of time series forecasting was [5], who concluded that there are significant differences in performance between the different methods. One of the first machine learning algorithms utilized was based on Artificial Neural Networks (ANNs). For example, [121] built an ANN used closing prices of different indices to predict the price movement of

the S&P-500 and DAX indices, showing that a significant improvement in predictive performance is obtained over an AR model. A different ANN model by [86], had a number of different economic and financial variables as inputs, in order to classify stocks based on their expected excess return, resulting in a trading strategy with high returns.

Support Vector Machines (SVM) is another widely used method. A model combining SVM with a wrapper method, was proposed by [141], which was shown to outperform an ANN model. SVM were also used by [209] to predict stock price direction using textual data from breaking financial news. Evolutionary algorithms, such as Genetic algorithms (GA) have also been employed in conjunction with machine learning methods, such as [131] who used a GA approach to feature discretization and ANN weight selection. Other utilized methods include variations of CART [243], Hidden Markov Models [108], ensemble methods [232], random forests [223] and naive Bayes classifier [165].

In more recent years, the more prominent machine learning methods are tailored towards time series data. Recurrent Neural Networks (RNN) are by far the most commonly used class of algorithms, and differ from ANN, in that they introduce skip connections which connect nodes from non-consecutive layers. A popular variant is the Long Short Term Memory (LSTM) algorithm [112] which has been used by [201] to model the equity NIFTY Index in India prices and [193] to model Google stock prices. A very high binary accuracy was achieved by [214] who used LSTM, along with feature engineering techniques, to predict price trends in the Chinese stock market. Another popular variant is the Gated Recursion Unit (GRU), which was used by [204] to predict stocks in the Nepal Stock Exchange, and shown to outperform vanilla RNN and LSTM networks. A hybrid model (combining LSTM and GRU) was used by [114] with promising performance.

In Chapter 3, we introduce OFTER, a forecasting algorithm which combined ideas from k -Nearest Neighbors, Generalized Regression Neural Networks, dimensionality reduction and ARIMA models. This method can provide forecasts for a variety of problems, but is especially efficient with financial forecasting problems which tend to operate in the low signal-to-noise (SNR) regime.

1.2.1 Networks

While financial data have a natural time series format, they can usually also be modeled as a network. A simple definition of a network is that of a mathematical construct composed of a set of objects (often called nodes) which can be connected

with each other (by what we call edges). Edges can be undirected (denoting that two nodes are connected), or directed (denoting that one node points to another). Modelling the financial market using networks can be an extremely useful way of analyzing the existence, stability and evolution of cross-sectional effects. Some examples of networks in the financial sector include an innovation network by using U.S patent data ([2]), the monetary flow between countries ([168]), a network based on the ownership on transnational companies ([228]). We refer the reader to [84] for a review of applications of networks in finance.

Another type of edge could be considered as the interplay between the price movement of different stocks. Consider for example the empirical correlation between the returns of two different instruments $\{f_{i,d}\}_{d=1}^D$, and $\{f_{j,d}\}_{d=1}^D$, over some time horizon. The empirical correlation (or any other measure of similarity) could be used to establish the weight or the existence of an edge (e.g by adding an edge to the network whenever the correlation exceeds some predetermined threshold). Due to the factor models, cross-sectionality between different stock is guaranteed. However, when building such networks, it is common practice to first residualize the returns, so that market factors are neutralized. Indeed, such networks have been widely used by many different studies. For example, [12] built a network using indices of different countries, [77] constructed it for the Greek stock market, while [245] used it for the Shanghai market, by using the partial mutual information to build the stock similarity matrix. Finally, [118] analyzed the dynamic correlation network of the Japanese market.

1.3 Derivative Products

A class of financial instruments that will be the central focus of this thesis is derivatives. The value of derivative products is derived from that of an underlying asset, such as stocks, bonds, commodities, or interest rates. Derivatives can be an alternative venue for speculation on the future price movements of the underlying assets, but they are also used for managing hedging risks or gaining access to otherwise hard-to-trade assets or markets. The history of derivatives can be traced back to Thales of Miletus, who, according to Aristotle, negotiated the right to use olive presses in the spring at a fixed price during the winter, speculating on a bountiful harvest using his expertise in astronomy.

Futures: One of the most common types of derivatives are futures. Futures are standardized and tradable agreements to buy or sell an asset at a predetermined price at a specified time in the future, and typically settled in cash. Futures had primarily composed of commodities, until the first currency and stock index futures were introduced in 1972 and 1982 respectively.

This market have been traditionally understood to play two major roles: price discovery and hedging [216]. The high liquidity, high leveraging potential and ability to lock prices, all make the futures' market an ideal market for hedging. The price discovery function of futures is believed to be a result of the lower transaction costs of the future's market. When the deviation between spot and futures market is large, it is the spot market that tends to adopt, rather than the other way around (see e.g [224]).

The fair value of a futures contract is the theoretical price at which the futures should be trading. This fair value is derived from the spot price of the underlying asset, adjusted for factors such as interest rates, dividends, storage costs, and convenience yield. As we are not dealing with the commodity market, the formula should take the simple form:

$$P_{i,t}^{(Future)}(T) = P_{i,t} \exp(r_t - D_{i,t})(T - t), \quad (1.11)$$

where r_t , $D_{i,t}$, $P_{i,t}$, $P_{i,t}^{(Future)}(T)$ are the interest rate, dividend yield of instrument i , price of instrument i , and price of the corresponding future contract (expiring at time T) at time t .

Options: The other type of derivative that will be considered in this thesis are options, which give the holder the right, but not the obligation, to buy or sell an asset at a predetermined price (called the strike price) before or at the expiration date. The two types of options are call options - which give the holder the right to buy the underlying asset, and put options which give the holder the right to sell the underlying asset.

On the one hand, a trader might buy call (put) options to speculate on the price increase (decline) of the underlying asset, similarly to longing a future contract. However, options generally offer more strategic flexibility. For example, longing both a put and a call option (called a straddle) can be used when speculating a large price movement, while being agnostic in the direction. A diverse list of strategies can be found in [127], which can give rise to various non linear payoff functions.

While future products are only identifiable by their expiration, option contracts also depend on the strike price. Furthermore, the ability to not exercise the option makes the payoff function non linear. Options which can be exercised only at the expiration are classified as **European**. Let $\mathbb{Q}$ be the risk neutral measure, the current value of a European option at time t , with strike price K and expiration at time τ is calculated as

$$e^{-r\tau} \mathbb{E}^{\mathbb{Q}} [\max(S_{\tau} - K, 0)] . \quad (1.12)$$

Options which can be exercised at any time before expiration are classified as **American**. In this case, the supremum is taken over all possible exercise points

$$\sup_{t \in [0, \tau]} e^{-rt} \mathbb{E}^{\mathbb{Q}} [\max(S_t - K, 0)] . \quad (1.13)$$

In either the case, the value depends on the distribution of the price process $(S_t : t \in [0, \tau])$. In chapters 2 and 4, we will consider two widely used models; the Black Scholes and Heston models.

1.4 Contributions

In this dissertation, we are concerned with the forecasting of the three major quantities of interest (price return, volatility, trading volume) using techniques from the literature of time-series analysis and machine learning. Forecasting the direction of price movements is one of the most well-studied problems in finance, as it can directly translate into profits, while the other two quantities can also be of crucial importance.

Prediction of volume has gained more attention in recent decades, with many attempts at forecasting both daily and intraday trading volumes. For instance, [203] and [159] used an ARMA and Kalman filter methodology, respectively. Accurate forecasts can help sell-side practitioners adjust their inventory and spread accordingly. Buy-side analysts also make use of such forecasts in conjunction with return forecasts by timing entry or exit times, or controlling the market impact of large trades.

All of these aforementioned benefits of volume forecasting are also valid reasons for volatility forecasting. Furthermore, volatility is a crucial factor for the pricing of derivative products, as well as in portfolio optimization tasks. Importantly, speculation on future levels of volatility can also be traded directly (see [127] for multiple such examples).

While a lot of research regarding the forecasting of the three main quantities has come out in recent years, there is still much room for further research. We have chosen

four such topics, tackling certain gaps in the literature, providing methodology and insights for practitioners. While these four topics lie in different subfields of quantitative research, they are connected by certain themes and principles that permeate this work.

- **Focus on data-driven methods:** As discussed earlier, there has been a growing interest in deep learning and other machine learning methods, with these models outperforming traditional models in the majority of cases [211]. There is, however, still a lot of space for further such research, with the projects included in this thesis following this pattern.
- **Avoiding overreliance on deep learning methods:** While deep learning methods are undoubtedly the most popular class of machine learning models, they are not always the optimal methodology [163]. In daily financial forecasting problems, where the number of data points is small, it is usually not sufficient to effectively train deep learning models, especially when the dynamical nature of financial markets makes older data unreliable. While this paper makes use of machine learning, it leans away from complex and computationally expensive models, focusing on simpler deep learning models and other machine learning methods.
- **Usage of economically significant metrics:** Traditional loss functions and evaluation metrics used in generic forecasting problems are sometimes inadequate to act as objective functions. In this thesis, we incorporate the concepts of the Profit and Loss and the Sharpe Ratio as objective functions and as evaluation metrics in a number of different cases when the forecasting problem regards returns, similar to [254, 249].

Chapters 2-5 of this thesis are reports of self-sufficient projects, centered on forecasting problems of quantitative finance, loosely connected via the themes discussed. Each chapter is a slight adaptation of a corresponding paper that is either accepted for publication or under review. In the table below we summarize some commonalities and differences between the four individual projects.

In the subsequent subsections, we outline the specific contributions made in Chapters 2-5. In Table 1.1 we compare and contrast the contribution, methodology and forecasting problem presented in each project.

Title	Option Volume Imbalance as a Predictor for Equity Returns	OFTER: An Online Pipeline for Time Series Forecasting	Options-driven Volatility Forecasting	A GCN-LSTM Approach for ES-mini and VX Futures Forecasting
Chapter	2	3	4	5
Forecasting Task	Equity Returns	Equity Returns and Volumes	Equity Realized Volatility	Futures' Returns, Realized Volatility, and Volumes
Forecast Horizon Method	Daily PNL Regression	Hourly OFTER	Daily LASSO	Hourly GCN-LSTM
Data Set (of extracted features)	Option Volume Data Set (NOTO + PHOTO)	NYSE minutely equity price data	NYSE intraday equity price data + Implied Volatility Surface Data set (OptionMetrics)	Futures minutely price dataset (Boltzmann Research)
Novelty	Novel option volume feature due to a novel segmentation by market participant	Novel forecasting pipeline tailored to mid-sized financial tasks	Novel features extracted using parametric models	Novel framework for analyzing futures of different expiration

Table 1.1: Overview of research projects showcased in this dissertation.

1.4.1 Option Volume Imbalance as a Predictor for Equity Returns

We investigate the use of the normalized imbalance between option volumes corresponding to positive and negative market views, as a predictor for directional price movements in the spot market. Via a nonlinear analysis, and using a decomposition of aggregated volumes into five distinct market participant classes, we find strong signs of predictability of excess market overnight returns. The strongest signals come from Market-Maker volumes. Among other findings, we demonstrate that most of the predictability stems from high-implied-volatility option contracts, and that the informational content of put option volumes is greater than that of call options.

Publication Status: This paper has been accepted by the Journal of Market Microstructure and Liquidity.

1.4.2 OFTER: An Online Pipeline for Time Series Forecasting

We introduce OFTER, a time series forecasting pipeline tailored for financial problems centered on mid-sized multivariate time series. OFTER utilizes the non-parametric models of k-nearest neighbors and Generalized Regression Neural Networks, integrated with a dimensionality reduction component. To circumvent the curse of dimensionality, we employ a weighted norm based on a modified version of the maximal correlation coefficient. The pipeline we introduce is specifically designed for online tasks, has an interpretable output, and is able to outperform several state-of-the-art baselines. The computational efficacy of the algorithm, its online nature, and its

ability to operate in low signal-to-noise regimes, render OFTER an ideal approach for financial multivariate time series problems, such as daily equity forecasting. Our work demonstrates that while deep learning models hold significant promise for time series forecasting, traditional methods carefully integrating mainstream tools remain very competitive alternatives with the added benefits of scalability and interpretability.

Publication Status: This paper is under review by the Quantitative Finance Journal.

1.4.3 Options-driven Volatility Forecasting

We augment the Heterogeneous Autoregressive Regression model for forecasting realized volatility, using various measurements for the daily, weekly, and monthly volatilities, in addition to other predictive features. The main focus is on novel methods for extracting volatility estimators using option price data. Firstly, we provide a dimensionality reduction method for implied volatility surfaces built under the Black–Scholes model, whereby we combine simple row-wise and column-wise decompositions of the implied volatility surface with principal component analysis. Secondly, we provide a method for extracting the implied volatility under a Heston model. This is achieved by a calibration of the Heston model while assuming that some of the model parameters remain constant. We demonstrate that these augmentations result in improved daily forecasts for realized volatility in a selection of different stocks. These volatility forecasts, can be also be utilized to further increase predictive performance for the realized volatility of other instruments, and can be combined to provide accurate forecasts for VIX.

Publication Status: This paper is under review by the Quantitative Finance Journal.

1.4.4 A GCN-LSTM Approach for ES-mini and VX Futures Forecasting

We propose a novel data-driven network framework for forecasting problems related to E-mini S&P 500 and CBOE Volatility Index futures, in which products with different expirations act as distinct nodes. We provide visual demonstrations of the correlation structures of these products in terms of their return, realized volatility,

	Chapter 2	Chapter 3	Chapter 4	Chapter 5
Price at opening (day d)	$P_{i,d}^{\text{open}}$	-	O_d	-
Price at closing (day d)	$P_{i,d}^{\text{close}}$	$P_{i,d}$	C_d	-
Highest price (day d)	-	-	H_d	-
Lowest price (day d)	-	-	L_d	-
Price at time t	-	-	S_t	$P_t^{(M)}/P_t^{(A)}/P_t^{(B)}$
Daily Close to Close Return	$f_{i,d}^{\text{CL_tmCL}}$	$R_{i,d}^1$	-	-
Daily Open to Close Return	$f_{i,d}^{\text{tmOP_tmCL}}$	-	-	-
Daily Close to Open Return	$f_{i,d}^{\text{CL_tmOP}}$	-	-	-
K-minute Return	-	-	-	$R_t^{K\text{-minute}}$
Daily Realized Volatility	-	-	$\text{RV}_{i,d}$	-
K-minute Realized Volatility	-	-	-	$\text{RV}_t^{K\text{-minute}}$
Trading Volume	$V_{i,j,d,m}^{\text{Up}} V_{i,j,d,m}^{\text{Down}}$	$V_{i,t}$	Volume_d	$V_t^{K\text{-minute}}$
Daily P&L	$P \& \text{amp; } L_d$	$P \& \text{amp; } \text{amp; } \text{amp; } L_d^1$	-	$P \& \text{amp; } \text{amp; } L_d$

Table 1.2: Notation used in the chapters of this dissertation.

and trading volume. The resulting networks offer insights into the contemporaneous movements across the different products, illustrating how inherently connected movements of the future products belonging to these two classes are. These networks are further utilized by a multi-channel Graph Convolutional Network to enhance the predictive power of a Long Short-Term Memory network, allowing for the propagation of forecasts of highly correlated quantities, combining the temporal and spatial aspects of the term structure.

Publication Status: This paper is under review by the Quantitative Finance Journal.

1.5 Outline and Note on Notation:

Note that as the following chapters are either published or submitted articles, they are self contained, and use notation tailored to their requirements. This is especially true for Chapter 4, which uses notation suited for stochastic calculus, as we are dealing with parametric volatility models. In Table 1.2, we differentiate between the notation used in the different chapters.

Chapter 2

Option Volume Imbalance as a Predictor for Equity Returns

2.1 Introduction

Since the opening of the CBOE in 1973, derivative markets have been a prominent presence in the financial system and have grown rapidly since. According to BIS estimates [19], the total global nominal amount outstanding for derivatives for the second half of 2020 amounts to USD 582 trn, of which 7 trn are equity-linked options¹. The latter is accounted for by multiple option exchanges across the world, including the CBOE, the CME group, Nasdaq, and Eurex. Equity options are appealing to traders due to the higher leverage they offer as well as the absence of restrictions such as the uptick rule. The limited downside risk of a long put position as opposed to a short futures position in the stock and the hedging opportunities are also attractive to investors.

Naturally, derivative markets are strongly intertwined with spot markets. It has been argued that, due to informed trading, option markets lead the spot markets (e.g. [33]). In particular, an interesting line of research is concerned with the information content of signed option volumes. A trader who (either from private information or by using predictive models) expects a spot-price rise can take a long call position or a short put position, while a trader who takes a short call position or long put position may expect a price fall instead. Therefore, aggregating the signed volumes into these two categories of positive and negative signals yields a metric of the discrepancy in trader anticipation. Our main question is whether this discrepancy can act as a predictor for the direction of future price movements in the spot market.

¹By comparison the global total GDP was estimated by the World Bank to be USD 88 trn [236].

To this end, we construct positive and negative “signals” (defined below) from the order flows and use their normalized difference as a predictor for future stock returns. We refer to this metric as the *Option Volume Imbalance (OVI)*. What differentiates this paper from other similar works is the use of two data sets, from the NOM and PHLX option exchanges, which disaggregate the daily option volume into 10-minute buckets, distinguishing between five different market participants and between call and put options. Similar research usually uses either customer data or generally non-market maker data, without distinguishing between classes of market participants. We place a special emphasis on the decomposition between different classes of traders (or market participant classes), and especially that of Market Makers. While this may create an instance of double counting orders for both sides of a trade, it provides insight into how the same transactions can be used from different angles to provide trading signals with varying performance. Furthermore, this shows that only certain trading classes can produce an OVI positively correlated with future equity returns. If, for example, Customers are at least partially successful in their speculative ability, this should translate into a financial signal that results in positive returns. On the other hand, Market Makers who profit in other ways (e.g spreads) would not be expected to provide a profitable signal but rather one with negative returns, reflecting the speculative ability of all those who trade with them.

We also empirically assess how the predictive power of OVI varies based on different values of the options’ Greeks and other characteristics. We compare the two different exchanges and provide a qualitative description of certain intraday patterns we observe. Lastly, we investigate the cross-sectional effects of this metric by modeling the interactions across the various assets using a directed network.

Throughout the paper, we use the cumulative Profit and Loss (P&L) and Sharpe Ratio as measures of predictability in future returns. However, as we consider multiple types of OVI (e.g different market participants), we also wish to consider their total predictive power. This motivates us to formulate a novel regression approach that directly maximizes the cumulative P&L, which we refer to as *P&L regression*. Unlike other research of this type, which usually tests the relationship between covariates and future returns by considering t or R^2 values, our approach is not limited to linear relationships and can deal effectively with correlated features. We summarize our main contributions as follows.

Summary of main contributions The main contribution of our work is an analysis of the Option Volume Imbalance (OVI), and in particular, its relationship with the future price of the underlying equity/ETF, as follows.

- (1) We provide evidence that the OVI metric exhibits predictive power on directional overnight price movements for the underlyings.
- (2) We show, by analyzing the OVIs of different market participants, that they bring different information to the market; this segmentation is novel in the literature.
- (3) We analyze the effect of various option features on the capacity of the OVI to predict equity returns.
- (4) Finally, we provide evidence of cross-impact within the universe of underlying equity/ETF assets.

As a secondary theme, we also give a short, high-level overview of the two markets, showing how most of the activity comes from Market Makers and Customers. We also provide an approximation of how option volume flows between the different market participants.

Paper outline. The remainder of the paper is organized as follows. In Section 2.2 we give an overview of related literature and define the Option Volume Imbalance metric (OVI). In Section 2.3, we describe the data sets used in our analysis. In Section 2.4, we discuss the evaluation methods used subsequently. In Section 2.5, using volumes from different market participants, we show the predictive power of the OVI for future returns in the spot market. In Section 2.6, we see the effects of various option features, such as the Greeks, on the predictive power of the OVI metric. In Section 2.7, we extend our analysis by comparing the two data sets (from different exchanges), and we consider results for longer time horizons as well as some intraday variants of the OVI-based strategies. In Section 2.7.5, we check for cross-impact effects, using a network model. Lastly, we summarize our conclusions and detail future research directions in Section 2.8. Additional numerical results are deferred to Appendix A.4.

2.2 Motivation and the Option Volume Imbalance

2.2.1 Related Literature

Traditionally, derivative markets have been used for hedging purposes, and have been shown to play a role in reducing the risk of portfolios. However, the high leverage offered by options naturally led to the suggestion that they are also an avenue for informed trading ([33] and [103]). The theoretical model of [83] showed that under

certain reasonable conditions, informed traders should choose to capitalize on their private information by trading in both the stock and option markets. Under this model, informed trading is more prevalent when the liquidity and leverage of the option markets are sufficiently high. Furthermore, under [31]’s model, the introduction of options reduces the informational efficiency of the market, making it harder for market makers to distinguish the information content of trades as the option market becomes a more attractive avenue for informed traders.

The early empirical investigation by [161] focused on the opportunity signals appearing whenever implied option prices from the Black-Scholes (BS) model differed significantly from observed prices, suggesting that there was information in option prices not reflected in the stock market. This was extended by [30] to intraday prices, with a pseudo-American BS model for the implied options, supporting and extending the results of [161]. They used a number of different holding periods for equities, with most intraday strategies resulting in losses due to transaction costs, even though they report profits in the overnight holding case.

Under the sequential information arrival model of [69], volumes can be used as a proxy for the rate of information arrival. Based on that, [11] showed that call volumes lead stock volumes, which indicates the presence of informed traders in the options market. Challenging these results, [220] showed empirically that the stock market leads the option market by at least 15 minutes, rather than vice versa. It was observed by [83] that while price changes in the stock market lead the option volumes (as articulated earlier by [220]), the results are quite different if one aggregates volumes into positive (call buys and put sells) and negative (call sells and put buys) flows. They rejected the hypothesis that these aggregated volumes hold no information about the spot market. This led to research on features based on option volumes and volume imbalances (the difference between volumes corresponding to positive and negative flows).

One of the most important studies focusing on the decomposition to positive and negative flows was [238], which showed that the signed option volume (which they referred to as *demand for volatility*) was correlated with the realized volatility and that market makers responded to imbalances by raising the spreads in order to shield themselves from informed traders. According to the findings of [206], signed option volumes have some predictive power for short-term returns, which reverse at longer timescales, in contrast to the predictive power they found in futures. Later, [187] used put/call ratios from the opening of new buyer-initiated options and showed their predictive power for future stock returns. They concluded that the source of

predictability was primarily informed trading by differentiating between publicly and non-publicly available data. Furthermore, they showed that deep out-of-the-money options had the highest predictive power, as did data that came from customers who traded using full-service brokerage houses; in contrast, firm proprietary trades had no predictive power.

Another option volume flow metric was defined by [115], who used delta-adjusted volumes and showed its significant predictive power for stock returns. The sign of the resulting coefficients did not reverse for longer horizons, as also observed earlier by [187]. Furthermore, they found that the predictability of future returns is stronger for assets with a higher ratio of informed traders, low analyst coverage, large bid-ask spreads, small market capitalization, and low institutional ownership.

Another promising metric proposed was the Option to Stock Volume Ratio (O/S), first constructed by [199], who concluded that the cross-sectional and time-series variation of this metric reflected the presence of informed trading. The predictive power of this metric for future returns was empirically verified by [124] and was shown to be stronger when short-sale constraints were higher. Also, [98] compared the different signed option volumes by using the O/S ratio and concluded that there was no significant difference in predictive power between long and short positions, thereby speculating that the source of predictability was the embedded leverage common to the two positions, while the short-sale costs played no significant role. Later, [148] decomposed the O/S metric into eight distinct ratios, differentiating between open and close, buy and sell, and calls and puts. Noticing how the informational content in both long and short positions was similar, they concluded that the most important channel of predictability was the embedded leverage in the options market.

A more recent study was [52], regarding the interplay between returns and net volume (buyer-initiated minus seller-initiated) across stock, call, and put markets. They found that it was the stock net volume that led the spot and option quote-price revisions, while the net option volumes had no effect. However, both net flows seemed to be negatively autocorrelated, perhaps indicating the effect of inventory control.

The hypothesis that informed trading is the source of predictability derived from the option market is strengthened by studies focusing on eventful dates. For example, [9] found that around announcement days, both the profitability and the effective bid-ask spread of options increased. [149] provided conclusive evidence that, prior to any type of corporate analyst event, both the skew and spread of implied volatility become significant predictors, especially in liquid markets. It has been conjectured that this is due to analyst tipping (analysts give traders tips on upcoming information). For

example, [48] investigated the effect of option volumes prior to takeover and merger announcements. They found that the lagged call option imbalance (the difference between buyer and seller-initiated volumes) is not a significant predictor for next-day returns, but that this was not the case for pre-announcement periods. They also showed that short-term out-of-the-money options experienced the largest increase in trading volume during these periods.

Contrary to previous research, [101] argued that the main driving force behind the predictive power of option prices stems from price pressure and not from informed trading, as the predictive power was not lowered when considering option contracts with zero volumes for a specific day.

2.2.2 Option Volume Imbalance

Denote by N_d the number of assets whose options were traded on financial day d , and by $C_{i,d}$ the number of different option contracts corresponding to the i^{th} asset. Given day d and asset i , let $T_{i,j,d,m}$ and $V_{i,j,d,m}$ denote the number of trades and volume for option contract j by investors of market participant class m . "Number of trades" refers to the total number of transactions (regardless of their size), while volume refers to the total number of option contracts traded. Also let $P_{i,j,d}^{option}$ denote the average of the opening and closing price of option contract j for day d .

As [83] observed, while the option volumes do not reliably predict movements in the stock market, this changes when one considers the volumes of positive and negative signals separately. Positive signals include call buys and put sells, while negative signals include call sells and put buys. If a speculator (with or without private information) predicts a positive movement in the price of the underlying, they can opt to enter into either or both of a long call position and a short put position. Conversely, if predicting a downward movement, they can take either or both of a long put position and a short call position². Thus, we define

$$V_{i,j,d,m}^{Up,\mathcal{F}} = V_{i,j,d,m}^{Call_Buy,\mathcal{F}} + V_{i,j,d,m}^{Put_Sell,\mathcal{F}} \quad \text{and} \quad V_{i,j,d,m}^{Down,\mathcal{F}} = V_{i,j,d,m}^{Call_Sell,\mathcal{F}} + V_{i,j,d,m}^{Put_Buy,\mathcal{F}}, \quad (2.1)$$

to be the volumes corresponding to speculation for upward and downward movement respectively. We similarly define

$$T_{i,j,d,m}^{Up,\mathcal{F}} = T_{i,j,d,m}^{Call_Buy,\mathcal{F}} + T_{i,j,d,m}^{Put_Sell,\mathcal{F}} \quad \text{and} \quad T_{i,j,d,m}^{Down,\mathcal{F}} = T_{i,j,d,m}^{Call_Sell,\mathcal{F}} + T_{i,j,d,m}^{Put_Buy,\mathcal{F}}, \quad (2.2)$$

²This excludes traders who enter the options market to hedge an existing stock position or those who trade option straddles and other more complicated strategies. The latter poses no issue as the two types of volumes cancel each other out, while, unfortunately, we are unable to identify traders purchasing options for hedging purposes.

to be the number of trades corresponding to speculation for upward and downward movement respectively. Note that we use $\mathcal{F}$ to encode different filters we apply to the transaction data before running different experiments (deliberately left general).

Therefore, using the volume of Call-Buy/Sell and Put-Sell/Buy transactions as positive and negative directional signals respectively, and normalizing by the total volume, we define the Option Volume Imbalance (OVI) for asset i and market participant class m (for day d) as

$$\text{OVI}_{i,d,m}^{[\mathcal{F}]} = \frac{\sum_{j=1}^{C_{i,d}} (X_{i,j,d,m}^{\text{Up},\mathcal{F}} - X_{i,j,d,m}^{\text{Down},\mathcal{F}})}{\sum_{j=1}^{C_{i,d}} (X_{i,j,d,m}^{\text{Up},\mathcal{F}} + X_{i,j,d,m}^{\text{Down},\mathcal{F}})}, \quad (2.3)$$

where X denotes a generic flow variable³. We mostly use the volume-based OVI, which is defined by setting $X = V$, i.e $X_{i,j,d,m}^{\text{Up},\mathcal{F}} = V_{i,j,d,m}^{\text{Up},\mathcal{F}}$ and $X_{i,j,d,m}^{\text{Down},\mathcal{F}} = V_{i,j,d,m}^{\text{Down},\mathcal{F}}$. The normalization ensures that OVI lies in the interval $[-1, 1]$, so that the OVI of different assets can still be directly compared, independent of the asset's liquidity or market cap. Note that due to the structure of the data set, each transaction is double-counted, from the perspective of both parties. For example, if a Customer buys a call option from a Market Maker, with the intention of creating a long position, it is counted both as a Customer-Buy and a Market Maker-Sell transaction. More details regarding the data set are given in Section 2.3.

One potential shortcoming of volume-based OVI is that more weight could be given to option contracts with lower prices (e.g deep OTM options) as they are sometimes transacted in larger volumes. To ameliorate this concern, we can define alternative OVI features by using the total number of trades (which we refer to as *trade-based* OVI) by setting $X = T$, i.e $X_{i,j,d,m}^{\text{Up},\mathcal{F}} = T_{i,j,d,m}^{\text{Up},\mathcal{F}}$ and $X_{i,j,d,m}^{\text{Down},\mathcal{F}} = T_{i,j,d,m}^{\text{Down},\mathcal{F}}$. We also define the *nominal volume-based* OVI, by weighting the volume by the option price, i.e set $X_{i,j,d,m}^{\text{Up},\mathcal{F}} = P_{i,j,d}^{\text{Option}} V_{i,j,d,m}^{\text{Up},\mathcal{F}}$ and $X_{i,j,d,m}^{\text{Down},\mathcal{F}} = P_{i,j,d}^{\text{Option}} V_{i,j,d,m}^{\text{Down},\mathcal{F}}$. Nominal-volume based OVI explicitly gives more weight to higher cost trades, thus preventing low-cost options from having a disproportionate influence⁴.

Our OVI-based approach is similar in spirit to but slightly different from the put-to-call volume ratio that was used by [187]; those authors restricted their analysis to Open-Buy transactions (see Appendix A.1 for definitions) and showed how the

³Note that for this definition we define $0/0 = 0$, i.e when the total volume is equal to 0, the OVI is also equal to 0.

⁴Note that the choice of sign in (2.3) is arbitrary; depending on the market participant, the OVI can be either positively or negatively correlated with future spot returns. The choice of signs was made in accordance with the intuition of positive and negative news, as characterizations of Call-Buys, Put-Sells and Call-Sells, Put-Buys respectively.

put-to-call ratio is a negative predictor for the price movement of the underlying. Since our data set splits transactions based on Market Participant, we do not need to restrict to Open-Buy transactions. Instead, by using (2.3), all of the available types of transactions are used to construct a metric for each Market Participant Class.

Importantly, unlike other similar studies in this area, we do not make a linearity assumption between OVI and future returns. In contrast, our analysis uses the sign of the OVI as a directional predictor for subsequent price movements, by making a hypothetical transaction in the corresponding direction. If a significant profit (or loss) is made from this transaction, then we conclude that the OVI contains significant information about future spot market returns.

2.3 NASDAQ Data Sets

2.3.1 Data Sets Description

We consider two data sets, from the Nasdaq Options Market (NOTO) and Nasdaq PHLX (PHOTO) exchanges. The data sets provide information about the aggregate option volumes for the period 02 Jan 2015–31 Dec 2019. For each day in this range, we are given a time series of intraday updates from the exchange, disseminated at 10-minute intervals, detailing the total cumulative volume of options transacted up to that intraday time point. As the market opens at 09:30 ET, the first report is disseminated at 09:40 ET, and the last report is disseminated at 16:00 ET (a total of 39 intraday time points). We are also given a daily summary report containing the total daily option volume for each contract. In the daily report, we are also given its low, high, opening, and closing price, its open interest, and total electronic volume, as well as the total volume for the specific exchange. Aggregate volumes are further broken down into five different market participant classes (hereafter MPCs) which are: Firm (Proprietary Trades), Brokers, Market Makers, Customers, and Professional Customers. Definitions for these are given in Appendix A.1.

Furthermore, for each MPC other than Market Makers, volumes are further decomposed into four parts based on the type of transaction and intention of the agent: Opening Buy, Opening Sell, Closing Buy and Closing Sell (definitions are given in Appendix A.1).

2.3.2 Descriptive Analysis of the Market

In the rest of this section, we show some preliminary analysis of the two data sets, including an analysis of the interaction between different MPCs. The purpose of

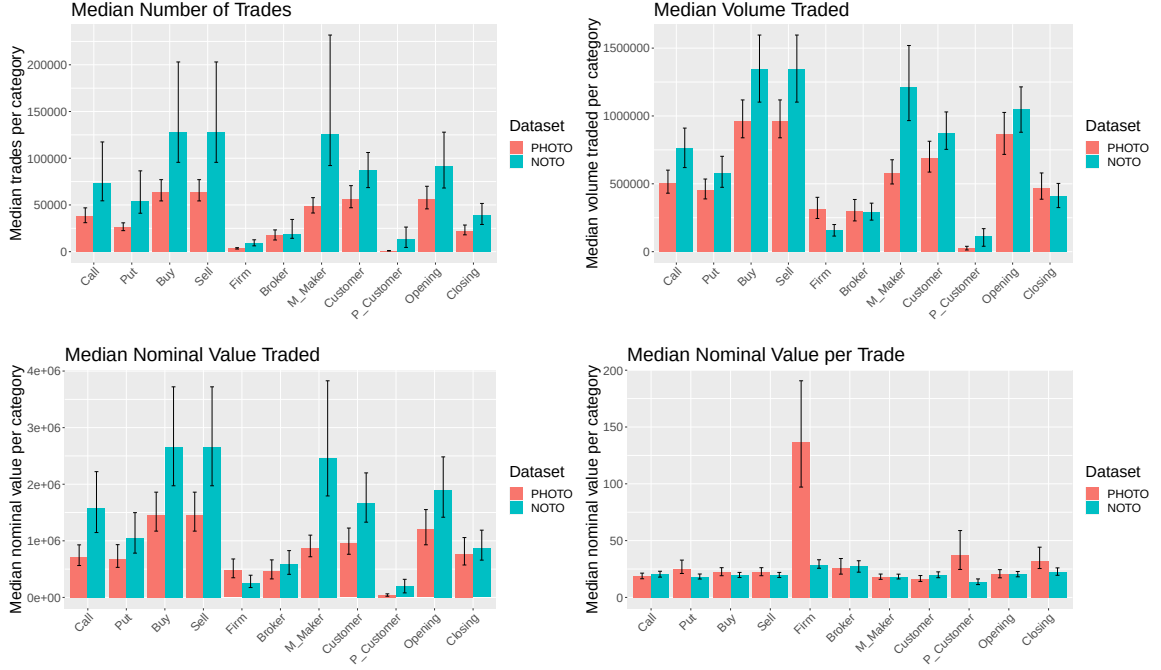


Figure 2.1: Breakdown of the two data sets into their various constituents: The plots show the median daily number of trades (top-left), median daily volume traded (top-right) and median nominal value traded (bottom-left) and the average *nominal size per trade* (bottom-right) which is the ratio between the values of the first and third quantities. The whiskers indicate the Inter-quartile Range across days for each constituent.

this section is twofold. First, the analysis in itself is novel and some interesting observations are noted. Second, these observations can aid with the interpretation of the results of the subsequent sections.

In Figure 2.1, we show summary statistics (on a daily basis) for the trades in the two data sets, disaggregated by trade type and MPC as defined in appendix A.1. Although NOTO has a higher average trade count across MPCs, the aggregates of the two data sets are qualitatively similar. In particular, call options have a higher median volume than put options for both data sets, and the most prevalent MPCs are Market Makers and Customers. Firms and Professional Customers have a very small median volume. However, when it comes to the nominal value per trade for the PHOTO data set, Firms have by far the highest ratio compared with other MPCs. Overall, we expect the OVI of Market Makers and Customers to be the most informative, as the OVI of the other MPCs is constructed from fewer transactions and thus more likely to be noisy (this is confirmed in Section 2.5, where we observe higher performance for Market Maker and Customers OVI).

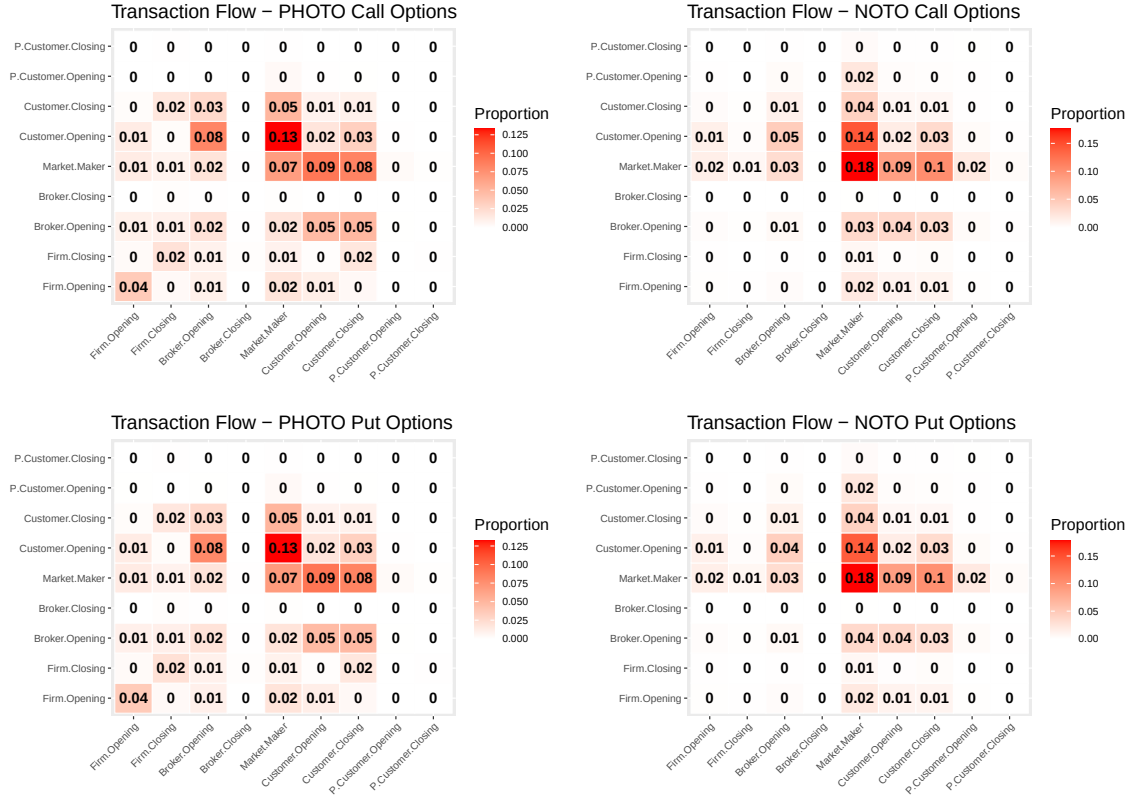


Figure 2.2: An approximation of the volume flow between MPCs for the two data sets (see (A.2) for the approximation). The rows/columns correspond to the Buyer/Seller side of transactions respectively. Cell values show the median of the approximated flow proportion between the two participants, to two decimal places. The red highlighting simply denotes higher values of a particular value.

To complement the overview of Figure 2.1, we also show the flow between different MPCs, i.e how much they trade with each other. Denote by $\widetilde{\text{NF}}_{m_1, m_2}$, the average trade flow from MPC m_1 (as a seller) to MPC m_2 (as a buyer), which represents the median percentage of trades that correspond to these MPCs. The full transaction details are not available, so we can approximate them by inspecting the 10-minute intraday aggregate volume reports, and then interpreting and matching the flow between different MPCs. For details, see Appendix A.2. The heatmap of the median flow for the two data sets is shown in Figure 2.2, where we differentiate between call and put options. Once again, the two data sets display many similarities in their flow structure. Firstly, in both cases, the call and put flow structures are very similar. Secondly, whenever a flow from MPC m_1 to m_2 ($\widetilde{\text{NF}}_{m_1, m_2}$) is relatively high, the reverse flow ($\widetilde{\text{NF}}_{m_2, m_1}$) also tends to be relatively high. In addition, a very large proportion of the flow involves either a Customer and/or a Market Maker in the

transaction. In both cases, the flow between Brokers and Customers is much higher than the flow between Brokers and Market Makers, but in the case of PHOTO the difference is much greater.

The proportion of trades involving the Firms MPC is much higher for PHOTO in comparison to NOTO. The opposite is true for Market Makers, where a higher proportion is for NOTO and they share a significant proportion of the flow with Professional Customers Opening.

In both data sets, Firms and Professional Customers trade mostly with Market Makers, while for Brokers their flow with Customers is larger than for Market Makers.

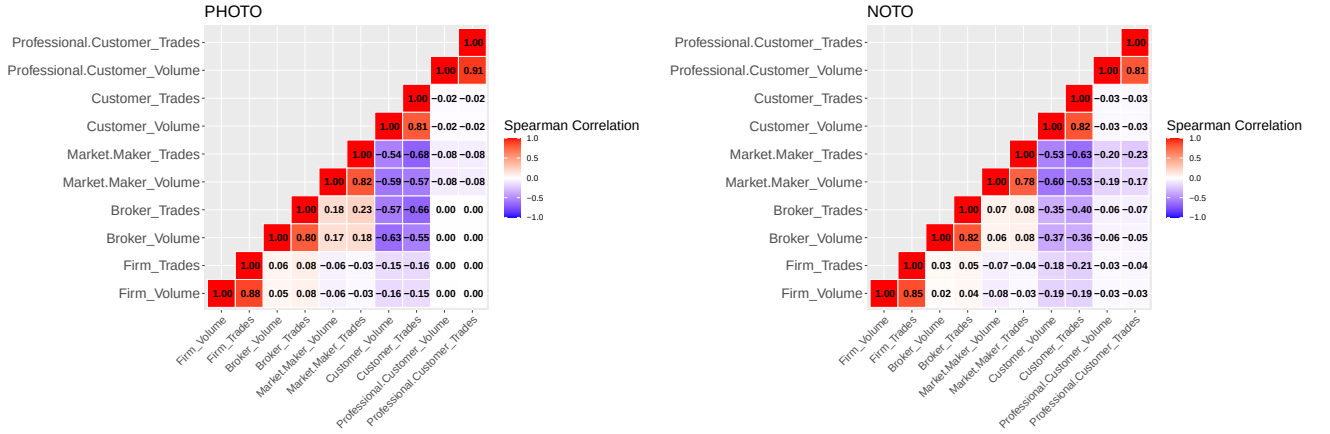


Figure 2.3: OVI Correlation Map: Shows the median Spearman correlation between OVI measures across days.

2.3.3 Correlation of Volume and OVI

The correlation between PHOTO and NOTO option volume ($\sum_m \sum_j V_{i,j,d,m}^{\text{PHOTO}}$ and $\sum_m \sum_j V_{i,j,d,m}^{\text{NOTO}}$) ranges from 0.00 to 0.93, for different assets. The correlation between the total volume for different MPCs ($\sum_i \sum_j V_{i,j,d,m}$ for different m) is also quite high (as expected), in the range 0.8–1.0. More interestingly, in Figure 2.3 we plot the median (across days) correlations between the OVI features built using the data from the five MPCs⁵. Unsurprisingly (as trades are double-counted), there is a strong correlation between the volume- and trade-based OVI when they correspond

⁵Note that the correlation map is very similar across days, with each entry having a standard deviation of 0.05 or less, across days. Also note that for this analysis, the Spearman correlation was chosen for its simplicity and ability to identify nonlinear relationships (the Pearson correlations were, however, very similar).

to the same MPC. More importantly, for both data sets, Customers have a high negative correlation of around -0.6 , with Market Makers and Brokers, and a low negative correlation of around -0.20 with Firms. We also observe a weak correlation of around 0.2 between Market Makers and Brokers for the PHOTO data set, as well as a weak correlation of around -0.2 between Market Makers and Professional Customers for the NOTO data set. The rest of the correlations are below 0.1 in magnitude. Therefore, we should keep in mind the strong multicollinearity present in the OVI features (which later on brings up the need for a regularization term in the P&L regression Section 2.4.3).

2.3.4 Principal Component Analysis

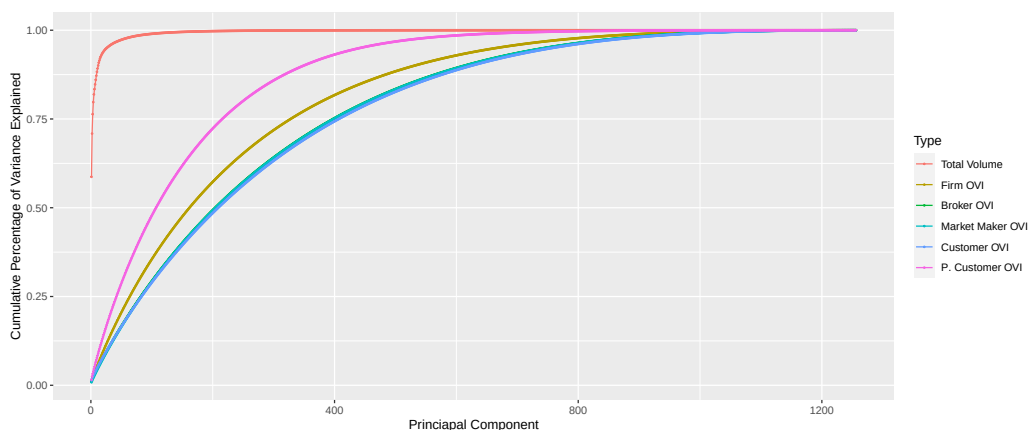


Figure 2.4: Volume and OVI Principal Components: This plot shows the cumulative percentage of variance captured by principal components for each of the different OVI or Volume.

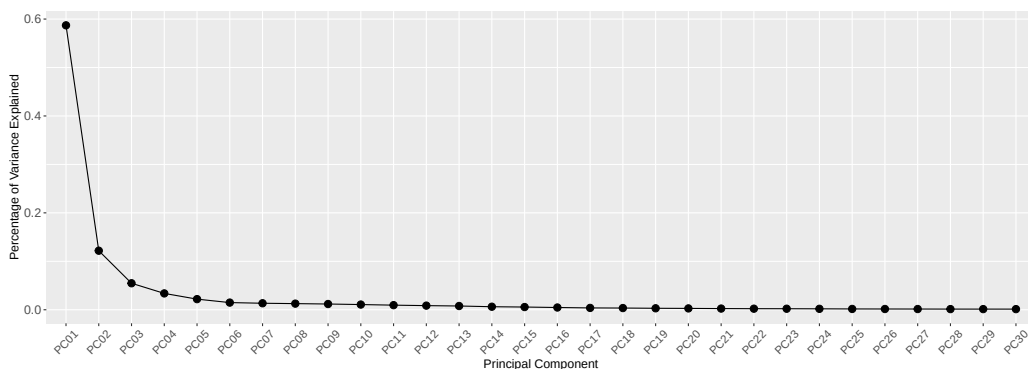


Figure 2.5: Volume Scree Plot: This plot shows the percentage of variance captured by the first few principal components for total PHOTO option volume.

We carry out PCA to investigate the intrinsic dimensionality of the option volumes and OVI observed for the PHOTO data set. In Figure 2.4, we plot the cumulative percentage of variance explained for volumes and each type of OVI. For volumes, the first principal component explains 59% of the variance, with the 12 first components explaining more than 90%. In Figure 2.5, we show a scree plot for the first few principal components corresponding to PHOTO option volumes. For all OVIs, more than 100 principal components are required to explain 50% of the variance. MPCs with lower daily volumes require fewer principal components, with Professional Customers requiring just 107 components, while Customers require 211 components. Thus, using option volumes to construct OVI significantly increases the intrinsic dimensionality of the data set. This could be an indication that OVI is a better feature for forecasting asset-specific returns when compared to option volumes (a data set with much lower intrinsic dimensionality).

Overall, we see that PHOTO and NOTO are similar in terms of the trade breakdown, while the breakdowns between call and put options are almost identical. We see that the major MPCs are Customers and Market Makers, while all the OVIs are highly interconnected. On the other hand, the higher intrinsic dimensionality observed from the OVI is further motivation for its use instead of the raw option volumes.

2.4 Evaluation Strategy

2.4.1 Profitability and Predictability

Evaluating the predictive power of features on returns is often dealt with using the traditional statistical treatment; that is, via metrics such as the Mean Squared Error, or the t and R^2 values, resulting from the linear relationship between the forecast and the target variable. However, the problem with traditional evaluation metrics often employed is that they do not take into account the peculiarities of the specific problem. In particular, correctly predicting the price movement direction but failing to accurately predict the magnitude should not be penalized as heavily as guessing the wrong sign, or perhaps should not be penalized at all. More importantly, given the very high error-to-signal ratio in this problem, the R^2 or t -values obtained are often too small to distinguish and compare, thus rendering these metrics unsuitable. A better alternative is to consider instead metrics with economic interpretation. In particular, we consider a hypothetical strategy in order to assess the predictability

of the OVI features with respect to future returns. As trading costs are trader-dependent, we choose to omit taking them into account. We also ignore any liquidity constraints. This ensures that any obtained predictability does not violate the efficient market hypothesis and is instead a realistic indication of the connection between the two quantities.

The best alternative is to instead consider metrics with economic interpretation. In particular, we consider a hypothetical strategy in order to assess the predictability of the OVI features with respect to future returns. As trading costs are trader-dependent, we choose to omit taking them into account. We also ignore any liquidity constraints. This ensures that any obtained predictability does not violate the efficient market hypothesis and is instead a realistic indication of the connection between the two quantities.

2.4.2 Types of Return and Evaluation Metrics

Let $P_{i,d}^{\text{open}}$ and $P_{i,d}^{\text{close}}$ denote the opening and closing price for asset i at day d . We define the following return types for asset i :

$$\text{Daily Close to Close Returns (CL_tmCL): } f_{i,d}^{\text{CL_tmCL}} = \frac{P_{i,d+1}^{\text{close}} - P_{i,d}^{\text{close}}}{P_{i,d}^{\text{close}}}, \quad (2.4)$$

$$\text{Daily Open to Close Returns (tmOP_tmCL): } f_{i,d}^{\text{tmOP_tmCL}} = \frac{P_{i,d+1}^{\text{close}} - P_{i,d+1}^{\text{open}}}{P_{i,d+1}^{\text{open}}}, \quad (2.5)$$

$$\text{Daily Overnight Returns (CL_tmOP): } f_{i,d}^{\text{CL_tmOP}} = \frac{P_{i,d+1}^{\text{open}} - P_{i,d}^{\text{close}}}{P_{i,d}^{\text{close}}}. \quad (2.6)$$

Whenever considering returns for any of the above types, unless specified otherwise, we use the overnight CL_tmOP returns. In this case, we drop the superscripts and simply write $f_{i,d}$. Also, denote the market-excess returns for asset i at day d

$$\tilde{f}_{i,d} = f_{i,d} - f_{\text{SPY},d}, \quad (2.7)$$

where $f_{\text{SPY},d}$ denotes the future return of the S&P-500 index⁷. The S&P-500 index is a widely used proxy for market returns due to its wide range of large-cap firms and quarterly updates. Considering market excess returns allows us to consider a market-neutral study by considering price movements relative to the market price movement. We refer to market excess returns as EMR and to raw returns as RR.

⁶Note that all returns considered are adjusted for splits and dividends.

⁷Using Risk-Adjusted return using the Fama and French 3-factor model ([88]) does not significantly change the results.

Generally, we are interested in predicting the directional movement of future prices. We denote by $s_{i,d} \in \mathbb{R}$ the observed signal, on day d , for the future price of asset i . Note that $s_{i,d}$ as defined here is deliberately left general and could be either some observed variable or the output of a predictive model. However, in the context of this paper, this will be either a single OVI metric or a linear combination of several OVI metrics (as seen later in this section). Then, the sign of this signal is used to select the direction in which to make a hypothetical transaction in the underlying, with $+$ and $-$ corresponding to buy and sell, respectively. A sequence of such trades constitutes a strategy. To assess the performance obtained using a series of signals, we consider the (excess-returns-adjusted) Profit and Loss (P&L) metric defined by

$$\text{P\&L}_d = \sum_{i=1}^{N_d} b_{i,d} \times \tilde{f}_{i,d} \times \text{sign}(s_{i,d}). \quad (2.8)$$

Unless specified otherwise, we allocate equal capital to each asset i for which $s_{i,d} \neq 0$, i.e $b_{i,d} = 1/N_d$, $\forall i, d$. In other words, we do not differentiate between the different underlying assets, allocating equal capital to each asset i with $s_{i,d} \neq 0$.

The P&L is the difference in profit or loss that would have been earned, had we invested $b_{i,d}$ units of currency (in our case dollars) in comparison to making an equivalent transaction in the benchmark SPY ETF (which is tracking the S&P-500 index).

The P&L is defined above for each day d , and this gives a vector $\mathbf{P\&L} = (\text{P\&L}_1, \text{P\&L}_2, \dots, \text{P\&L}_{D-1})$ of daily P&L's which measures the outcome of the hypothetical strategy over the whole time horizon. A measure of the long-term performance of a strategy is the Sharpe Ratio [213], defined as

$$\tilde{\text{SR}} = \frac{\text{mean}(\mathbf{P\&L})}{\text{sd}(\mathbf{P\&L})}. \quad (2.9)$$

Alternatively, the annualized Sharpe Ratio (hereafter SR) is defined as

$$\text{SR} = \tilde{\text{SR}} \times \sqrt{252}. \quad (2.10)$$

Lastly, we define the P&L per dollar traded (PPD) as

$$\text{PPD} = \frac{\sum_{d=1}^{D-1} \text{P\&L}_d}{\sum_{d=1}^{D-1} B_d}, \quad (2.11)$$

where $B_d = \sum_{i=1}^{N_d} b_{i,d}$ is the total bet size for day d ⁸.

When assessing whether a hypothetical strategy is significantly profitable, we use a test with the null hypothesis $H_0 : \text{SR} = 0$. We make use of this test throughout the

⁸Due to the normalization used in Equation (2.8), this is equal to $D - 1$

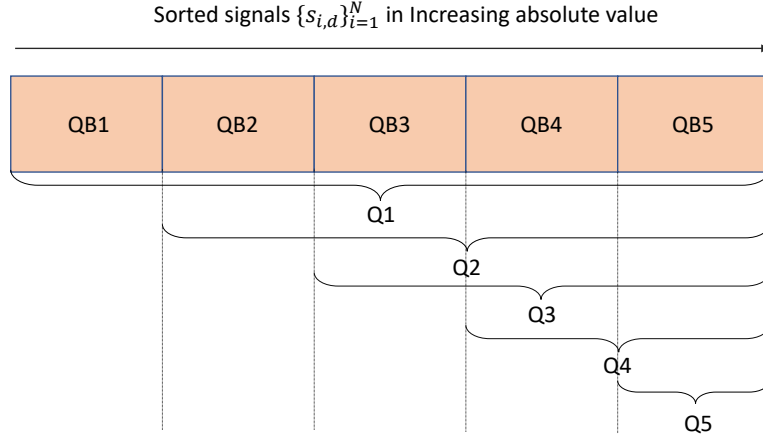


Figure 2.6: Quantile Rank Portfolio Diagram: QB1 - QB5 denote the five quantiles of $\{|s_{i,d}|\}_{i=1}^n$ and Q1-Q5 the upward containing quantile rank portfolios.

paper, by conducting a test in the same fashion as [17], which takes into account the non-normality of returns. In particular, we use the test statistic

$$\frac{\tilde{\text{SR}}(\mathbf{P\&L})}{\sqrt{\frac{1 - \text{skewness}(\mathbf{P\&L}) * \text{SR}(\mathbf{P\&L}) + (\text{kurtosis}(\mathbf{P\&L}) - 1) * (\text{SR}(\mathbf{P\&L})^2) / 4}{T - 1}}}, \quad (2.12)$$

where the denominator term is based on the estimate provided by [20], for the variance of the Sharpe Ratio.

Whenever comparing the SRs of two different strategies, we need to test the hypothesis of the form of $H_0 : \text{SR}_1 = \text{SR}_2$. For that purpose, we use the hypothesis test of [139]. To adjust for the non-normality of returns, they use a studentized bootstrap test on the residuals of a VAR model (in a circular block approach) to adjust for the non-normality of returns.

Quantile rank portfolios. The P&L (2.8) only uses the sign of the observed signal (for example, the OVI), and does not take into account its magnitude. Naturally, a larger magnitude for the signal should entail a higher degree of confidence in the prediction. To that end, for each MPC and each day, we split underlying assets into five groups in the following way.

For each financial day, we divide the underlying assets into five disjoint quantile buckets on the basis of the magnitude (strength) of the corresponding option-market signal, $|s_{i,d}|$. Denote by QB1_d the group of assets with the 20% of weakest signals, and so on up to QB5_d , which corresponds to the 20% strongest signals. We then define the quantile rank portfolios Q1 – Q5 in the following manner; $\text{Q1} = \text{QB1} \cup \text{QB2} \cup \text{QB3} \cup \text{QB4} \cup \text{QB5}$, $\text{Q2} = \text{QB2} \cup \text{QB3} \cup \text{QB4} \cup \text{QB5}$, and so on

up to QB5 = Q5. In other words, the first quantile rank portfolio corresponds to all 100% of signals, and the fifth to just the strongest 20% signals. We demonstrate this in Figure 2.6. Note that when calculating the PnL (2.8), unless using $Q1$, we re-define $b_{i,d} = \frac{1}{|Qk_d|}$, where Qk_d is the Qk rank portfolio formed for day d .

2.4.3 P&L Regression

In our analysis, we consider several variants of OVI based on different option transaction types (Buys, Sells or both), data sets (PHOTO, NOTO or both) and MPCs. We are also interested in analyzing the joint predictive power of such OVI features in explaining future returns. Usually, a linear method is employed (OLS or LASSO Regression), and the corresponding t or p -values are used to compare the explanatory power of different covariates (e.g [187]).

However, a linear model is suitable only for finding linear relationships. Furthermore, the objective function of OLS regression is based on the Euclidean distance between observed and fitted values. Such an objective function can be unsuitable for directional predictions (as predicting the direction is an inherently nonlinear problem). A predictor that is successful in capturing the signs of future returns may be deemed uninformative with respect to the Euclidean distance. In the case of linear models, while a predictor’s sign may match the observations perfectly, we may still observe a small corresponding R^2 value. Finally, some of the options corresponding to an asset may not be traded at all on a given day, meaning that the OVI can be zero for a large number of days. Linear methods are also unable to deal with feature sparsity, as they are unable to ignore zero-valued features. Indeed, while OLS regression is able to show that all of the OVIs are significant predictors of future returns, the corresponding R^2 values are quite low (check Appendix A.3) and it is hard to distinguish between the explanatory power of the different OVIs.

For these reasons, we propose a novel method for regressing the future returns against different variables (in our case, different variants of OVI), where we directly set an objective function equal to the cumulative P&L (2.8), thus searching for a combination of covariates that maximizes net profit. Equivalently, we can minimize the negative P&L. Inspired by the popular LASSO regression [225], we also impose an ℓ_1 penalty term on the objective function. This penalization allows distinguishing between significant and non-significant covariates by assigning larger penalties to non-significant covariates. Similar previous research has optimized the Sharpe Ratio by including it in the objective function for portfolio optimization problems such as

[254] and [50], but in this case, we are looking to combine multiple features for the forecasting of future returns for different assets.

Suppose that we have a total of K variants of OVI. Denote by $A_{i,d,k}$ the k^{th} type of feature/OVI for asset i at day d (e.g $A_{i,d,k} = \text{OVI}_{i,d,m}^{[\mathcal{F}]}$ for some MPC m and some restrictions $\mathcal{F}$). In this case, we define a signal which is a linear combination of the different OVI

$$s_{i,d}(\beta; \mathbf{A}) = \beta_0 + \sum_{k=1}^K \beta_k A_{i,d,k}, \quad (2.13)$$

for some vector $\beta = (\beta_0, \dots, \beta_K)$. The goal is to choose β , in a way that maximizes the profit of a strategy based on the signal in (2.13). In particular, we look for a vector β , which maximizes the total P&L (denoted as PL) defined as

$$\text{PL}(\beta; \mathbf{A}, f) = \sum_{d=1}^{D-1} \text{P\&L}_d = \sum_{d=1}^{D-1} \left(\sum_{i=1}^N b_{i,d}(\beta; \mathbf{A}) f_{i,d} \text{sign}(s_{i,d}(\beta; \mathbf{A})) \right). \quad (2.14)$$

Thus, we look for an objective function that incorporates the total P&L. In order to use gradient methods, we require a smooth objective function. The sign function used in (2.14) is not smooth, therefore we wish to substitute it with a differentiable activation function $g : \mathbb{R} \rightarrow \mathbb{R}$, such as

$$g(x; \alpha_1) = 2 \left(\frac{1}{1 + e^{-\alpha_1 x}} - \frac{1}{2} \right), \quad \alpha_1 > 0. \quad (2.15)$$

This is equivalent to the *tanh* function for $\alpha_1 = 2$, and as $\alpha_1 \rightarrow \infty$ it converges pointwise to the sign function. We can get rid of the sign function by choosing

$$b_{i,d}(\beta; \mathbf{A}) = \frac{|g(s_{i,d}(\beta; \mathbf{A}))|}{\sum_j |g(s_{j,d}(\beta; \mathbf{A}))|}. \quad (2.16)$$

Therefore, Equation (2.14) can be rewritten as

$$\text{PL}(\beta; \mathbf{A}, f) = \sum_{d=1}^{D-1} \text{P\&L}_d = \sum_{d=1}^{D-1} \left(\sum_{i=1}^N \frac{1}{\sum_j |g(s_{j,d}(\beta; \mathbf{A}))|} f_{i,d} g(s_{i,d}(\beta; \mathbf{A})) \right). \quad (2.17)$$

This brings another differentiability issue, with the absolute value function. This is used twice, once for the bet sizes (2.16), and for the ℓ_1 regularization term, and is non-differentiable at 0. There exist many methods to circumvent this, but we shall make use of one of the simpler ones, which is to replace these functions with a smooth approximation, in the same spirit as how we used the modified sigmoid. We use the function h , defined as

$$h(x; \alpha_2) = \sqrt{x^2 + \alpha_2}, \quad \alpha_2 > 0, \quad (2.18)$$

for some small value of α_2 . In summary, the objective function to be minimized takes the form

$$L(\beta; \mathbf{A}, f, q, \alpha_1, \alpha_2, \lambda) = - \sum_{d=1}^{D-1} \left(\sum_{i=1}^N \frac{1}{\sum_j h(g(s_{j,d}(\beta; \mathbf{A})))} f_{i,d} g(s_{i,d}(\beta; \mathbf{A})) \right) + \lambda \sum_{k=1}^K h(\beta_k; \alpha_2). \quad (2.19)$$

To minimize (2.19), we employ the ADAM algorithm [132]⁹. The derivative of the objective function is given by

$$\begin{aligned} \frac{\partial L}{\partial \beta_k} = & - \sum_{d=1}^{D-1} \sum_{i=1}^N f_{i,d} \left(\frac{1}{\sum_j h(g(s_{j,d}(\beta; \mathbf{A})))} \frac{\partial g(s_{i,d}(\beta))}{\partial \beta_k} - \right. \\ & g(s_{i,d}(\beta)) \left(\frac{1}{\sum_j h(g(s_{j,d}(\beta; \mathbf{A})))} \right)^2 \sum_{j=1}^N h'(g(s_{j,d}(\beta))) \frac{\partial g(s_{j,d}(\beta))}{\partial \beta_k} \left. \right) \\ & + \lambda (1 - 1_{(k=0)}) h'(\beta_k), \end{aligned} \quad (2.20)$$

where

$$\frac{\partial g(s_{i,d}(\beta))}{\partial \beta_k} = g' \left(\beta_0 + \sum_{k=1}^K A_{i,d,k} \beta_k \right) A_{i,d,k}. \quad (2.21)$$

Equivalently, written in vector form, this amounts to

$$\begin{aligned} \frac{\partial L}{\partial \beta} = & - \sum_{d=1}^{D-1} \sum_{i=1}^N f_{i,d} \frac{1}{\sum_j h(g(s_{j,d}(\beta; \mathbf{A})))} \frac{\partial g(s_{i,d}(\beta))}{\partial \beta} \\ & + \sum_{d=1}^{D-1} \left(\sum_{i=1}^N f_{i,d} g(s_{i,d}(\beta)) \left(\frac{1}{\sum_j h(g(s_{j,d}(\beta; \mathbf{A})))} \right)^2 \right) \\ & \times \left(\sum_{j=1}^N h'(g(s_{j,d}(\beta))) \frac{\partial s_{j,d}(\beta)}{\partial \beta} \right) \\ & + \lambda [0, h'(\beta_1), \dots, h'(\beta_K)]^T. \end{aligned} \quad (2.22)$$

Lastly, once the algorithm converges and a vector $\hat{\beta}$ is obtained, we rescale the coefficients so that the non-intercept coefficients have a total absolute value of one, i.e

$$\tilde{\beta} = \frac{\hat{\beta}}{\sum_{k=1}^K |\hat{\beta}_k|}. \quad (2.23)$$

⁹Using the ADAM optimizer, requires choosing two parameters, the so-called moment parameters. For this paper, we use the recommended values given in the original paper (0.9 and 0.999).

Since β only affects the P&L only through the signal s (see (2.13)), rescaling all of the coefficients does not affect the objective function if we are using the sign function. However, note that due to using the proxy *tanh* function, there is in fact a small difference that is caused by the rescaling.

As already mentioned, this method has various advantages over standard linear regression. For example, it is better able to uncover monotonic relationships than standard linear regression. The disadvantage is that, unlike OLS, the minimization of Equation (2.19) does not have a closed-form solution nor is it convex, and we are also burdened with the choice of λ . On the other hand, the magnitude of the fitted coefficient β has a direct interpretation as the weighting factor on which the buy/sell/neutral decision is made.

Throughout the rest of the paper, signals $s_{i,d}$ are either equal to a particular type of OVI or take the form (2.13). In the first case, the strategy does not require any parameters and so remains the same throughout the whole evaluation period. In the second case, the model is trained on more recent data via a sliding-window approach.

In particular, we set a window size of l of 500 trading days (about 2 years), a testing window size of $T = 100$ days¹⁰. A window at day d , contains the range of days $\{d - l + 1, \dots, d\}$. For each window, we regress the future returns against the OVI features, using the P&L regression described above, to obtain coefficients $\{\beta_k^{(d-l+1):d}\}_{k=1}^K$. After obtaining the fitted coefficients $\{\beta_k^{(d-l+1):d}\}_{k=1}^K$, we calculate the P&L in (2.17) using the days in the testing set for the window at day d

$$\text{Test}_d = \{d + 1, \dots, d + T\}, \quad (2.24)$$

i.e $\text{PL}(\{\beta_k^{(d-l+1):d}\}_{k=1}^K; A_{[\cdot, \text{Test}, \cdot]}, f_{[\cdot, \text{Test}]}, q, \alpha_1, \alpha_2)$. This is an out-of-sample metric, with the direct interpretation of being the hypothetical profit one would make in 100 days of trading by following such a strategy (in the absence of trading costs).

2.5 Main Empirical Results

In this section, we assess the predictive power of OVI against the future returns across different MPCs, across different return types (overnight, intraday and close to close) for the PHOTO data set. Among the plethora of testing carried out, we get a clear indication of significant results, which while they seem to vary in terms of profitability, remain mostly consistently statistically significant.

¹⁰Note that performance is not significantly affected by the choice of parameters l and T .

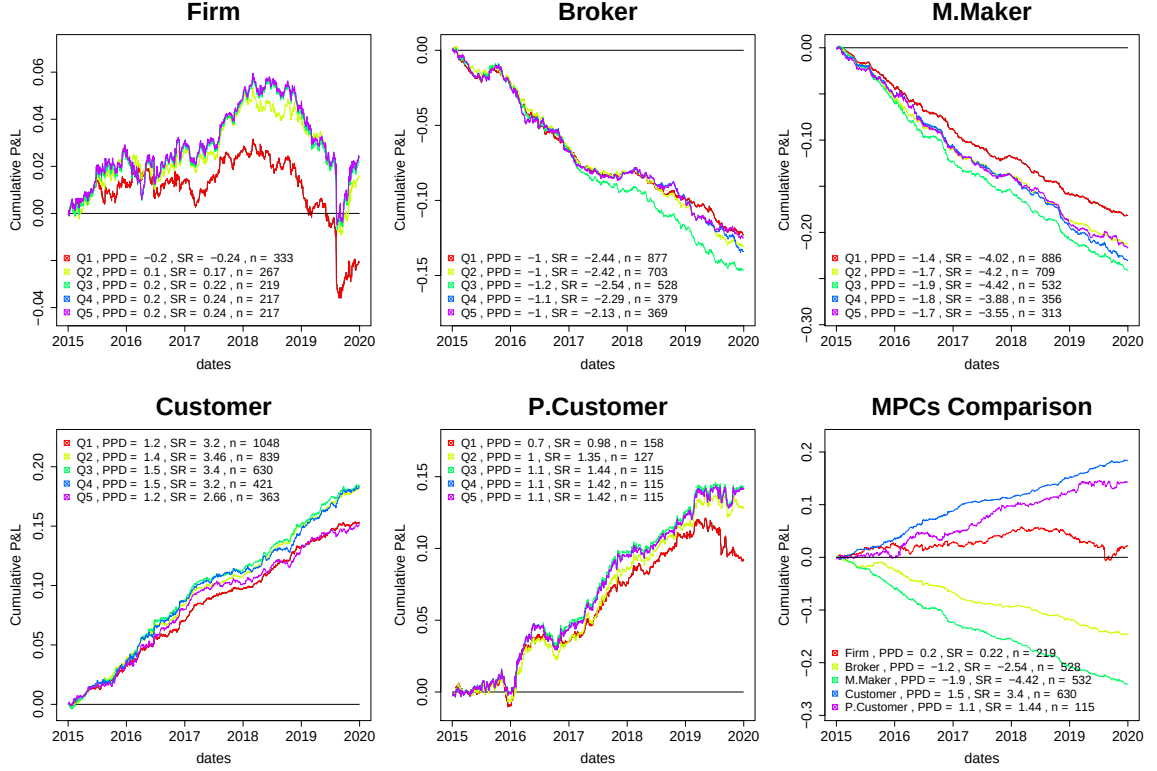


Figure 2.7: Cumulative P&L plot - Overnight Returns: This figure shows the cumulative P&L by following a uniform scheme - one-day holding period strategy, for the five QR groups Q1-Q5 and each MPC, for overnight returns (EMR_CL_tmOP). Also given, are SR (Annualized Sharpe Ratio), daily PPD (Profit per dollar in basis points), and n (the average number of instruments in the portfolios). The first five subplots show the 5 QR groups for each MPC, while the last subplot corresponds to using the 3rd QR group for each MPC.

2.5.1 Comparison across MPCs

We consider the OVI for each of the 5 MPCs separately by using their sign as the signal for future returns ($s_{i,d} = \text{OVI}_{i,d,m}$ for each m). We then evaluate their corresponding P&L using Equation (2.8), for their excess market overnight returns (EMR_CL_tmOP), and for different quantile ranked portfolios (Q1-Q5) built as described in Section 2.4. In particular, we plot the time series for the cumulative P&L incurred over the period 2015-2019 in Figure 2.7.

The highest performance is observed for Market Makers, where the cumulative returns are between 17% and 25% and annualized SRs lie between 3.5 and 4.4 depending on the quantile rank portfolio used. The second best performance is observed for Customers, whose cumulative P&L display a clear upward trend and SRs with magnitudes in the range 2.7–3.5. Strong positive signals are also observed for Brokers,

with SRs between 2.1 and 2.5. Lastly, Professional Customers display a downward trend with SRs of magnitudes 1–1.4. For Firms, while the total P&L is slightly positive, there does not appear to be a clear upward trend, with SRs very close to 0. Note that there seems to be a trend shift happening after 2018 for Firms, from a somewhat upward trend to a downward trend (a trend shift can also be noticed for other MPCs).

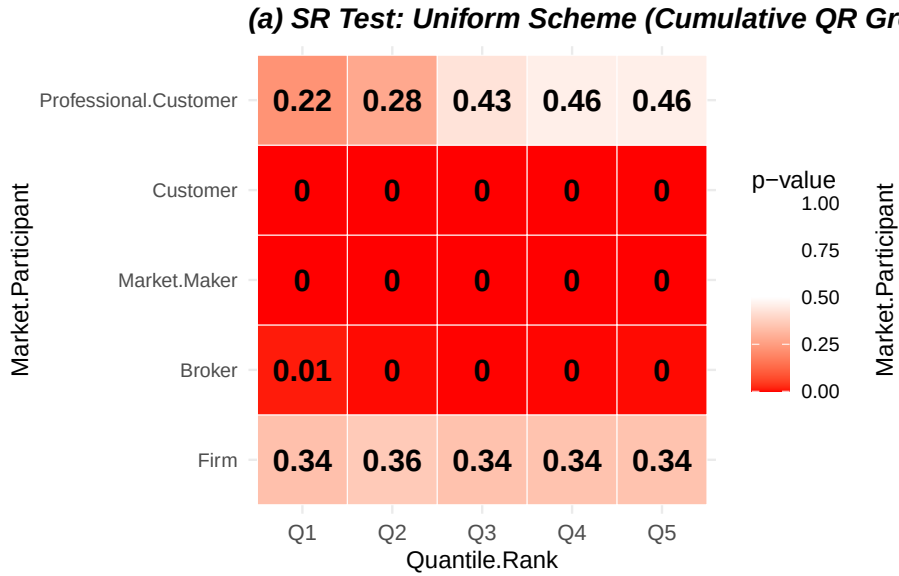


Figure 2.8: Sharpe Ratio Significance tests: The grid shows the p -value for the significance of the Sharpe Ratio for the different strategies. The coloring shows the significance level for the strategies, where red denotes statistically significant returns associated with that strategy (either positive or negative). Sub-figure(a) uses the cumulative QR groups ranks where as sub-figure(b) uses the (distinct) quantile buckets.

In Figure 2.8, we report the p -values obtained when running a significance test for the SRs corresponding to excess market CL_tmOP returns. The tests for *Brokers*, *Market Makers*, and *Customers* are rejected with a p -value of 0 (up to 2 decimal places) across all quantile ranks. *Firm* and *Professional Customers* have p -values close to 0.5 across all ranks. Thus, in the span of five years, we observe a clear signal from three out of five of the MPCs, with the *Market Maker* achieving a return of about 30% in a period of 5 years. The two participants without a clear signal are also the MPCs with the lowest mean volume per day (see Figure 2.1), which partly accounts for the reduced performance. The clearest signals come from *Market Makers*, followed by *Customers*, which have on average the highest volume per day. These observations clearly mark the OVI as having predictive power for future returns. Note that this

verifies the findings of the earlier study [238], who reported that the informational context of Customers volumes is higher than Firm volumes.

For all MPCs except Firms, the best performing portfolio is Q3 both in terms of PPD and in terms of SR. The worst performing portfolio in terms of PPD is Q1, which in most cases has a noticeably worse performance than other QR groups. In terms of SR, the worst performing portfolio is either Q1 (all assets) or Q5 (top 20% of strongest signals) depending on the MPC. Therefore, the strategies of taking into account all the non-zero OVIs or only considering a very small subset of the strongest signals are both outclassed by following a portfolio consisting of the top 60% or 80% of the highest-magnitude OVIs. This confirms that not only is the sign of the OVI informative on future returns, but that the predictability observed is stronger for higher magnitudes of the OVI, which is typical for most profitable financial signals.

2.5.2 Comparison across Return Types

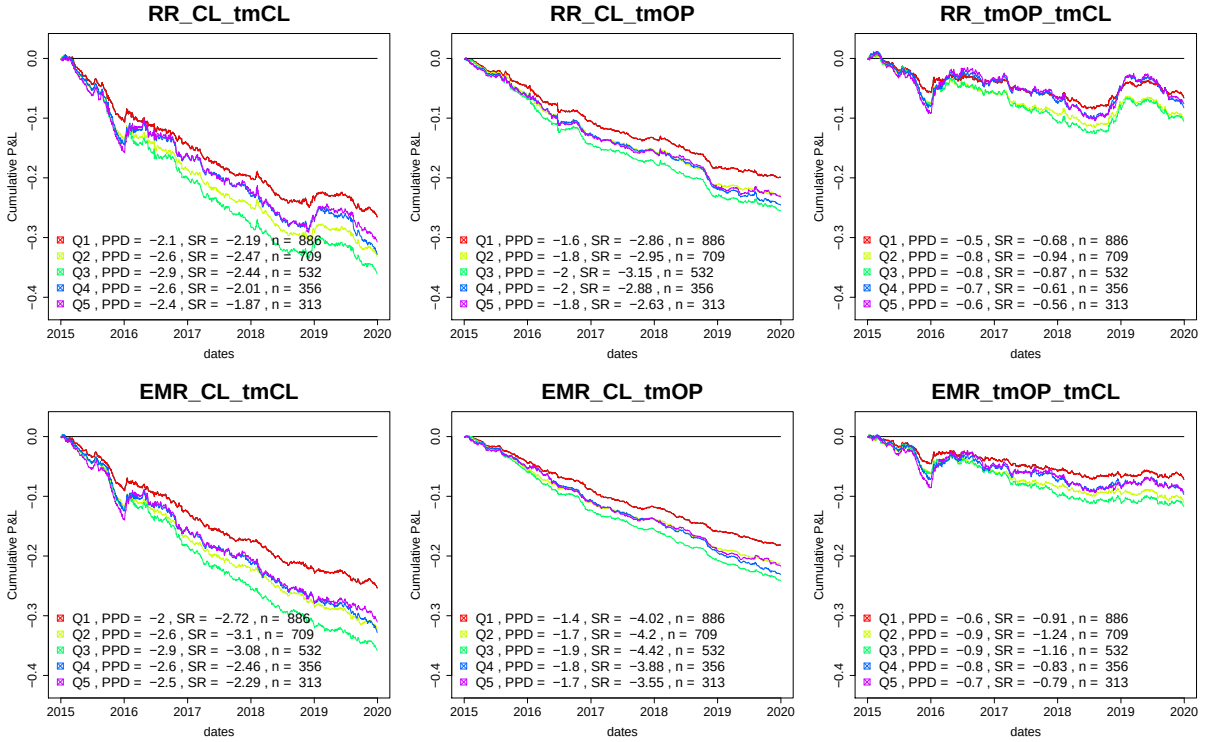


Figure 2.9: Cumulative P&L plot - Market Maker: This shows the cumulative P&L from following a uniform weighting scheme - one-day holding period strategy, for the five quantile rank groups Q1-Q5, using the OVI of Market Makers for six different types of returns. Also given, are SR (the Annualized Sharpe Ratio), PPD (Profit per dollar in basis points), and (the average number of instruments in the portfolios).

In Figure 2.9, the cumulative P&L corresponding to each type of return is shown, when using the Market Maker OVI. The highest predictability is clearly observed for the overnight (CL_tmOP) returns, followed by the close-to-close (CL_tmCL), and then by the open-to-close (tmOP_tmCL) returns. Close-to-close return is a combination of the other two types of returns, with most of its predictability stemming from the overnight returns rather than the next-day returns. Higher performance is shown for market excess returns, but significant results across MPCs and quantile rank portfolios are also shown for raw returns.

2.5.3 Comparison Across OVI Types

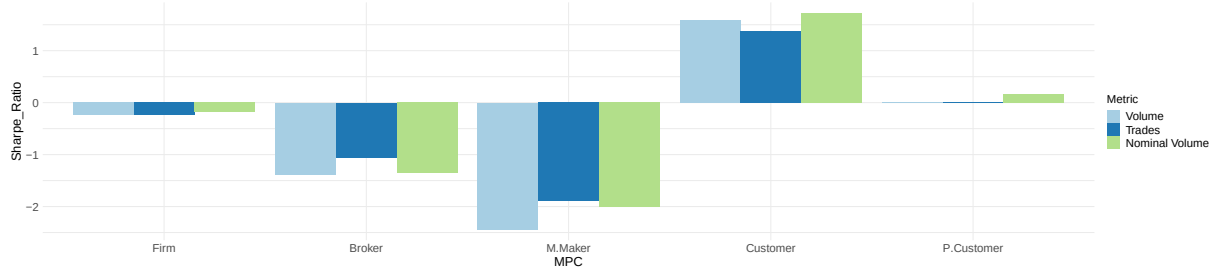


Figure 2.10: Comparison across OVI types: For each MPC, the SR corresponding to EMR CL_tmOP returns are shown, for each OVI type (as defined in Section 2.2.2). We use a uniform betting scheme, Q3 strategy for each are shown for each OVI type.

Throughout this section, we have been using the Volume-based OVI. In Figure 2.10, we compare the different types of OVI (Q3 ranked portfolio) across MPCs. Trade-based OVI has the lowest performance for Brokers (significantly lower than both volume and nominal volume based OVI), Market Makers (significantly lower than volume based OVI) and Customers (significantly lower than nominal volume based OVI). This is expected, as it is naturally the least informative of the three types of features, compared to volume and nominal volume OVI. Volume-based OVI scores the highest (in magnitude) SR for Market Makers and Brokers, and the difference with the other two types for Market Makers is significant when running a pairwise SR Difference test.

2.5.4 Multi-factor Analysis and Stability Analysis across time

In predicting future overnight returns, using the signal given by the sign of the Market Maker OVI has resulted in the highest performance. The Market Maker signal can

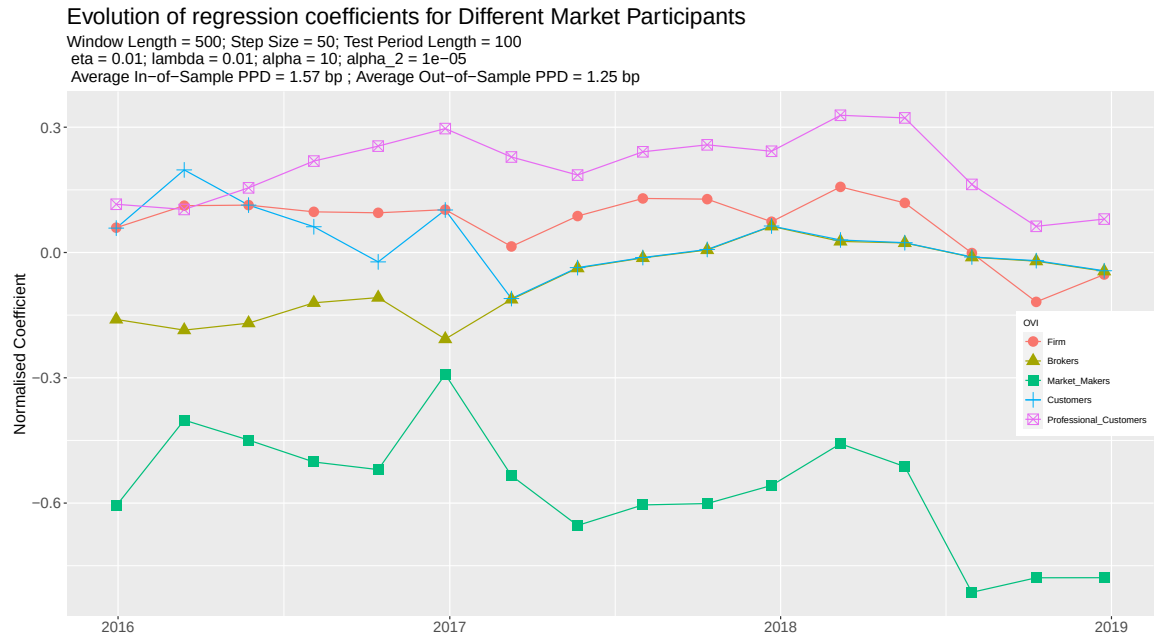


Figure 2.11: Multiple-factor regression for different MPCs: The line graphs, show the evolution of the coefficients in equation (2.26), fitted using the P&L regression for M2 across the different windows, as indicated by the plotted points (a point indicates the end of a window of length $L = 500$). Note that coefficients are normalized to have a total ℓ_1 norm equal to 1 (excluding the intercept term).

be written as

$$M1 : s_{i,d}^{(1)} = \text{OVI}_{i,d,\text{M.Maker}}^{\text{PHOTO}}. \quad (2.25)$$

In this subsection, we evaluate the joint predictive power of the OVI corresponding to the five different market participants (K=5) using the P&L regression method described in Section 2.4 to predict the future EMR CL_tmOP return by using a signal of the form

$$\begin{aligned} M2 : s_{i,d}^{(2)}(\beta) = & \beta_0 + \beta_1 \text{OVI}_{i,d,\text{Firm}}^{\text{PHOTO}} + \beta_2 \text{OVI}_{i,d,\text{Broker}}^{\text{PHOTO}} \\ & + \beta_3 \text{OVI}_{i,d,\text{MarketMaker}}^{\text{PHOTO}} + \beta_4 \text{OVI}_{i,d,\text{Customer}}^{\text{PHOTO}} + \beta_5 \text{OVI}_{i,d,\text{P.Customer}}^{\text{PHOTO}}. \end{aligned} \quad (2.26)$$

We regress in a sliding window manner and record the evolution of the normalized coefficients during 2016-2019, as shown in Figure 2.11. Note that at each time point, the coefficient of *Market Maker* OVI dominates, with a magnitude of 0.3-0.75. Generally, the Market Maker OVI seems to have a higher weight for the later years, which is partly due to the coefficients of Broker and Customer OVI shrinking in magnitude. Firms show mostly negative coefficients with low magnitudes for the entire 5-year period. Interestingly, the OVI for Professional Customers has a consistently negative coefficient throughout the 5-year period, with a relatively high magnitude. Since the volumes of Professional Customers are quite low (i.e the OVI is equal to zero in most cases), the high coefficient does not disrupt the model (which would not have been possible if we used a linear regression model). On the other hand, as seen in the correlation map in Figure 2.3, the OVI for Professional Customers is not correlated with the other features, which potentially shows that additional information is brought when it is used in conjunction with Market Makers' OVI.

Since running the P&L regression results in a coefficient vector dominated by the Market Maker's OVI, it is worth comparing the performance of signals $s^{(1)}$ and $s^{(2)}$. The two strategies give similar mean in-sample PPDs with 1.50bps and 1.57bps respectively. However, using $s^{(1)}$ results in a slightly higher out-of-sample mean PPD of 1.33bps compared to 1.25bps for $s^{(2)}$, indicating some overfitting in the second case.

2.6 Influence of Options Features on OVI Predictability

In the previous section, we showed how the OVI can be used as a predictor for next-day returns. In this section, we investigate which options carry most of the predictability by segmenting options according to the value of certain option features, namely a subset of the "Greeks", *moneyness*, *maturity*, and *implied volatility*; we calculate

all of these features in a standard Black–Scholes model without dividends. One might naturally expect option moneyness, being a measure of the relative sizes of the spot and strike prices, to influence OVI predictability. Indeed, [187] and [98] both find that the highest predictability over comparatively long timescales stems from volumes of OTM options, which is attributed to the relatively high leverage that they offer speculative traders targeting (or hedging) large price moves over these timescales. However, as we consider next-day returns, we expect the performance of ATM options to be higher, as their informational content is more geared towards small price moves than OTM options. On the other hand, ITM options’ intrinsic value results in a higher price, hindering the high leverage advantage offered by the option market and resulting in relatively low volumes.

Note that there is not a one-to-one map between moneyness and other Greeks; while the Greeks are related in various ways, different Greeks can serve to segment the universe of options in different ways. For example, Γ is highly correlated with the absolute value of moneyness, while Δ depends monotonically on moneyness. As we see below, this enables us to consider a range of perspectives.

For each day, we divide the option contracts into four disjoint buckets based on the quantiles of the feature in question. The first bucket contains options corresponding to the lowest 25% of values for that feature, and so on up to the 4th bucket, which contains options corresponding to the highest 25%. The OVI pipeline described in the previous sections is then applied to each bucket separately, and the performance of the resulting buckets is compared. For each feature, we run the pairwise SR difference test between the four buckets (at the 5% level). We only follow this process for Brokers, Market Makers and Customers, as they are the MPCs whose OVI was concluded to be a significant predictor of future returns.

Below we show some interesting results based on the Δ , as well as the implied volatility of options, calculated (for simplicity and ease of computation) using the Black–Scholes (BS) model for European options without dividends [34]¹¹. We also provide a short discussion on moneyness and maturity to interpret the results.

¹¹Note that since equity options are mostly American, the Black–Scholes model may systematically undervalue them. However, as these differences are typically of the order of 1–2%, we accept the inaccuracy and avoid large-scale computations.

Breakdown of Data Set based on Option Greeks

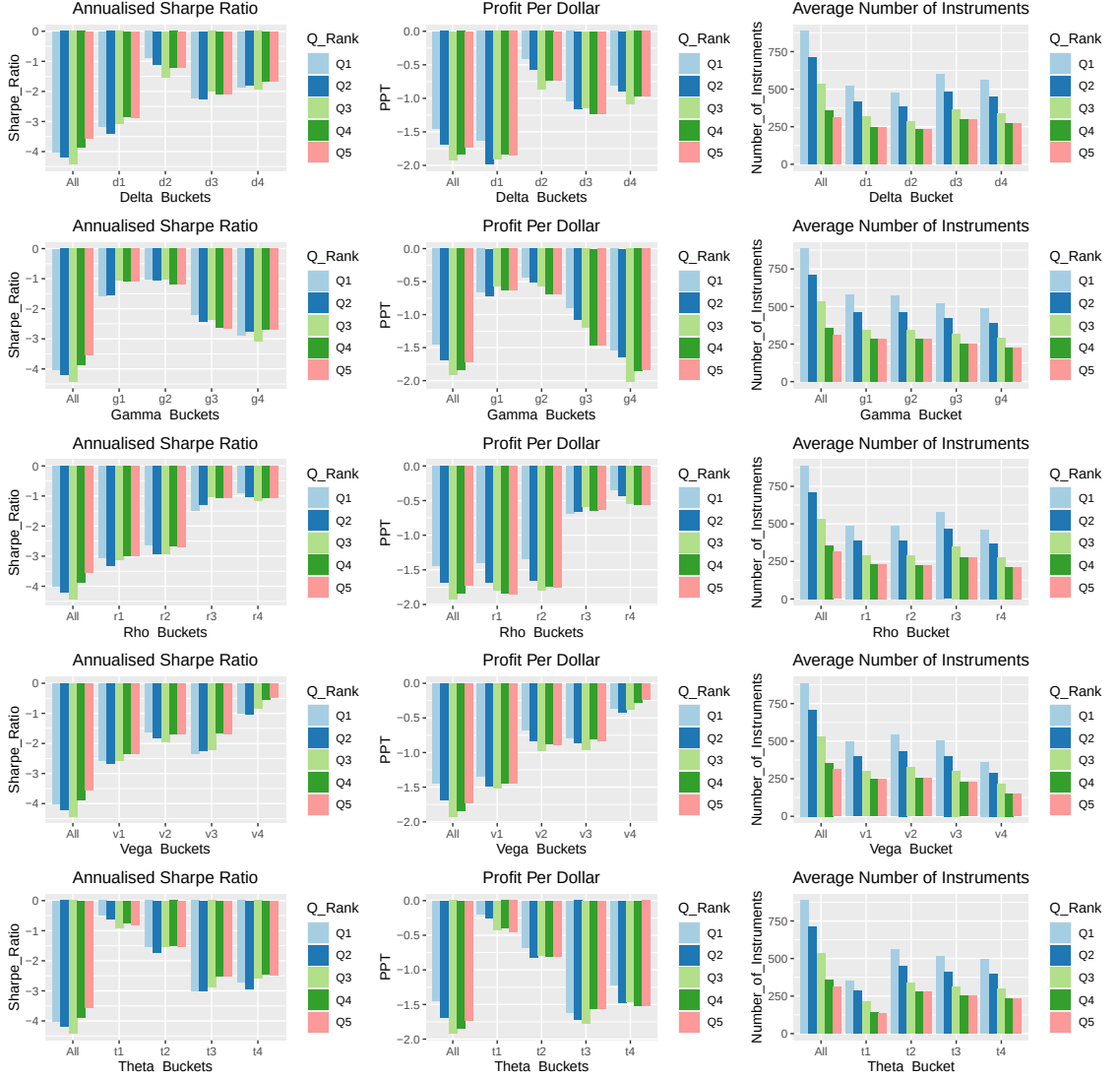


Figure 2.12: Summary statistics from the breakdown of the PHOTO option data set for Market Makers into four buckets based on the value of some Greek. For each feature, the first bucket corresponds to option contracts whose features have the lowest values, whereas the last bucket corresponds to option contracts whose features have the highest values. The label 'All' corresponds to using the complete data set, without conditioning on the values of a Greek. Quantile Groups are defined as in Section 2.4, while buckets are based on the quantiles of the Greek corresponding to each option contract. The sub-figures for each Greek correspond to the Annualised Sharpe, Profit Per Dollar, and Average Number of Instrumtents used in the portfolio.

Under the BS model, a call option has a price P^{Call} equal to

$$P^{\text{Call}}(S, K, \tau, r, \sigma) = Se^{-r\tau}\Phi(d_1) - Ke^{-r\tau}\Phi(d_2), \quad (2.27)$$



Figure 2.13: Summary statistics from the breakdown of the PHOTO option data set for Market Makers into four buckets based on the value of some option feature. For each feature, the first bucket corresponds to option contracts whose features have the lowest values, whereas the last bucket corresponds to option contracts whose features have the highest values. The label 'All' corresponds to using the complete data set, without conditioning on the values of a Greek. Quantile Groups are defined as in Section 2.4, while buckets are based on the quantiles of the Greek corresponding to each option contract. The sub-figures for each Greek correspond to the Annualised Sharpe, Profit Per Dollar, and Average Number of Instrument used in the portfolio.

and a put option has a price P^{Put} equal to

$$P^{\text{Put}}(S, K, \tau, r, \sigma) = Ke^{-r\tau}\Phi(-d_2) - Se^{-r\tau}\Phi(-d_1), \quad (2.28)$$

where

$$d_1(S, K, \tau, r, \sigma) = \frac{\log(S/K) + (r + \sigma^2/2)\tau}{\sigma\sqrt{\tau}} \text{ and} \quad (2.29)$$

$$d_2(S, K, \tau, r, \sigma) = d_1(K, S, \tau, r, \sigma) - \sigma\sqrt{\tau}. \quad (2.30)$$

Here, S refers to the current price of the underlying asset (in our case, we use the average of close and open), r refers to the risk-free interest rate, τ to the time until maturity, K to the strike price, σ the unobserved volatility of the underlying and Φ the cumulative distribution function of a normal distribution. Lastly, $P^{\text{Call}} / P^{\text{Put}}$ denote the average of the opening and closing prices. The implied volatility σ_I , is the value of σ that solves (2.27) / (2.28) (with $P^{\text{Call}} / P^{\text{Put}}$ being the average of the opening and

Breakdown of Data Set based on Option Features

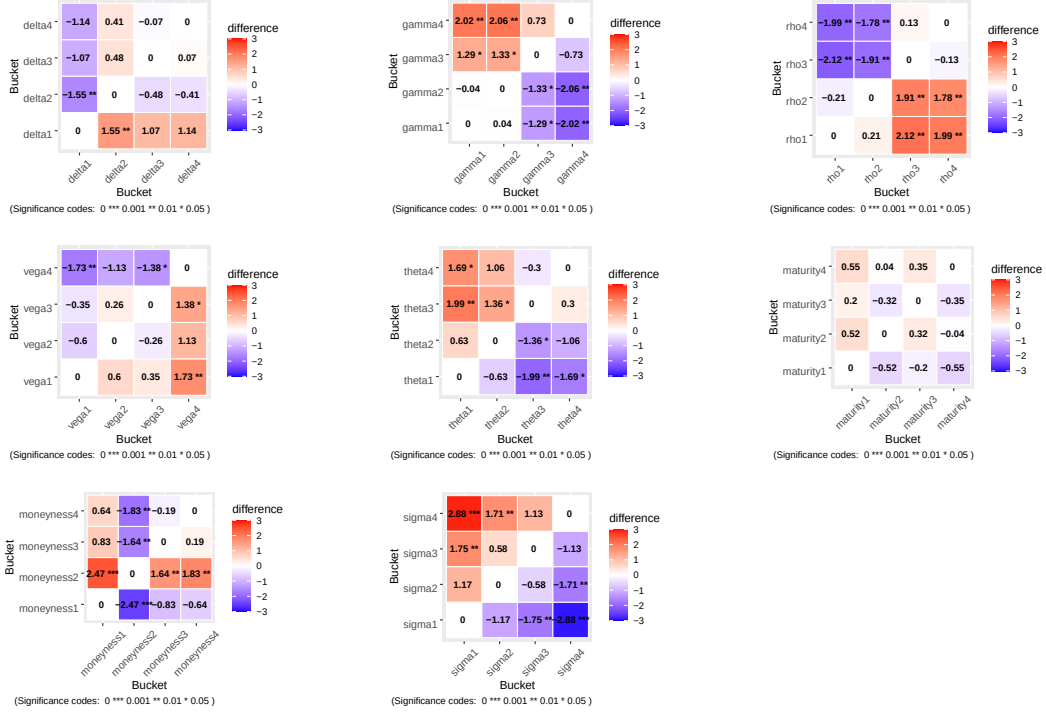


Figure 2.14: This figure shows the pairwise SR difference tests. In each of the grids depicted in the above figure, the results of running the pairwise SR difference tests for different buckets are shown. The difference shown in each cell is the difference between the row and column bucket and is colored red (resp. blue) if the row bucket has a higher (resp. lower) SR than the column bucket. The significance level is also given by the number of stars (1-3), if applicable.

closing prices), and can be obtained using a simple iterative method. Importantly, note that all these quantities describe an option from the buyer-perspective (e.g all call options have a positive Δ).

In Figure 2.13 and Figure 2.12, we plot the results for the option Maturity τ , standardized moneyness defined as

$$\text{Mon} = \begin{cases} +\log(Se^{r\tau}/K)/(\sigma\sqrt{\tau}) & \text{for call options} \\ -\log(Se^{r\tau}/K)/(\sigma\sqrt{\tau}) & \text{for put options,} \end{cases} \quad (2.31)$$

and implied volatility σ_I , as well as the five main Greeks (Delta, $\Delta = \partial P/\partial S$; Gamma, $\Gamma = \partial^2 P/\partial S^2$; Theta, $\Theta = \partial P/\partial t$; Rho, $\rho = \partial P/\partial r$; and Vega, $\mathcal{V} = \partial P/\partial \sigma$).

Moneyness, Gamma and Theta. Consistent with our expectations noted above, ATM options show the highest predictability. For example, across MPCs, higher Gamma options/higher Gamma buckets (corresponding to ATM options) tend to

perform better. For Market Makers, the 4th Gamma bucket significantly outperforms the others, clearly indicating how a higher Γ is linked to higher predictability. Likewise, higher Θ (also closer to being ATM) have the highest performance, with the 3rd and 4th buckets significantly outperforming the first two ones. Results are not as clear with moneyness buckets, with no consistent pattern followed by the buckets except that the 1st bucket is outperformed by the rest, hinting that deep OTM options have the lowest informational content.¹²

(Non) Effect of Maturity (τ). Maturity is often expected to play a significant role in predictability. Similar to how high moneyness options are sometimes filtered out, one might consider excluding options with very high maturities due to potential poor informational content. However, analysis across different maturity buckets reveals no consistent pattern in MPCs, suggesting that maturity does not affect predictability. Therefore, it is unnecessary to filter option contracts based on their maturity, unlike what is commonly done in similar studies.

Difference between call and put options. By definition, call options have a positive Δ , while put options have a negative Δ . In addition, as the future value of assets rises with interest rates, call options also have a positive ρ , while put options have a negative ρ .

We can clearly observe that put options play a more important role in predicting future returns compared to call options. Across all MPCs, the first bucket for Delta (consisting mostly of ITM put options) outperforms the other buckets (consisting mostly of call and OTM put options). For Market Makers and Customers, the 1st bucket significantly outperforms the other buckets. In addition, the first two buckets for ρ (corresponding to the Put Options) are the best performing, with the difference being statistically significant for all MPCs.

Implied Volatility (σ) and Vega ($\mathcal{V}$). For all of Brokers, Market Makers, and Customers, higher implied volatility results in higher predictability. In particular, the 4th bucket is the best performing bucket across all MPCs, and significantly outperforms all other buckets except the 3rd bucket for Brokers. The 4th bucket also outperforms the complete data set in terms of PPD for Brokers and Market Makers, as well as in terms of the SR for Brokers. The 3rd bucket is also consistently the

¹²Replicating this analysis with five buckets still fails to produce any significant results. This is mainly because the five buckets do not exactly correspond to deep OTM, OTM, ATM, ITM and deep ITM, as the moneyness distribution is skewed towards OTM options.

second best performing, significantly outperforming the first two buckets for Market Makers, as well as the 1st bucket for Brokers. Overall, these results clearly indicate that most of the predictability stems from high implied volatility contracts.

For Vega, no consistent pattern can be observed across MPCs. Thus, while implied volatility plays a big role in determining the informational content of an option's volume, the sensitivity of options to implied volatility does not.

2.7 Extensions

In this section, we build on the results of Section 2.5, by performing further tests for more general scenarios. These include checking the price persistence for longer holding periods, comparing the PHOTO and NOTO data sets, and conducting a cross-sectional analysis.

2.7.1 Short-Term Versus Long-Term Price Persistence

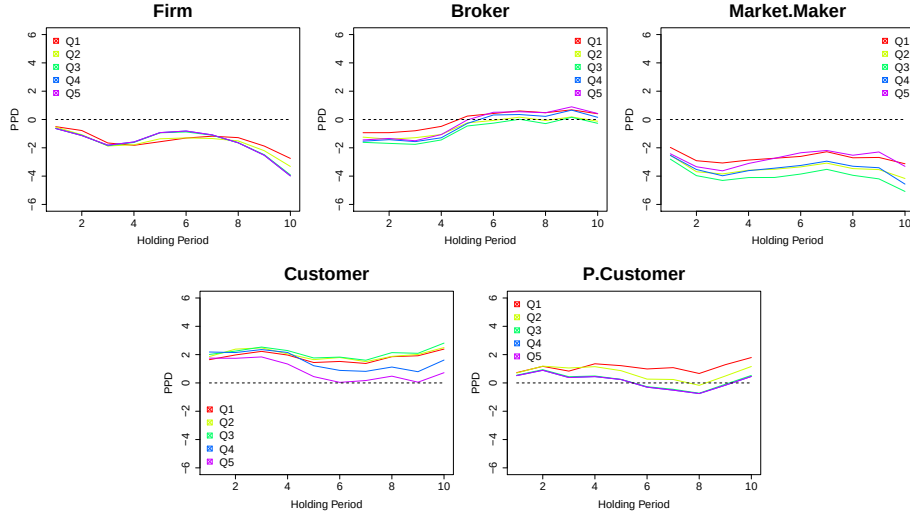


Figure 2.15: Price Impact over longer holding periods: This figure shows the cumulative returns obtained for different holding periods ranging between 1 and 10 days. Results are shown for each MPC, when the corresponding OVI strategy is followed, for each QR Group and the Uniform Weighting Scheme.

Previously, we observed that for Market Makers, Customers and Brokers, a strategy based on the OVI with a holding period of one day results in either significant profits or losses. In a similar fashion to [187], we test how the results change for different holding periods. For each strategy, we record the cumulative returns obtained

when using a longer holding period. Therefore, for a holding period of h days we calculate the P&L as

$$\text{P\&L}_d^h = \sum_{i=1}^{N_d} b_{i,d} \times \left(\sum_{k=1}^h \tilde{f}_{i,d+(k-1)}^{\text{CL_tmCL}} \right) \times \text{sign}(s_{i,d}), \quad (2.32)$$

and plot the cumulative returns in Figure 2.15, for holding periods of up to 10 days.

For Market Makers, aside from a significant SR displayed during the first day, the SR corresponding to the (non-cumulative) returns of the second day after execution is also statistically significant (with a p-value of 0.00). The cumulative returns do not revert over the following days, clearly demonstrating persistence in price changes.

For the rest of the MPCs, only the first-day (non-cumulative) returns are significant. There is a long-lasting effect shown for Customers' Q4 and Q5, where cumulative returns do not significantly decrease after the first day, not displaying any mean reversion. For Brokers, this is not the case, as the cumulative returns revert signs from around the fifth day onward. The other two MPCs did not produce significant results for the first day.

Overall, as the returns of a strategy (based on the Market Maker OVI) with a longer holding period do not appear to revert, the predictive power does not seem to stem from the dynamics of price pressure but rather unravels a persisting price trend, supporting the observations of [187] that speculators and informed traders utilize the options market. Note that an analysis like [101] is impossible as in our case a contract with zero volume simply has a zero OVI, which is be unusable for a trading strategy.

2.7.2 Difference Between Buys and Sells

Recalling the definition of OVI (2.3), the terms $V_{i,j,d,m}^{\text{Up}}$ and $V_{i,j,d,m}^{\text{Down}}$ are defined as using the volumes from both the buying and selling activities of a MPC. In order to gain a better understanding of which type of transaction is the major source of predictive power we calculate the OVI restricted to only the “buy” transactions (resp. sells) denoted as $\text{OVI}_{i,d,m}^{\text{Buy}}$ (resp. $\text{OVI}_{i,d,m}^{\text{Sell}}$). We compare both of these with the standard volume-based OVI feature ($\text{OVI}_{i,d,m}$), by using the SR difference test. As can be seen from Figure 2.16, $\text{OVI}_{i,d,\text{Customer}}^{\text{Buy}}$ performs significantly worse than $\text{OVI}_{i,d,\text{Customer}}$ across QR Groups, with its OVI being almost uninformative. $\text{OVI}_{i,d,\text{Customer}}^{\text{Sell}}$ also has a lower SR than $\text{OVI}_{i,d,\text{Customer}}$, but the difference is not statistically significant. This implies that for Customers most of the predictability comes from the selling activity. The opposite is true for Brokers, where $\text{OVI}_{i,d,\text{Broker}}^{\text{Sell}}$ performs significantly worse than $\text{OVI}_{i,d,\text{Broker}}$, but $\text{OVI}_{i,d,\text{Broker}}^{\text{Buy}}$ has a slightly higher performance than $\text{OVI}_{i,d,\text{Broker}}$.

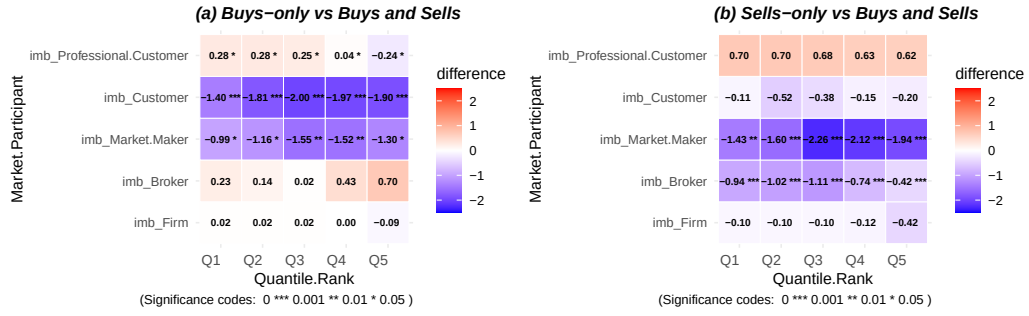


Figure 2.16: SR Difference Tests for Buy / Sell: The left (resp. right) figure shows the difference in SRs obtained between using only the buys (resp. sells) vs using the whole data set. Blue cells denote cases where the using only the buys (resp. sells) is outperformed by the complete data set, and red when it outperforms the complete data set instead. The significance level is also given after the number by 1-3 stars, if applicable.)

When it comes to Market Makers, both differences are significant across all quantile rank portfolios. Recall that the differences in volume for buying and selling activities were small for all of the MPCs (see Figure 2.2). Therefore, the above results could indicate that Customers utilize their speculative ability (or possibly inside information) mostly by writing option contracts. This serves to illustrate an important benefit of decomposing the OVI into different MPCs.

2.7.3 Robustness Under Different Market Conditions

To further demonstrate the robustness of our results, we consider how they change under different market conditions. By focusing on an overnight strategy corresponding to the Market Maker OVI, we consider the following criteria corresponding to proxies of the trend, volatility, and volume of the market

- Daily Momentum – Previous close-to-close return of SPY,
- Weekly Momentum – Sum of five most recent close-to-close SPY returns,
- Monthly Momentum – Sum of 21 most recent close-to-close SPY returns,
- Volatility – Previous day value of the VIX index,
- Trading Volume – The average turnover of 600 NYSE stocks.

For the first four criteria, we consider whether the current value falls above the 66th percentile (in which case we classify the day as *High*), below the 33rd percentile (in

which case we classify the day as *Low*), or in between (in which case we classify the day as *Medium*). For Trading Volume we only consider the 33rd and 66th percentiles of the corresponding values of the last 100 trading days.

As it can be seen in Figure 2.17, the strategy achieves clear trends with statistically significant Sharpe Ratios across all of these regimes. Interestingly, in the High Daily and Weekly Momentum regimes, returns appear to be significantly lower than on days with low momentum, which denotes a significant decrease in the informational content of the option volumes during the bullish states of the market (classified as such based on daily or weekly data).

2.7.4 Additional Features from the NOTO Data Set

We have so far considered only the PHOTO data set corresponding to the PHLX option exchange. We now consider the NOTO data set, which corresponds to the NASDAQ exchange. In Figure 2.18, we show a complete overview of the SR corresponding to the three types of returns for the five MPCs across quantile rank groups, using either buys, sells, or both for the calculation of the OVI feature. Furthermore, we show these values when using NOTO, PHOTO or the union of the two data sets. About half of the values in the table have p-values less than 0.05 for the SR significance. In addition, the colored cells denote the cells that have a significant SR when Bonferroni corrected across all tests, i.e using a p-value of $0.05/(5 \times 5 \times 3 \times 3 \times 3) \approx 7.4 \times 10^{-5}$.

Similarly to the PHOTO data set, the highest performance corresponds to the CL_tmOP future returns for Customers and Market Makers, while Brokers also have a significantly good performance. Another similarity is that the middle quantile rank groups (Q2-Q4) tend to outperform Q1 and Q5. In cases of statistical significance, CL_tmOP and CL_tmCL future returns have identical signs, with the CL_tmOP returns consistently scoring higher SRs.

For the NOTO data set, tmOP_tmCL returns display higher SRs than for PHOTO, with some of them being significant (when ignoring the Bonferroni correction), such as Broker sells or Customers buys. Note that in both of these cases, the sign of SR is in the opposite direction of what is displayed by the CL_tmOP returns, showing a price reversal effect between the overnight and tmOP_tmCL returns, which is not as strongly present in the PHOTO data set. In fact, one can further notice that no result corresponding to CL_tmCL returns is significant for the NOTO data set. This seems to indicate a sharper price reversal present in the NOTO data set. Coupled with the highest SRs scored by the PHOTO data set, this could be indicating that a

higher proportion of traders in the PHLX exchange are informed when compared to the NOM exchange.

Due to the higher median daily volume displayed for NOTO (as seen in Figure 2.1), the combined OVI's SR takes the same sign as the SR of the corresponding NOTO value in almost all cases. When a combined OVI is significant, we observe a weak result when using PHOTO's OVIs and a strong result when using NOTO's OVIs (that is weaker than the result of using the combined OVI). An interesting example is when using the selling transactions for Firms corresponding to the CL_tmCL returns, where there is a significantly positive SR for the PHOTO data set, but a significantly negative SR when using the combined data set.

Most of the predictability comes either from the buying or selling transactions, and in almost no cases do they both appear to result in a significant SR. The one exception is the NOTO OVI of Market Makers for CL_tmOP returns where both $OVI_{i,d,M. \text{ Maker}}^{\text{Buy}, \text{NOTO}}$ and $OVI_{i,d,M. \text{ Maker}}^{\text{Sell}, \text{NOTO}}$ are negative, resulting in a statistically significant performance for $OVI_{i,d,M. \text{ Maker}}^{\text{NOTO}}$. Therefore, using only Buys or Sells outperforms the OVI obtained by using the complete data set in almost all cases.

Taking into account the plethora of different OVI features, we consider another model, in order to test whether these 45 features can be used to build a better signal (for overnight returns)

$$M3 : s_{i,d}^{(3)} = g \left(\beta_0 + \sum_{M \in \text{MPCs}} \sum_{S \in \{\text{Buy}, \text{Sell}, \text{Both}\}} \sum_{D \in \{\text{NOTO}, \text{PHOTO}, \text{Both}\}} \beta_{M,S,D} OVI_{i,d,M}^{S,D} \right), \quad (2.33)$$

where $OVI_{i,d,m}^{S,D}$ is the OVI corresponding to the i^{th} stock for the d^{th} day, for MPC M , data set D and type of transaction S of the data set (Buys, Sells or Both). In Figure 2.19, we plot the evolution of some of the coefficients (specifically the ones with the highest median magnitude across time). Note that M3 raises the overall PPD compared to M1-M2, with 2.48bp in-sample and 1.68 out-of-sample mean PPD, which signifies that the predictive power can be increased by considering the NOTO data set and decomposing the data set into Buys and Sells.

Looking at the significant coefficients, we draw the following conclusions

- $OVI_{i,d,M. \text{ Maker}}$ gives a consistently negative coefficient (similar to M2), and the same holds true for $OVI_{i,d,M. \text{ Maker}}^{\text{Sell}}$ and $OVI_{i,d,M. \text{ Maker}}^{\text{Sell}, \text{NOTO} \cup \text{PHOTO}}$.
- $OVI_{i,d, \text{ Broker}}^{\text{Sell}}$ starts with a significantly positive coefficient, which falls close to zero right after 2018, which is similar to the case of $OVI_{i,d, \text{ Broker}}$ for M2.

- $\text{OVI}_{i,d,\text{Customer}}^{\text{NOTO,Buy}}$ yields consistently positive coefficients, similar to $\text{OVI}_{i,d,\text{Customer}}$ for M2.
- $\text{OVI}_{i,d,\text{Customer}}^{\text{Buy}}$ is also consistently negative, but its coefficient shrinks for later years, similar to $\text{OVI}_{i,d,\text{Broker}}^{\text{Sell}}$.
- $\text{OVI}_{i,d,\text{Firm}}^{\text{NOTO}\cup\text{PHOTO,Buy}}$ has a fluctuating coefficient for the early years, but a positive one for the latter years, reflecting the pattern observed earlier in Figure 2.7.

2.7.5 Cross-Impact Analysis

We perform a cross-sectional analysis, where the OVIs of all the assets are tested for predictive power against the future returns of all the underlying equities. With N stocks in the universe, there are N^2 relationships to be tested. These interactions can be represented as an unweighted directed network, where the existence of an out-edge from vertex i to vertex j indicates that the OVI of asset i can be used as a predictor for the equity returns of asset j .

Our approach is similar to the one employing Bonferroni networks ([226]). Such a network was constructed by [73] to investigate the intraday return correlations across stocks. Instead of correlations, we employ the cumulative P&L obtained by following a simple strategy of using the Market Maker OVI of an asset i as a predictor for the overnight (CL_tmOP) returns of another asset j . An edge is added to the network if the SR differs significantly from 0, according to the [17] hypothesis test.

In particular, let G denote the adjacency matrix of the network. For each potential edge (i, j) , we iterate across days and trade equity $j = \{1, \dots, N\}$, for days where asset i 's OVI belongs to the Q3 portfolio (top 60% of the highest OVI's in magnitude). In that case, we buy asset j if the OVI of asset i is positive and sell it if it is negative. Let $Q(i, d, q) = 1$, if the i^{th} asset is in the q^{th} quantile group for day i and 0 otherwise. Then, the daily return for following this simple strategy is equal to $\text{PnL}_{i,j,d} = 1_{Q(i,d,q)} \text{sign}(s_{i,d}) \tilde{f}_{j,d}$. We perform the SR significance test on $\{\text{PnL}_{i,j,d}\}_{d=1}^D$, and set $G_{i,j} = 1$ if the obtained p-value is less than a predetermined value p (and $G_{i,j} = 0$ otherwise). By varying $q = 1, 2, \dots, 5$, we obtain five different networks for each threshold.

To account for multiple testing, we use the conservative method of a Bonferroni-corrected test by choosing a uniform significance level $\alpha_1 = p/(\text{Number of tests})$ for all individual tests, giving a bound for the family-wise error rate (FWER) $\leq p$ (see Appendix A of [73]). We use three different types of p-values corresponding to 0.05, $0.05/N$ (applying the correction to the row-level), and $0.05/N^2$ (applying a full

Bonferroni correction). Depending on the p-value and quantile group, we build a different network, resulting in 15 different networks.

For this analysis, we only use assets listed in CRSP for the whole duration from 01 January 2015-31 December 2019, leaving a reduced universe of 1792 NYSE assets. In Figure 2.20, we summarize the information regarding each of these networks. Obviously, as the p-value grows the network also grows, with the networks corresponding to a p-value of 0.05 having a density of 7%. For higher QR groups, the network is larger, with Q5 resulting in the largest networks. The fully Bonferroni-corrected networks have a total number of edges between 209 and 1207, which clearly rejects any hypothesis of no cross-impact effects.

Furthermore, out of the 1207 edges of the Q5 network, only two are *self-loops* (i.e. using an asset's own OVI feature to predict its future returns). This is higher than what would be expected under an Erdős-Rényi network with the same density (which is 0.67 here). *Bi-directed* edges (i.e. OVI of two different assets being good predictors of each other's future returns) are also quite rare, with only two such edges, which is, again, much higher than the expected value of 0.23.

Lastly, we test whether this network is structurally similar to any of the main types of networks commonly observed in practice. The degrees of many real-world networks follow a *power-law* distribution; to this end, we employ the test of [64] to test the null hypothesis that in/out-degrees follow a power-law. As shown in Figure 2.20, the network degrees do not appear to follow such a power-law distribution. In particular, five of the networks reject the null hypothesis that in-degrees follow a power-law distribution, while six of the networks reject the equivalent hypothesis for out-degrees.

Let $d_j^{\text{IN}} = \sum_{i=1}^N G_{i,j}$, and $d_j^{\text{OUT}} = \sum_{i=1}^N G_{j,i}$, denote the in- and out-degrees of node j , respectively. Furthermore, let

$$\hat{p}_j^{\text{IN}} = \frac{\sum_{l=1}^N G_{l,j}}{N}, \quad \text{and} \quad \hat{p}_j^{\text{OUT}} = \frac{\sum_{l=1}^N G_{j,l}}{N},$$

denote the empirical edge densities corresponding to node j . Another question is whether the differences between network degrees are statistically significant. In particular, assume that a node's in- and out-degrees follow a binomial distribution, i.e

$$d_j^{\text{IN}} \sim \text{Binomial}(N, p_j^{\text{IN}}) \quad \text{and} \quad d_j^{\text{OUT}} \sim \text{Binomial}(N, p_j^{\text{OUT}}), \quad \forall 1 \leq j \leq N. \quad (2.34)$$

Then, we can form the following hypotheses

$$\text{H0: } p_j^{\text{IN}} = p_i^{\text{IN}}, \quad \forall 1 \leq i, j \leq N, \quad \text{and} \quad \widetilde{\text{H0}}: p_j^{\text{OUT}} = p_i^{\text{OUT}}, \quad \forall 1 \leq i, j \leq N. \quad (2.35)$$

If the network is a simple Erdős-Rényi network, then both hypotheses should hold in conjunction. To test whether node i and node j degrees are similar, we use the conservative two-sample binomial t-test, with the test statistic

$$\text{Test}_{i,j} = \frac{\hat{p}_i - \hat{p}_j}{\sqrt{\frac{2\hat{p}(1-\hat{p})}{N}}}. \quad (2.36)$$

Under the null hypothesis, we have $\text{Test}_{i,j} \sim N(0, 1)$. To account for multiple testing, we once again use the Bonferroni correction by setting the p-value equal to $0.05/N^2$.

Note that such a test is very conservative, both due to the Bonferroni correction and the choice of a t-test for the pairwise difference tests. Indeed, for both in- and out-degrees, smaller networks (fully Bonferroni corrected) do not result in any rejections of the null hypothesis. This is different for the larger networks, as we see a 5-12% rejection rate for the row-level Bonferroni corrected networks and a 41-58% rejection rate for the uncorrected networks. Thus, the networks at hand seem to neither follow a power law nor form an Erdos-Renyi network, as there are significant differences between node degrees.

In Appendix A.4, we provide additional results based on the Q3 row-level Bonferroni corrected network. We show differences in the average degrees of different sectors and list some of the instruments with the highest centralities, demonstrating a more central role within the network.

2.8 Conclusion and Future Research Directions

In this paper, we have defined the OVI feature and shown how it can act as a predictor for future equity returns. Focusing on the PHLX exchange, we compared OVI across MPCs, and found that the Market Maker's OVI consistently provides the highest predictability, yielding annualized Sharpe Ratios of up to 4.5 (without taking into account transaction costs). In terms of PnL, the tail portfolios corresponding to the strongest signals can attain up to 4bpts per day. We have shown that some level of predictability is also present for Customer and Broker OVIs, while no predictability was concluded for Firm Proprietary trades and Professional Customers. We demonstrated how to improve performance by taking the OVI's magnitude into account. In particular, when using quantile rank groups, we found that the 2nd-4th quantile rank groups are typically the best performing.

We showed that most of the predictability is present in either Buys or Sells depending on the MPC and the data set used. Decomposing into opening or closing

transactions did not significantly affect the observed predictability. However, decomposing transactions based on certain option features significantly affects predictability. In particular, the OVI from high implied volatility contracts is significantly more informative than the OVI from option contracts with low implied volatility. Other option characteristics that seem to consistently induce higher performance include negative values for Rho and negative values for Delta (i.e put options). We have also shown that cross-impact effects are present and, in fact, significantly outnumber the cases of direct impact when applying a conservative Bonferroni correction. Results obtained for the NOM exchange (NOTO data set) are consistent with those obtained on the PHLX exchange (PHOTO data set), albeit weaker, with Market Makers still showing a significant annualized SR of around 2.5–3.

The obtained results strongly support previous claims that the options market is leading the underlying stock market. Our results point away from the explanation that the predictive power is a result of some form of price pressure. Higher volumes do not lead to higher performance, as indicated by using different volume weighting schemes (see appendix Appendix A.3), where we allocate higher capital for instruments with a higher option volume. This seems to indicate that the predictability is not due to positive (negative) price pressure from call (put) options. On the other hand, as shown in Appendix A.4, holding the positions open for longer horizons than one day does not significantly alter the P&L, indicating a longer-term return persistence.

Another contribution of this paper was the introduction of a new approach to tackling future return regression, namely the P&L regression, which was motivated by the needs of this particular research but may also be of independent interest to other financial studies. This approach, while slightly more intricate than a simple OLS method, is problem-tailored for directional predictions, bypasses the requirement for linear effects and is able to work with sparse data. We have, for the most part, avoided the usage of linear methods and focused on the usage of metrics like the SR, providing a general framework for analyzing return prediction tasks.

This research could lead to further studies that examine option volumes from different data sets and at different time horizons. This research could also be extended to an intraday analysis (given that data comes in batches of 10-minute bucket increments), which we consider in Appendix A.5. This led to promising observations which we have not been able to quantitatively approach, due to our limited access to intraday data. Another interesting angle would include a study of the interplay between option and stock volumes and ultimately the prediction of future volumes

(for both the option and stock markets), which would be of potential interest for market making and optimal execution.

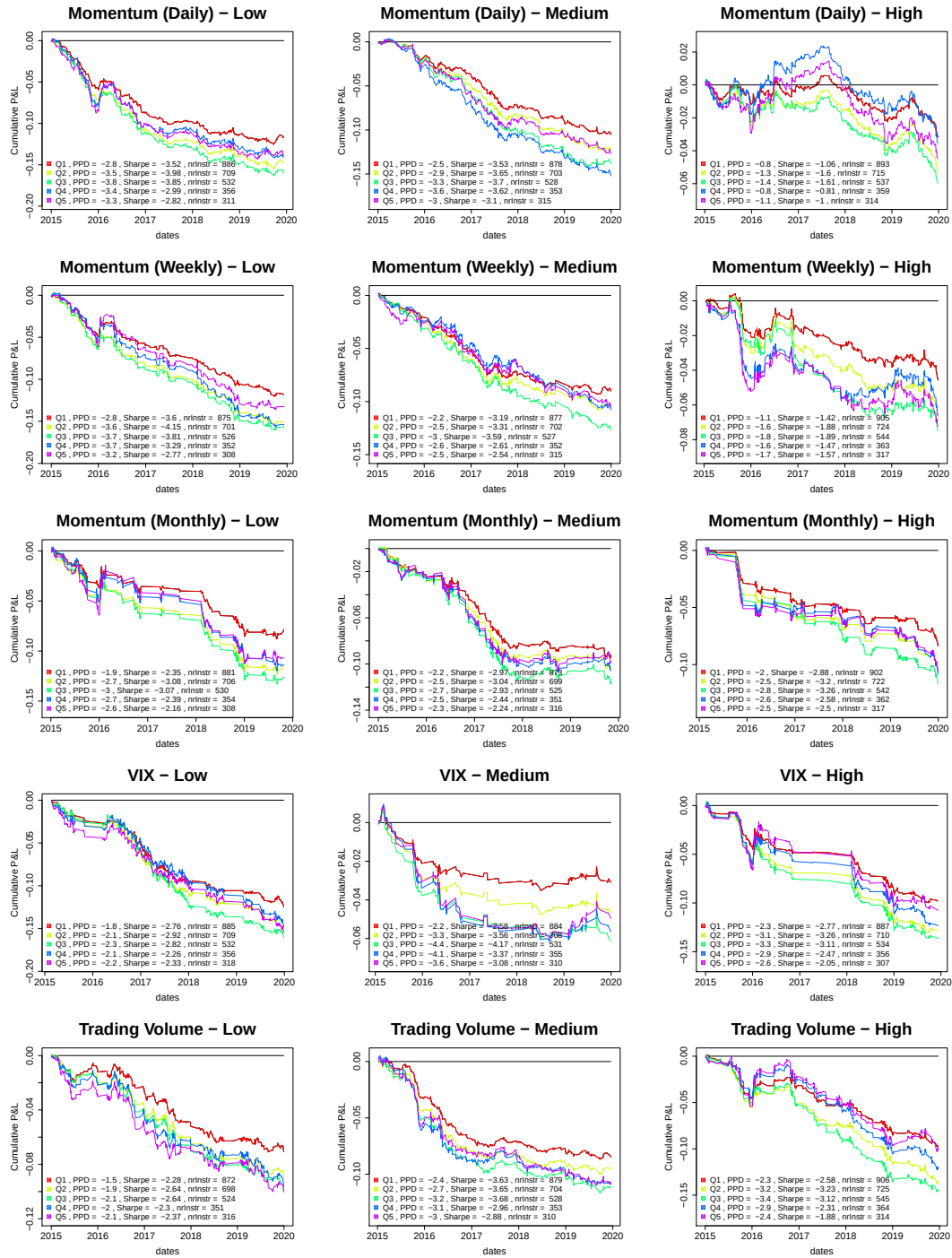


Figure 2.17: Cumulative P&L plot - Market Maker: This shows the cumulative P&L from following a one-day holding period strategy for the five quantile rank groups Q1-Q5, using the OVI of Market Makers for fifteen different market regimes, as explained in Section 2.7.3. Also given are SR (the Annualized Sharpe Ratio), PPD (Profit per dollar in basis points), and n (the average number of assets in the portfolios).

	NOTO									PHOTO									NOTO + PHOTO								
	CLCL			CLOP			OPCL			CLCL			CLOP			OPCL			CLCL			CLOP			OPCL		
	Buy	Sell	Both	Buy	Sell	Both	Buy	Sell	Both	Buy	Sell	Both	Buy	Sell	Both	Buy	Sell	Both	Buy	Sell	Both	Buy	Sell	Both	Buy	Sell	Both
	Firm									Brokers									Market Makers								
Q1	-0.12	0.70	0.07	-0.34	0.00	-0.43	-0.02	0.50	0.52	0.63	-1.72	-0.38	1.26	-1.58	-0.24	0.11	-0.87	-0.34	-0.53	2.01	0.34	-1.29	1.19	-0.35	-0.08	1.13	0.80
Q2	-0.10	0.52	0.31	-0.41	0.29	-0.41	-0.17	0.59	0.57	0.71	-1.98	-0.77	1.35	-1.22	0.17	-0.01	-0.76	0.00	-0.53	1.87	0.85	-1.36	1.17	-0.61	-0.19	1.17	0.67
Q3	-0.21	0.42	-0.08	-0.44	0.32	-0.42	0.01	0.88	0.53	0.52	-2.10	-1.12	1.31	-1.10	0.22	0.14	-0.74	-0.23	-0.54	1.78	0.58	-1.38	1.14	-0.67	-0.06	1.50	0.75
Q4	-0.21	0.41	0.03	-0.44	0.32	-0.42	0.02	0.82	0.66	0.48	-2.06	-1.06	1.31	-1.10	0.24	0.12	-0.71	-0.15	-0.52	1.73	0.65	-1.38	1.14	-0.68	-0.04	1.41	0.85
Q5	-0.21	0.41	0.03	-0.44	0.32	-0.42	0.02	0.82	0.66	0.48	-2.06	-1.06	1.31	-1.10	0.24	0.12	-0.71	-0.15	-0.52	1.73	0.65	-1.38	1.14	-0.68	-0.04	1.41	0.85
Brokers																											
Q1	0.13	0.33	0.02	-1.83	0.71	-1.13	-0.46	1.43	1.16	1.29	-2.47	-0.54	0.38	-2.97	-2.44	0.07	-0.79	-0.68	-0.58	1.82	0.34	-2.34	2.19	0.01	-0.56	2.00	1.75
Q2	0.00	0.34	0.06	-1.78	0.61	-1.34	-0.59	1.54	1.09	1.27	-2.58	-0.69	0.57	-2.97	-2.42	-0.04	-0.92	-0.85	-0.72	1.80	0.45	-2.38	2.00	-0.30	-0.65	2.18	1.77
Q3	0.01	0.23	0.08	-1.58	0.45	-1.56	-0.51	1.52	1.24	1.10	-2.41	-0.79	0.47	-2.77	-2.54	0.00	-0.72	-0.62	-0.58	1.55	0.50	-2.08	1.68	-0.53	-0.58	2.07	1.83
Q4	0.03	0.24	0.28	-1.57	0.41	-1.43	-0.40	1.50	0.88	1.01	-2.41	-0.65	0.39	-2.60	-2.29	0.07	-0.66	-0.44	-0.50	1.55	0.66	-2.00	1.55	-0.53	-0.50	2.05	1.28
Q5	0.03	0.24	0.29	-1.57	0.41	-1.34	-0.40	1.50	1.03	1.01	-2.41	-0.64	0.39	-2.60	-2.13	0.08	-0.65	-0.25	-0.50	1.55	0.65	-2.01	1.55	-0.51	-0.50	2.05	1.36
Market Makers																											
Q1	-0.82	-0.05	-0.93	-1.42	-0.78	-2.72	-0.86	1.24	0.27	0.46	-2.90	-1.60	0.64	-4.13	-4.02	0.13	-0.58	-0.39	-1.22	1.55	-0.13	-1.98	1.17	-0.91	-1.06	1.70	0.55
Q2	-0.83	-0.27	-1.14	-1.58	-0.96	-3.10	-0.96	1.33	0.34	0.61	-3.06	-1.81	0.54	-4.49	-4.20	0.15	-0.16	-0.15	-1.30	1.33	-0.27	-2.08	1.08	-1.24	-1.19	1.58	0.49
Q3	-0.93	-0.47	-1.47	-1.53	-0.82	-3.08	-1.12	1.18	0.38	0.57	-3.12	-2.18	0.25	-4.17	-4.42	0.04	-0.54	0.01	-1.41	1.12	-0.50	-1.85	1.04	-1.16	-1.32	1.60	0.45
Q4	-0.61	-0.58	-1.55	-1.31	-0.86	-2.46	-0.92	1.35	0.14	0.32	-3.13	-2.13	0.07	-3.93	-3.88	0.42	-0.19	0.12	-0.87	0.97	-0.62	-1.49	0.85	-0.83	-1.28	1.65	0.10
Q5	-0.52	-0.60	-0.93	-1.30	-0.86	-2.29	-1.02	1.18	0.23	0.36	-2.77	-1.59	0.07	-3.93	-3.55	0.37	-0.16	0.33	-0.79	0.76	-0.19	-1.48	0.85	-0.79	-1.38	1.44	0.08
Customers																											
Q1	-0.28	0.76	0.52	-0.11	1.81	2.01	-1.33	0.69	-0.83	3.06	-0.91	1.35	3.51	-0.65	3.20	0.69	-0.07	0.29	-1.92	1.39	-0.16	-1.69	2.41	0.67	-1.83	0.82	-1.13
Q2	-0.35	0.76	0.49	-0.05	1.86	2.01	-1.38	0.77	-0.86	3.12	-0.66	1.41	3.49	-0.64	3.46	0.84	-0.08	0.19	-2.02	1.25	-0.21	-1.58	2.43	0.58	-1.94	0.91	-1.12
Q3	-0.03	0.60	0.59	-0.01	1.63	2.02	-1.47	0.89	-0.80	3.18	-0.52	1.81	3.47	-0.64	3.40	0.52	0.07	0.25	-1.62	0.98	-0.27	-1.48	2.15	0.67	-1.90	0.98	-1.08
Q4	-0.04	0.39	0.32	-0.22	1.52	2.04	-1.45	0.58	-0.80	2.83	-0.67	1.58	2.98	-0.46	3.20	0.39	-0.16	0.06	-1.44	0.81	-0.46	-1.49	1.92	0.80	-1.82	0.74	-0.97
Q5	-0.04	0.39	0.14	-0.22	1.52	1.62	-1.45	0.58	-0.82	2.83	-0.67	1.37	2.98	-0.47	2.66	0.37	-0.13	-0.05	-1.44	0.81	-0.56	-1.49	1.92	0.59	-1.82	0.73	-0.93
Professional Customers																											
Q1	0.20	0.34	0.63	-0.31	1.17	0.55	-0.01	0.17	-0.19	0.55	-0.27	0.03	0.93	0.74	0.98	0.19	-0.08	-0.03	-0.12	0.65	0.74	-1.03	0.83	-0.08	-0.12	0.25	-0.22
Q2	-0.09	-0.08	0.12	-0.52	1.12	0.49	-0.27	0.05	0.00	0.63	-0.59	-0.27	0.89	0.97	1.35	0.22	-0.15	0.29	-0.51	0.34	0.32	-1.21	0.57	-0.35	-0.44	0.15	-0.17
Q3	0.05	-0.06	0.00	-0.61	1.04	0.36	-0.34	-0.05	-0.35	0.62	-0.64	-0.11	0.82	0.91	1.44	-0.15	-0.20	-0.05	-0.33	0.39	0.07	-1.25	0.53	-0.57	-0.33	0.06	-0.39
Q4	0.04	-0.19	-0.12	-0.62	1.04	0.33	-0.44	-0.04	-0.66	0.79	-0.78	-0.17	0.82	0.91	1.42	-0.06	-0.32	-0.26	-0.44	0.33	-0.03	-1.27	0.53	-0.59	-0.49	0.16	-0.64
Q5	-0.02	-0.25	-0.35	-0.62	1.04	0.33	-0.59	0.03	-0.60	0.69	-0.83	-0.48	0.82	0.91	1.42	-0.15	-0.30	-0.39	-0.44	0.30	-0.10	-1.27	0.53	-0.59	-0.62	0.23	-0.49

Figure 2.18: Overview of predictability across data sets: This table shows the SRs obtained for each OVI measure. The columns are grouped by the data set, where PHOTO corresponds to the PHLX and NOTO to the NOM option data set. At a lower level, the columns are grouped by the return type used for their evaluation, and further down by the type of transactions considered (Only Buys / Only Sells / Both). The rows are broken down at the level of MPC and then finally by the QR Group. Cells colored red (blue) when they are significantly negative (positive), at the Bonferroni Corrected level.

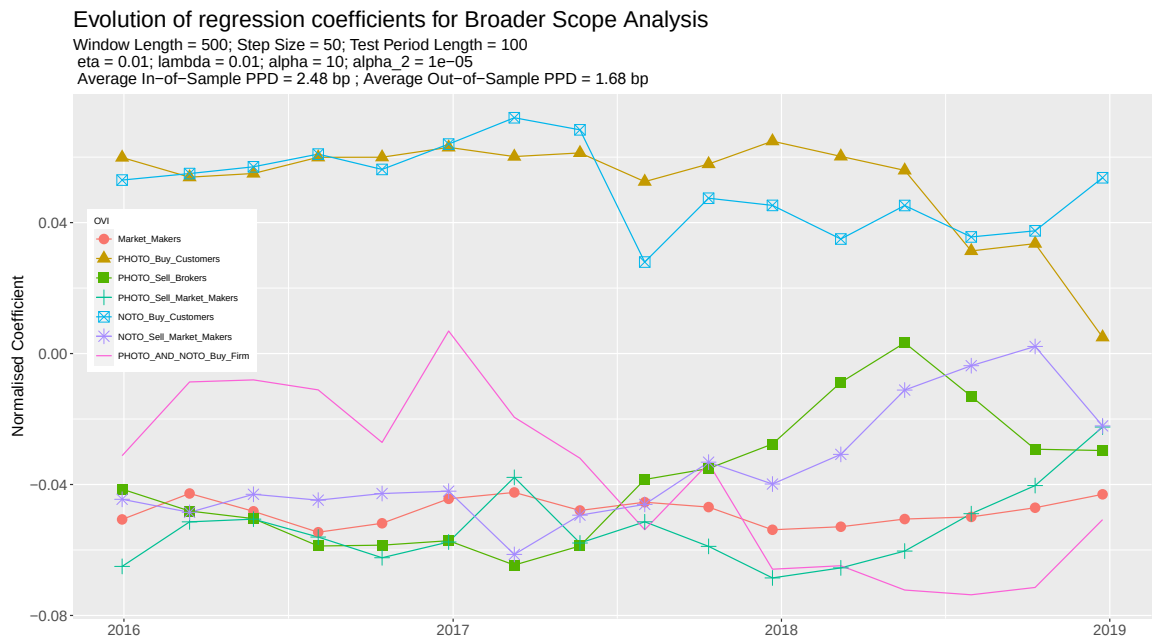


Figure 2.19: Multiple Factors Regression for Broader Scope Analysis: The line graph shows the evolution of the coefficients fitted for model (2.33) using the P&L regression for QR group 1, across the different windows. The points are shown at the midpoint of their corresponding window. Note that coefficients are normalized to have a total L1 norm equal to 1 (excluding the intercept term). This figure only shows the coefficients with the highest average absolute values, with most of the other factors not being consistently positive or negative.

Network Type Hypothesis								Motifs			
Power Law (p value)				ER (proportion of tests rejected)				Self-Loops		Bi-Directed Edges	
Number of Edges	Density	Average Degree	In Degrees	Out Degrees	In Degrees	Out Degrees		Observed	Expected	Observed	Expected
p = 1.6e-08: Fully Bonferonni Corrected											
Q1	209	0.00	0.12	0.21	0.72	0.00	0.00	0	0.12	0	0.01
Q2	343	0.00	0.19	0.55	0.00	0.00	0.00	0	0.19	0	0.02
Q3	415	0.00	0.23	0.03	0.00	0.00	0.00	1	0.23	0	0.03
Q4	782	0.00	0.44	0.01	0.00	0.00	0.02	1	0.44	2	0.10
Q5	1207	0.00	0.67	0.00	0.00	0.00	0.04	2	0.67	2	0.23
p = 2.8e-05: Row Level Bonferonni Corrected											
Q1	4898	0.00	2.73	0.01	0.06	0.08	0.05	1	2.73	8	3.73
Q2	5027	0.00	2.81	0.47	0.03	0.10	0.05	3	2.81	6	3.93
Q3	5486	0.00	3.06	0.00	0.13	0.11	0.06	3	3.06	8	4.68
Q4	6456	0.00	3.60	0.85	0.00	0.11	0.08	2	3.60	10	6.49
Q5	7471	0.00	4.17	0.26	0.00	0.10	0.12	5	4.17	20	8.69
p = 0.05: No Correction											
Q1	227277	0.07	126.83	0.99	0.85	0.41	0.58	149	126.83	18078	8038.27
Q2	222773	0.07	124.32	0.49	0.81	0.42	0.57	151	124.32	17328	7722.83
Q3	226820	0.07	126.57	0.95	0.49	0.43	0.57	158	126.57	18226	8005.98
Q4	229658	0.07	128.16	0.68	0.11	0.45	0.57	148	128.16	18896	8207.57
Q5	231413	0.07	129.14	0.12	0.21	0.45	0.57	157	129.14	19250	8333.49

Figure 2.20: Summary Information for Cross Impact Networks: The above table shows information on 15 cross impact networks based on three p-values (no Bonferroni correction, row-level Bonferroni Corrected and fully Bonferroni Corrected) as well as five Quantile Ranks. The summary information includes the in and out degrees, the parameter fitted for a power-law distribution and the p-value based on the null hypothesis that the network follows a power-law.

Chapter 3

OFTER: An Online Pipeline for Time Series Forecasting

3.1 Introduction

Time series are temporally structured data sets that play a central role in finance. For forecasting tasks, the temporal structure can be leveraged to improve performance but also to avoid look-ahead biases in the training phase. For univariate time series, parametric linear time series models such as the Autoregressive Integrated Moving Average (ARIMA) and the Autoregressive Conditional Heteroscedasticity (ARCH) have attained sufficiently good results in most applications of interest. While these algorithms can be extended to the setting of multivariate time series, the rapid rise in computing power in recent decades has rendered machine learning techniques an attractive alternative to traditional time series methods. An early comparison between different machine learning models was [5], where the authors concluded that the choice of method had a significant effect on performance.

Deep learning models have been a popular choice for multivariate time series forecasting, including Long-Short Term Memory (LSTM) [112], Bidirectional LSTM [70] and Gated Recursion Unit [62] (GRU). Various extensions of such deep learning models have been proposed in recent years, such as DeepAR [95] and Temporal Fusion Transformers [147], which generally achieve state-of-the-art performances across various domains. However, such approaches often employ thousands of trainable parameters, which can be difficult to train with a limited number of observations, thus leading to degraded out-of-sample forecasting performance. Despite such advancements in deep learning, simple regression-based models are still widely used and sought after. This is especially true for mid-low frequency financial tasks, such as

predicting daily returns, volumes, and realized volatilities, partly due to the complexity, unpredictability and constantly shifting dynamics of the financial markets. For example, Ordinary Least Squares (OLS) is still widely used for both forecasting and interpretive statistical analysis due to its simplicity, closed-form solution, and the fact that it requires no choice of hyper-parameters. Its regularized versions, such as the LASSO and the Elastic Net are some of the most widely utilized regression models in finance, as they require only a few hyperparameters.

On the other hand, various recent lines of work have demonstrated the merit of combining traditional time series models and machine learning models, such as [75, 92, 188, 185, 250]. In this paper, we consider one of the simplest types of hybrid models, which builds on the results of a univariate ARIMA model and enhances it by performing a novel regression method on the residuals of an ARIMA model.

Non-parametric models and dimensionality reduction Two simple models that have been relatively overlooked by the machine learning community, are k-Nearest neighbors (kNN) and Generalized Regression Neural Networks (GRNN), both of which are simple, non-parametric, fast methods requiring only a single hyperparameter. This mainly stems from the curse of dimensionality, which affects both techniques. GRNNs have been utilized primarily in engineering applications, such as [256] for logo recognition, [8] for forecasting feeding patterns, and [186] for modeling the pile-bearing capacity of soil. An overview of its different applications for dynamical systems has been pursued by [7]. In this work, we aim to circumvent the curse of dimensionality that these two methods suffer from by using the following two techniques. Firstly, instead of the Euclidean norm employed by kNN and GRNN, we consider a thresholded weighted norm that is driven by the feature’s predictive potential. Secondly, before employing these well-known models, we consider a dimensionality reduction component, which is by far the most common remedy for the curse of dimensionality.

Dimensionality reduction is one of the main unsupervised machine learning paradigms, with the past couple of decades having witnessed seminal contributions in terms of non-linear dimensionality reduction methods, such as locally linear embedding (LLE) [120], ISOMAP [222], Hessian LLE [80], Laplacian eigenmaps [27], and diffusion maps [65]. In more recent years, deep learning models have also been increasingly used as a means of dimensionality reduction, such as the class of autoencoders [136], which can be seen as a non-linear extension of PCA. Performing dimensionality reduction via

the use of autoencoders before employing a kNN technique, [195] has shown a significant improvement in performance over kNN for classification. Partial Least Squares is a popular method of incorporating information from the target into the initial features, during the embedding stage. More recently, [116] achieved improvements in the performance of PCA by scaling each factor based on its predictive slope with the target.

Uses in financial applications Examples of dimension reductionality use cases in time series forecasting tasks include [57], who used a truncated diffusion map to construct a global manifold for the slow dynamics of a reduced kinetic model, and [189] who utilizes manifold learning techniques in conjunction with various supervised learning models to build a profitable strategy for the Forex market. Along the same lines, [241] proposed a hybrid approach integrating the Midas-RV model and PCA to achieve better results in forecasting the volatility of crude oil futures in the Chinese market. In this paper, we consider the forecasting of daily returns and traded volumes for ETF instruments. Being able to accurately predict next-day volumes is of interest to practitioners and especially to market makers. Previous works include [126], who showed how search engine data can be leveraged to forecast future trading volumes. Similarly to our work is [151], who utilizes Support Vector Machines to predict the daily volume percentages with promising performance, in the spirit of [146] who combine SVM and LSTM after residualizing with an AR model.

Main contribution OFTER - a novel integrated pipeline We consider the task of one-step time series forecasting and propose an approach that integrates ARIMA models, dimensionality reduction, and simple non-parametric methods into a single pipeline, which we denote as *Online Forecasting via Temporal Embedding Regression (OFTER)*. Such a pipeline is of great interest to practitioners working with mid-sized data sets that are low signal-to-error in nature, with no significant seasonality patterns, with low frequency trading strategies being a common example that we explore.

To alleviate the curse of high dimensionality and enhance our dimensionality reduction component, we conduct feature selection on the embedding features based on the concept of *maximal correlation*. We apply OFTER to the financial tasks of forecasting next-day volume and returns by building an embedding based on various features of 10 Exchange Traded Funds (ETFs). Through an extensive set of numerical simulations on both synthetic and real data, we demonstrate that OFTER provides

a powerful alternative to other mainstream machine learning models mid-sized data sets, even performing better than established time series-tailored deep learning models. In particular, it can work effectively in low signal-to-noise regimes for mid-sized data sets, while still being computationally efficient.

Furthermore, our proposed methodology provides the following additional benefits.

- Once an embedding has been computed, the algorithm can easily perform out-of-sample extensions and incorporate newly arrived time observations. Therefore, OFTER can naturally be viewed as an online algorithm that operates efficiently in an online setting.
- Certain configurations of our method fall under the unsupervised learning setting, as the embedding is independent of the target. In other words, the same embedding can potentially be employed to forecast different target variables of interest, and thus OFTER is amenable to extensions to the multi-target setting.
- Our method can further provide an avenue for change-point detection and anomaly detection tasks, which can be seen as natural byproducts. A previous or newly arrived observation without any local neighboring points in the embedding, could indicate either an extreme measurement or a change-point in the time series.

Paper Structure The rest of this paper is structured as follows. In Section 3.2, we provide an overview of some of the essential concepts used in the proposed pipeline, including the kNN and GRNN algorithms, principal component analysis, and maximal correlation. In Section 3.3, we introduce our proposed methodology, along with its main components and their integration into an algorithm. In Section 5.5, we analyze the performance of OFTER in both synthetic experiments and real-world data. We also discuss the additional properties of OFTER in Section 3.5. Finally, we conclude in Section 5.6.

3.2 Preliminaries and Related Works

Consider a d -dimensional time series represented as a matrix $X = (x_1, \dots, x_T) \in \mathbb{R}^{T \times d}$, with rows denoting observations for T time points, which we refer to as the feature matrix. For the remainder of this paper, assume that we are interested in

forecasting a single univariate target variable, denoted by $y = (y_1, \dots, y_T) \in \mathbb{R}^T$, which we refer to as the target variable. Furthermore, let $x' \in \mathbb{R}^d$ and $y' \in \mathbb{R}$ represent the values of the features and the target variable, respectively, associated with a new observation. We denote by e_d the all-ones vector of length d . In terms of matrix indexing, we generally denote the (i, j) entry of X by $X_{i,j}$, the submatrix of X including rows from t_1 to t_2 by $X_{t_1:t_2}$ (i.e., $(x_{t_1}, \dots, x_{t_2})$), and the j^{th} column of $X_{t_1:t_2}$ by $X_{t_1:t_2,j}$.

3.2.1 K-nearest Neighbors and Generalized Regression Neural Network

K-Nearest Neighbors (kNN) is a non-parametric supervised learning algorithm that dates back to the work of Fix and Hodges in 1951 [94] and has grown to become one of the most popular machine learning techniques. The main idea is that a similarity of values in the feature space implies a similarity in the values for the target variable. Denote by $x_{(1)}, \dots, x_{(T)}$ a reordering of the time series entries, such that

$$\|x_{(1)} - x'\| \leq \dots \leq \|x_{(T)} - x'\|, \quad (3.1)$$

where $\|\cdot\|$ denotes the Euclidean norm. The k-nearest neighbors set for the new entry x is then defined as

$$\text{NN}_k(x') = \{i \in \{1, \dots, T\} : \|x_{(i)} - x'\| \leq \|x_{(k)} - x'\|\}. \quad (3.2)$$

Based on $\text{NN}_k(x')$, the forecast for y' is computed as

$$\hat{y} = \frac{1}{k} \sum_{i \in \text{NN}_k(x')} y_i. \quad (3.3)$$

A closely related method is the Generalized Regression Neural Network (GRNN), introduced in [219], based on the Nadaraya-Watson kernel regression estimator [179, 230]. When forecasting y' , GRNN attaches a weight to each of the previous time series entries, determined by their distance to x' . In particular, define the weight w_i of the i^{th} observation as

$$w_i = \frac{\exp\left(\frac{-\|x_i - x'\|^2}{h}\right)}{\sum_{j=1}^T \exp\left(\frac{-\|x_j - x'\|^2}{h}\right)}, \quad (3.4)$$

where h is a bandwidth hyper-parameter to be optimized. The prediction for the response of a new observation $\hat{y}$ is then simply the weighted average of the responses

of the training observations, defined as

$$\hat{y} = \sum_{i=1}^T w_i y_i. \quad (3.5)$$

Consider the observations $(x_1, y_1), \dots, (x_T, y_T)$, to be realizations of an $\mathbb{R}^{d+1}$ random vector, and let f denote the corresponding true underlying density of the random variable. The theoretical motivation of GRNN stems from the Parzen estimators [219], used to estimate unknown probability density functions f as

$$\hat{f}_h(x', y') = \frac{1}{T} \frac{1}{(\pi h)^{\frac{d+1}{2}}} \sum_{i=1}^T \exp \left[-\frac{(x' - x_i)^T (x' - x_i)}{h} \right] \exp \left[-\frac{(y' - y_i)^2}{h} \right]. \quad (3.6)$$

If h varies with T such that $\lim_{T \rightarrow \infty} h(T) = 0$ and $\lim_{T \rightarrow \infty} T h^d(T) = 0$, these estimators are shown to be consistent, i.e

$$\text{plim}_{T \rightarrow \infty} \hat{f}_{h(T)}(x', y') = f(x', y'), \quad \forall (x', y') \in \mathbb{R}^{d+1}. \quad (3.7)$$

Informally, [219] argues that if $\hat{f}(x, y) \approx f(x, y)$, then

$$\hat{y} = \sum_{i=1}^T w_i y_i = \int y' \frac{\hat{f}(x, y')}{\int \hat{f}(x, y) dy} dy' \approx \int y' \frac{f(x, y')}{\int f(x, y) dy} dy' = \mathbb{E}[y'|x']. \quad (3.8)$$

We caution the reader that, despite GRNN bearing “Neural Network” in its name, it is quite different from other popular variants. Specifically, aside from the width hyper-parameter, there are no ad-hoc hyperparameters to be chosen or optimized. Furthermore, it is a single-pass algorithm, as the weights are not optimized. Thus, unlike other neural network structures, it is much faster and easier to train but lacks the expressive power of neural networks.

3.2.2 Dimensionality Reduction

Both kNN and GRNN suffer from the so-called *curse of dimensionality* which refers to the phenomenon that for certain methods, the amount of data points required to estimate an unknown function with a certain accuracy, grows exponentially with the dimensionality of the data set.¹ To alleviate this issue, kNN or GRNN is often

¹For a common illustration of this problem, consider points that lie uniformly on a d -dimensional hypercube of side length $(1 + \epsilon)r$. In this setup, only a proportion of $(1 + \epsilon)^{-d}$ data points lie in the central inner hypercube of side length r , with the rest of the points lying in the outer shell. This demonstrates that the distance between data points becomes unsuitable for a meaningful description of the structure of high-dimensional data sets.

applied after projecting the data points from $\mathbb{R}^d$ into a lower-dimensional space $\mathbb{R}^p$ ($p \leq d$), while somewhat retaining the structure of the original data set. By far the most popular such method is Principal Component Analysis (PCA), which applies a linear transformation to the original features, chosen to maximize the variance of the projected features in the newly reduced space while ensuring the resulting features remain uncorrelated with each other.

Denote by $C(X)$ the covariance matrix of X , and its eigen-decomposition by $C(X) = U\Lambda U^T$, where Λ is a diagonal matrix capturing the $\lambda_1 \geq \dots \geq \lambda_d \geq 0$ eigenvalues of $C(X)$, and the columns of U are given by the $u_1, \dots, u_d$ orthonormal eigenvectors (in the same order as their corresponding eigenvalues). The principal component scores are given by the columns of XU , with the variance of each of the new components given by its corresponding eigenvalue (component i has variance λ_i). Selecting the first p components retains a proportion of the total variance in the feature matrix equal to $\sum_{i=1}^p \lambda_i / \sum_{i=1}^d \lambda_i$. Given the proportion of variance to retain $\delta \in (0, 1)$, we set

$$p = \min \left\{ i : \frac{\sum_{k=1}^i \lambda_k}{\sum_{k=1}^d \lambda_k} \geq \delta \right\}. \quad (3.9)$$

We remark that random-matrix theory based techniques could also be employed for delimiting the informative part of the spectrum from the bulk of the eigenvalues that correspond to the noise component [39]. Note that before calculating $C(X)$, we ensure that X is of full rank by removing any columns of X that result in rank deficiency.² Before applying PCA, we standardize X (subtracting by each feature its mean and dividing by its standard deviation). We shall denote the transformed features as $\tilde{X} = XU$.

3.2.3 Maximal Correlation

Consider two univariate random variables denoted as V_1 and V_2 . Loosely stated, the correlation between V_1 and V_2 is a measure of their association. Popular types of correlation include the Pearson, Spearman and Kendall correlations. Another type is the maximal correlation (see e.g [41]) defined as

$$\operatorname{argmax}_{\phi, \psi \in \mathcal{F}} \operatorname{Cor}(\phi(V_1), \psi(V_2)), \quad (3.10)$$

²Using a QR-decomposition algorithm, we can express $X = QR$, where Q is an orthogonal matrix and R is an upper-triangular matrix. Then we remove any feature i , where $|R_{i,i}| < \epsilon$ for some small $\epsilon > 0$.

where $\mathcal{F}$ is a suitable function class and Cor is the Pearson correlation defined as

$$\text{Cor}(V_1, V_2) = \frac{\mathbb{E}[V_1 V_2] - \mathbb{E}[V_1] \mathbb{E}[V_2]}{\sqrt{\mathbb{E}[V_1^2] - \mathbb{E}[V_1]^2} \sqrt{\mathbb{E}[V_2^2] - \mathbb{E}[V_2]^2}}. \quad (3.11)$$

A solution to the problem of determining maximal correlation can be computed by optimizing the functions ϕ and ψ iteratively. We consider a related but simpler problem, where we aim to calculate

$$\arg\max_{\phi \in \mathcal{F}} \text{Cor}(\phi(V_1), V_2). \quad (3.12)$$

The reason for considering Equation (3.12) instead of Equation (3.10), is that in a regression setting, we are only interested in transforming the feature, rather than the target. By solving Equation (3.10), one can transform the features in the embedding space so as to be maximally correlated to the target.

We consider a solution in a restricted function class $\mathcal{F}$.³ In particular, we consider the class of Bernstein polynomials (see e.g [154]), which are defined as

$$b_{\nu,n}(x) = \binom{\nu}{n} x^\nu (1-x)^{n-\nu}, \quad \nu = 0, 1, \dots, n. \quad (3.13)$$

The choice of this basis is justified via the Stone-Weistrass theorem.⁴ Bernstein polynomials have been widely used, especially in the context of Bezier curves [113], and B-splines [194], often arising in the development of modeling tools in computer-aided geometric design.

For the solution of Equation (3.12), we assume that we observe vectors v_1 and v_2 , which capture N observations of the random variables V_1 and V_2 , respectively. Denote by $\bar{v}_1$ and $\bar{v}_2$ the mean of v_1 and v_2 respectively. Similarly denote by σ_{v_1} and σ_{v_2} standard deviation of v_1 and v_2 respectively. Consider the function basis $(\phi_1, \dots, \phi_K)$, where $\phi_i(x) = b_{K,i}(x)$ and define the $N \times K$ matrix Φ s.t $\Phi_{i,j} = \phi_j(v_{1,i})$. We assume Φ is centered, i.e $\Phi^T e_N = \mathbf{0}$.

With the function basis defined, we consider the function class

$$\mathcal{F} = \left\{ \sum_{k=1}^K c_k \phi_k : c_1, \dots, c_K \in \mathbb{R} \right\}, \quad (3.14)$$

³If $\mathcal{F}$ is chosen to be the class of all functions, then the solution to the above problem can easily be shown to have the trivial solution $\phi(V_1) = \frac{V_2 - \mathbb{E}[V_2]}{\text{Var}(V_2)}$.

⁴In its most usual form it says that for any real continuous function f on the interval $[a, b]$ and $\epsilon > 0$, $\exists$ a polynomial function g such that $\forall x \in [a, b]$, we have $|f(x) - g(x)| < \epsilon$.

so that for $f \in \mathcal{F}$, $f(v_{1,i}) = \sum_{k=1}^K c_k \phi_k(v_{1,i})$, and problem Equation (3.12) reduces to finding the suitable vector c .

To simplify calculations, we also require the transformed vector to be standardized, i.e. $c^T \Phi^T e_N = 0$, $\frac{1}{N-1} c^T \Phi^T \Phi c = 1$. Thus, we aim to solve

$$\underset{c \in \mathbb{R}^K, \ c^T \Phi^T e_N = 0, \ c^T \Phi^T \Phi c = N-1}{\operatorname{argmax}} \operatorname{Cor}(\Phi c, v_2). \quad (3.15)$$

Note that

$$\operatorname{Cor}(\Phi c, v_2) = \frac{\operatorname{Cov}(\Phi c, v_2)}{\sigma_{v_2} \sigma_{v_1}} = \frac{\frac{(\Phi c)^T v_2}{N-1} - \left[\frac{(\Phi c)^T e_N}{N} \right] \left[\frac{v_2^T e_N}{N} \right]}{\sigma_{v_2} \sigma_{v_1}}, \quad (3.16)$$

and thus, under the constraints of Equation (3.15), we have

$$\operatorname{Cor}(\Phi c, v_2) = \frac{(\Phi c)^T v_2}{(N-1) \sigma_{v_2}}, \quad (3.17)$$

for which a solution can be found by considering the Lagrangian

$$L = \frac{c^T \Phi^T v_2}{(N-1) \sigma_{v_2}} - \lambda (c^T \Phi^T e_N) - \mu (c^T \Phi^T \Phi c - (N-1)). \quad (3.18)$$

Extremizers of the above Lagrangian satisfy

$$\mu = \frac{c^T \Phi^T v_2}{2(N-1)^2 \sigma_{v_2}}, \text{ and} \quad (3.19)$$

$$\lambda = \frac{\bar{v}_2}{(N-1) \sigma_{v_2}}. \quad (3.20)$$

Substituting back to Equation (3.18), and writing $z = \Phi^T v_2$, $\bar{z} = z - \bar{v}_2 \Phi^T e_N$ and $A = \frac{\Phi^T \Phi}{N-1}$, we reduce the problem to solving

$$\bar{z} = (c^T z) A c. \quad (3.21)$$

By considering a solution of the form $c = \tau A^{-1} \bar{z}$ for $\tau \in \mathbb{R}$, we arrive at the solution

$$\tau = \frac{1}{\sqrt{\bar{z}^T (A^{-1})^T \bar{z}}}. \quad (3.22)$$

Overall, the solution to Equation (3.15) is given by

$$c = \frac{A^{-1} \bar{z}}{\sqrt{\bar{z}^T (A^{-1})^T \bar{z}}} = \frac{(\Phi^T \Phi)^{-1} \Phi^T (v_2 - \bar{v}_2 e_N)}{\sqrt{(v_2 - \bar{v}_2 e_N)^T \Phi \bar{z}^T [(\Phi^T \Phi)^{-1}]^T \Phi^T v_2}}. \quad (3.23)$$

We refer to the correlation between the target y and the transformed x as the *one-sided maximal correlation* (OSMC) of x and y , as

$$\operatorname{OSMC}_K(x, y) = \operatorname{Cor}(\Phi c, y), \quad (3.24)$$

where Φ and c are depended on x and the choice of the number of basis functions used K .

3.3 Methodology

Figure 3.1 is a pictorial diagram of the main components of the OFTER pipeline, each of which is described in detail in the remainder of this section.

3.3.1 Online Update of the Embedding

In order for the algorithm to run in an online setting, the embedding needs to be updated as new data arrives. The initial embedding is calculated based on the first $L_0 < T$ entries, and then updated at every time step. However, re-computing the embedding at every single iteration can be quite expensive (in fact, of complexity $O(nd^2 + d^3)$ at each step⁵). For PCA, there exists a computationally efficient approach that updates U and Λ . In particular, if $C(X_t)$ denotes the covariance matrix built from the first t data points, one can compute the eigenvectors of the updated covariance

$$C(X_{1:t+1}) = \frac{t-1}{t}C(X_{1:t}) + \frac{1}{t}\|x_{t+1}\|^2 \frac{x_{t+1}}{\|x_{t+1}\|} \frac{x_{t+1}^T}{\|x_{t+1}\|}, \quad (3.25)$$

by using the previously computed eigenvectors of $C(X_{1:t})$ (we assume centered data for now). The online algorithm proposed in [171] achieves just that, with a complexity of $O(p^2d)$ at each iteration (recall that p denotes the chosen dimension of the embedding).

The above is an instance of a more general problem, which is the rank-one update of an eigensystem [45]. Consider again the symmetric matrix A , whose eigendecomposition is known, with $\lambda_1 \geq \dots \geq \lambda_d$ denoting its eigenvalues, and u_i be the eigenvector corresponding to $\lambda_i, \forall i \in \{1, \dots, d\}$ such that $A = U\Lambda U^T$. The task at hand is to evaluate the eigensystem of $A + \rho vv^T$, by using the known eigensystem of A . As shown by [45], the eigenvalues of $A + \rho vv^T$ are the roots $\kappa_1, \dots, \kappa_d$ of the *secular equation*

$$1 + \rho \sum_{i=1}^d \frac{z_i^2}{\lambda_i - \kappa} = 0, \text{ where } z = U^T v. \quad (3.26)$$

Once the updated eigenvalues are computed, each updated eigenvector u_k can be calculated as

$$u_k = \frac{U (\Lambda - \kappa_k I)^{-1} z}{\|U (\Lambda - \kappa_k I)^{-1} z\|}. \quad (3.27)$$

⁵This is due to PCA consisting of a covariance estimation step with complexity $O(nd^2)$ and a spectral decomposition step of complexity $O(d^3)$.

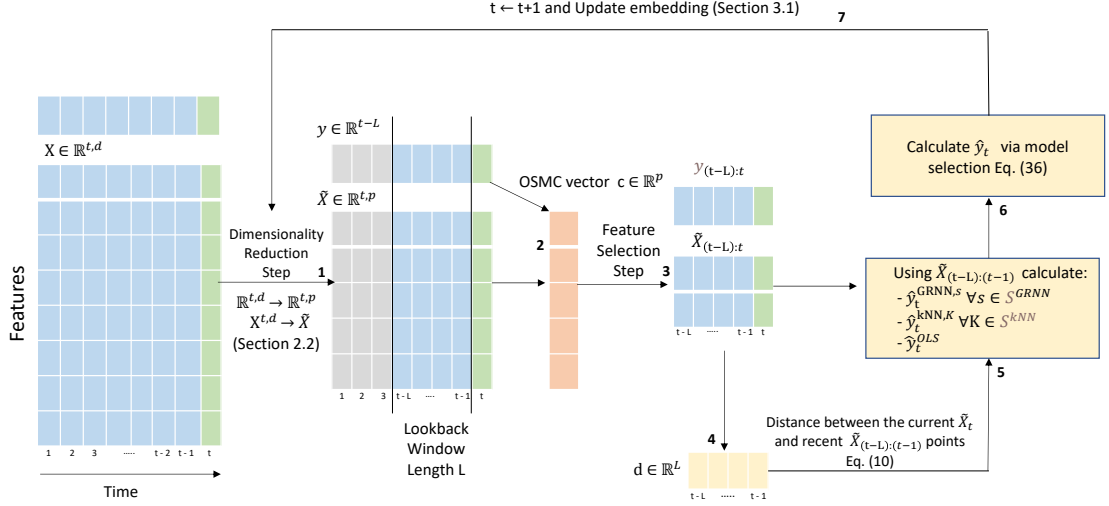


Figure 3.1: Pictorial illustration of OFTER, where the main components of its iterative part are displayed. Note that some components are omitted to simplify the diagram.

(1) In this step the feature matrix $X \in \mathbb{R}^{n,d}$ is transformed to $\tilde{X} \in \mathbb{R}^{n,p}$ ($p < d$) using Principal Component Analysis.

(2) Each entry of the vector c is calculated as $c_j = \text{OSMC}(\tilde{X}_{1:L_0,j}, y_{1:L_0})$, with OSMC as formulated in Section 3.2.3.

(3) Each entry of the vector v is calculated as $v_j = \frac{c_j^2 1(|c_j| \geq c_{\min})}{p + |I^{\text{Original}}| \sum_{l=1}^p c_l^2 1(|c_l| \geq c_{\min})}$, as is

explained in Section 3.3.3. In other words, features whose correlation with the target is low are discarded.

(4) The distance between observations of time points j and t is calculated as $d_{j,t} = \|\tilde{X}_t - \tilde{X}_j\|_v$, where $\|\cdot\|_v$ is later defined in Equation (3.44). This is calculated for $j = t - L, \dots, t - 1$ where L is the lookback window.

(5) All the different forecasts from GRNN, kNN and OLS are calculated using $\tilde{X}_{(t-L):(t-1)}$, and the distance vector d .

(6) All the individual forecasts are combined into $\hat{y}_t$. In particular, the forecast of the highest performing model is selected as explained in Section 3.3.2.

(7) The PCA embedding is updated using the methodology described in Section 3.3.1. In particular, the eigenvalues and eigenvectors of the covariance matrix is updated.

Going a step further, one could reduce computations by only updating the first few eigenvalues. A good approximation of the solution is given by [170], who shows how this can be achieved by solving

$$1 + \rho \left(\sum_{i=1}^p \frac{z_i^2}{\lambda_i - \kappa} + \frac{1 - \sum_{i=1}^p z_i^2}{\mu - \kappa} \right) = 0. \quad (3.28)$$

The solutions of this equation are shown to approximate the first m eigenvalues, with an error of the order $O(\max_{m+1 \leq i \leq p} |\lambda_i - \mu|)$. Note that μ is another parameter to be chosen, and should be made so as to minimize $O(\max_{m+1 \leq i \leq p} |\lambda_i - \mu|)$, for example $\mu = \text{median}(\lambda_{m+1}, \dots, \lambda_p)$. The parameter p will have already been chosen at an earlier part of the algorithm using Equation (3.9), and does not change during the update step. While there are various existing techniques to solve equation Equation (3.28), we employ the convexifying method described in [166], adopting the earlier work of [45].

Note that initially the feature matrix is standardized based on the initial L_0 entries (i.e we standardize $X_{(1:L_0)}$). Denote the mean vector of $X_{1:t}$ by $\mu_t \in \mathbb{R}^d$, and the correctly centered $X_{1:t}$ by $\bar{X}_{1:t} \in \mathbb{R}^d$. Assuming that $X_{(1:t-1)}$ has been centered by μ_{t-1} , we need to re-center based on μ_t . Following [171], the update takes the form $\bar{X}_{1:t} = X_{1:t} + e_t(\mu_{t-1} - \mu_t)^T$. It turns out that we can adjust for the re-centering, via two more rank-one updates of the covariance matrix, by expressing $\bar{X}_{1:t}^T \bar{X}_{1:t}$ in the following way.

$$\begin{aligned} \bar{X}_{1:t}^T \bar{X}_{1:t} &= X_{1:t}^T X_{1:t} + X_{1:t}^T e_N (\mu_{t-1} - \mu_t)^T + (\mu_{t-1} - \mu_t) e_N^T + t(\mu_{t-1} - \mu_t)(\mu_{t-1} - \mu_t)^T \\ &= X_{1:t}^T X_{1:t} + [\mu_{t-1} - \mu_t \quad X_{1:t}^T e_N] \begin{bmatrix} t & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} (\mu_{t-1} - \mu_t)^T \\ e_N^T X_{1:t} \end{bmatrix} \\ &= X_{1:t}^T X_{1:t} + [\mu_{t-1} - \mu_t \quad X_{1:t}^T e_N] \begin{bmatrix} \frac{\rho_1}{\sqrt{1+\rho_1^2}} & \frac{\rho_2}{\sqrt{1+\rho_2^2}} \\ \frac{1}{\sqrt{1+\rho_1^2}} & \frac{1}{\sqrt{1+\rho_2^2}} \end{bmatrix} \\ &\quad \times \begin{bmatrix} \rho_1 & 0 \\ 0 & \rho_2 \end{bmatrix} \begin{bmatrix} \frac{\rho_1}{\sqrt{1+\rho_1^2}} & \frac{\rho_2}{\sqrt{1+\rho_2^2}} \\ \frac{1}{\sqrt{1+\rho_1^2}} & \frac{1}{\sqrt{1+\rho_2^2}} \end{bmatrix}^T \begin{bmatrix} (\mu_{t-1} - \mu_t)^T \\ e_N^T X_{1:t} \end{bmatrix} \\ &= X_{1:t}^T X_{1:t} + \begin{bmatrix} \frac{\rho_1(\mu_{t-1} - \mu_t) + X_{1:t}^T e_N}{\sqrt{1+\rho_1^2}} & \frac{\rho_2(\mu_{t-1} - \mu_t) + X_{1:t}^T e_N}{\sqrt{1+\rho_2^2}} \end{bmatrix} \begin{bmatrix} \rho_1 & 0 \\ 0 & \rho_2 \end{bmatrix} \begin{bmatrix} \frac{\rho_1(\mu_{t-1} - \mu_t) + X_{1:t}^T e_N}{\sqrt{1+\rho_1^2}} \\ \frac{\rho_2(\mu_{t-1} - \mu_t) + X_{1:t}^T e_N}{\sqrt{1+\rho_2^2}} \end{bmatrix} \\ &= X_{1:t}^T X_{1:t} + \rho_1 b_1 b_1^T + \rho_2 b_2 b_2^T, \end{aligned} \quad (3.29)$$

where in the above

$$\rho_1 = \frac{t - \sqrt{t^2 + 4}}{2}, \quad (3.30)$$

$$\rho_2 = \frac{t + \sqrt{t^2 + 4}}{2}, \quad (3.31)$$

$$b_1 = \frac{\rho_1(\mu_{t-1} - \mu_t) + X_{1:t}^T e_N}{\sqrt{1 + \rho_1^2}}, \quad (3.32)$$

$$b_2 = \frac{\rho_2(\mu_{t-1} - \mu_t) + X_{1:t}^T e_N}{\sqrt{1 + \rho_2^2}}. \quad (3.33)$$

3.3.2 Choice of Hyperparameters

The hyperparameter k of kNN in Equation (3.5) and the smoothing parameter hyperparameter h in Equation (3.4) are both of crucial importance to the performance of their respective algorithms. Notice that, as $h \rightarrow \infty$, $w_i \rightarrow 1/T$ (Equation (3.4)) for all i . On the other hand, as $h \rightarrow 0$, GRNN becomes equivalent to 1NN (kNN with $k = 1$). To demonstrate the effect of the smoothing parameter on the predictive task, we consider the following toy model

$$y_t = 0.5 + \cos\left(\frac{\pi}{64}t\right) + \epsilon_t, \text{ where } \epsilon_t \sim N(0, \sigma^2). \quad (3.34)$$

In Figure 3.2 we demonstrate how the choice of the h affects the nature of the obtained fit. For low values of h , the obtained fit is close to giving constant fitted values equal to the mean, while for high values of h the obtained fit matches the observed data points almost perfectly.⁶

To choose a suitable value of h , [219] proposes the use of the holdout method, where the optimal parameter is found via calculating a loss function by removing a sample point at a time and taking the average. More modern applications use K-fold validation instead, and these methods can be modified for time series data (see [29] for a discussion on the different approaches).

To remove any scale dependencies, we re-parameterize h as

$$h = \frac{\text{median}(\{|x_j - x'|\}_{j=1}^T)}{s}. \quad (3.35)$$

Dynamic Selection Since this is an online algorithm, we could update the hyperparameters in an online manner as well. Instead of choosing a single value from a candidate set, we use each of the candidate values for the hyperparameter to fit a different model, updating the performance of all models at each time point. The model with the highest performance up to the previous time point is chosen for the forecast of the current time point, as is done in dynamic model selection [176, 217].

⁶Note that this is a pattern for this specific toy data set and not a general pattern.

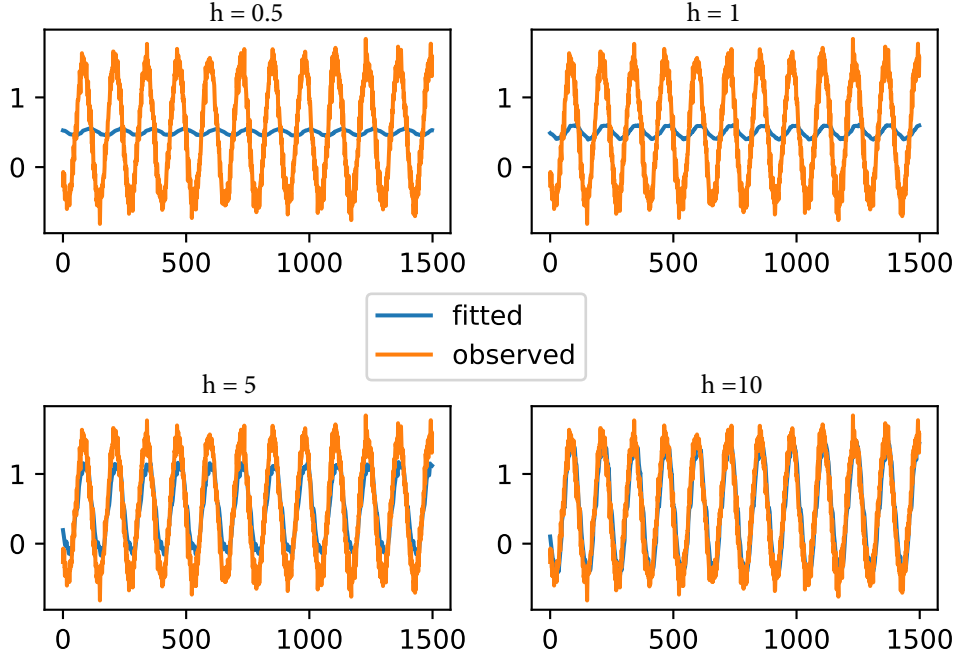


Figure 3.2: Visualization of the forecasted values when using GRNN with different smoothing parameters for the model Equation (3.34). In a high signal-to-noise problem like this, as h becomes larger GRNN becomes almost perfectly capable of forecasting future values.

The region of competence includes the last L_D time points (we refer to this as the lookback window).⁷

The choice of the parameter s can be performed by the use of a loss function. If the fitted value corresponding to parameter s for day t is $\hat{y}_t^s$, the forecast $\hat{y}_t$ is calculated as

$$\hat{y}_t = \sum_{s=1}^{|S|} \eta_t^s \hat{y}_t^s, \text{ where } \eta_t^s = \begin{cases} 1, & R_t^s = \min(R_t^1, \dots, R_t^S) \\ 0, & \text{otherwise} \end{cases}. \quad (3.36)$$

By R_t^s , we define the loss corresponding to parameter s , i.e $R_t^s = \sum_{i=1}^{t-1} \text{Loss}(\hat{y}_i^s, y_i)$. The loss function can be the Mean Squared Error (MSE) or Mean Absolute Error (MAE) for general applications. For the application of forecasting future returns, we use the loss function of minus P&L (we define P&L in Section 3.4.3).

Similarly, a large k for kNN algorithm would produce similar forecasts to using low values of h . Instead of using either GRNN or kNN, we use both methods, each

⁷Alternatively, we can take a weighted average of the forecasts of each model. However, while not reported here for brevity, the performance is inferior.

with their own hyperparameter candidate set, and then select based on performance, in a similar manner to Equation (3.36). Aside from these two regression methods, we also consider a standard OLS component.

Denote by S^{GRNN} the hyperparameter candidate for GRNN and S^{kNN} the hyperparameter candidate set for kNN. Denote by $\hat{y}_t^{\text{GRNN},s}$, the forecast when using GRNN with parameter s , $\hat{y}_t^{\text{kNN},k}$ the forecast when using kNN (with parameter k), and $\hat{y}_t^{\text{OLS}}$ the forecast when using OLS. Then define

$$\begin{aligned} R_t^{\text{GRNN},s} &= \sum_{i=t-L_D}^{t-1} \text{Loss}(\hat{y}_i^{\text{GRNN},s}, y_i). \\ R_t^{\text{kNN},k} &= \sum_{i=t-L_D}^{t-1} \text{Loss}(\hat{y}_i^{\text{kNN},k}, y_i). \\ R_t^{\text{OLS}} &= \sum_{i=t-L_D}^{t-1} \text{Loss}(\hat{y}_i^{\text{OLS}}, y_i). \end{aligned} \quad (3.37)$$

Based on the losses up to time point t , the minimum loss is calculated as

$$R_t^{\text{optimal}} = \min(\{R_t^{\text{GRNN},s} : s \in S^{\text{GRNN}}\} \cup \{R_t^{\text{kNN},k} : k \in S^{\text{kNN}}\} \cup \{R_t^{\text{OLS}}\}), \quad (3.38)$$

and the number of models achieving the optimal performance as

$$N_t^{\text{optimal}} = |\{s : R_t^{\text{GRNN},s} = R_t^{\text{optimal}}\}| + |\{k : R_t^{\text{kNN},k} = R_t^{\text{optimal}}\}| + 1(R_t^{\text{OLS}} = R_t^{\text{optimal}}), \quad (3.39)$$

where $1(A)$ is an indicator function taking the value 1, if condition A is satisfied and 0 otherwise. Then we reformulate Equation (3.36), to include the kNN and OLS models as

$$\hat{y}_t = \sum_{s \in S^{\text{GRNN}}} \eta_t^{\text{GRNN},s} \hat{y}_t^{\text{GRNN},s} + \sum_{k \in S^{\text{kNN}}} \eta_t^{\text{kNN},k} \hat{y}_t^{\text{kNN},k} + \eta_t^{\text{OLS}} \hat{y}_t^{\text{OLS}}, \quad (3.40)$$

where

$$\begin{aligned} \eta_t^{\text{GRNN},s} &= \begin{cases} \frac{1}{N_t^{\text{optimal}}}, & R_t^{\text{GRNN},s} = R_t^{\text{optimal}} \\ 0, & \text{otherwise} \end{cases}, \\ \eta_t^{\text{kNN},k} &= \begin{cases} \frac{1}{N_t^{\text{optimal}}}, & R_t^{\text{kNN},k} = R_t^{\text{optimal}} \\ 0, & \text{otherwise} \end{cases}, \\ \eta_t^{\text{OLS}} &= \begin{cases} \frac{1}{N_t^{\text{optimal}}}, & R_t^{\text{OLS}} = R_t^{\text{optimal}} \\ 0, & \text{otherwise} \end{cases}. \end{aligned} \quad (3.41)$$

3.3.3 Overcoming the Shortcomings of kNN and GRNN

One downside of applying dimensionality reduction methods is that they are agnostic of the target to be forecasted. In transforming the initial space in order to retain a large proportion of the variability with a reduced number of features, an initial feature that is already highly correlated with the target but also weakly correlated with other features could end up having a relatively small weight in the final forecasting step. To circumvent this shortcoming, we employ two additional steps.

Firstly, we make sure that all features from the original feature space X that have a high enough correlation with the target are retained in the forecasting step. This way, we ensure that selecting the components that capture most of the variability is not done at the expense of dropping original features that provide predictive power on their own. Therefore, we define the following set, after calculating the OSMC between each of the original features and the target

$$I^{\text{Original}} = \{j : \text{OSMC}(X_{1:L_0,j}, y_{1:L_0}) \geq c_{\text{Original}}\}, \quad (3.42)$$

where any feature whose OSMC is higher than the pre-determined c_{Original} , is chosen to be included in the feature set.

During the pre-processing part of the algorithm, we replace $\tilde{X}$ with $(\tilde{X}, X_{1:T, I^{\text{Original}}})$, with a number of features equal to $p + |I^{\text{Original}}|$. Note that for mid-sized, low signal-to-error data sets like the daily data set we use in Section 3.4.3, I_{Original} is usually null.

Secondly, certain features in the embedding space ought to have more influence on the neighbor selection. This can be achieved by using a weighted metric rather than the standard Euclidean norm. In particular, if v represents the vector of weights, the weighted Euclidean distance between X_{t_1} and X_{t_2} is given by

$$\|\tilde{X}_{t_1} - \tilde{X}_{t_2}\|_v = \sqrt{p + |I^{\text{Original}}| \sum_{j=1} v_j (\tilde{X}_{t_1,j} - \tilde{X}_{t_2,j})^2}. \quad (3.43)$$

To increase robustness we threshold any correlations with absolute values smaller than $c_{\min}$ (which is another hyper parameter) to 0. The weights are based on the OSMC between each feature and the target

$$v_j = \frac{c_j^2 1(|c_j| \geq c_{\min})}{p + |I^{\text{Original}}| \sum_{l=1} c_l^2 1(|c_l| \geq c_{\min})}, \quad (3.44)$$

where $c_j = \text{OSMC}(\tilde{X}_{1:L_0,j}, y_{1:L_0}) \quad \forall j \in \{1, \dots, p, p+1, \dots, |I^{\text{Original}}|\}$.

3.3.4 Algorithm Outline

To ensure stationarity of the target y , we first perform an ADF test with significance level p_{ADF} , and if the null hypothesis is not rejected, we difference the time series, and repeat until we reject the null hypothesis. We then perform ACF and PACF tests to check if there are any significant ARMA effects.⁸ If such effects are identified, we perform a grid search over the orders of the AR and MA orders (p and q respectively) to find the optimal ARIMA(p,r,q) model in terms of their Bayesian Information Criterion (BIC), where r denotes the number of times differencing was applied. Otherwise, we simply fit ARIMA(0, r ,0).

Using this procedure, we decompose the target y into a time series signal (ARIMA) and a residual component

$$y = \underbrace{y^{\text{TS}}}_{\text{Time Series Signal}} + \underbrace{\epsilon}_{\text{Residual}}. \quad (3.45)$$

Overall, the algorithm consists of three main steps: (1) computing an embedding, (2) choosing features based on the OSMC, and (3) forecasting based on kNN and GRNN. The complete algorithm is given in Algorithm 1.

3.4 Empirical Results

We consider and compare four different configurations of OFTER, which we denote as follows

- OFTER (no dimensionality reduction and no feature transformation). This entails setting $\tilde{X} = X$, using the Pearson correlation instead of the OSMC, and skipping steps 4,5, 6, 7 and 17 of Algorithm 1.
- OFTER-DR (dimensionality reduction and no feature transformation). This is almost the same as Algorithm 1, but we use the Pearson correlation instead of the OSMC.
- OFTER-FT (feature transformation and no dimensionality reduction). This entails setting $\tilde{X} = X$ as well as skipping steps 4,5, 6, 7 and 17 of Algorithm 1.

⁸We use a significance level of 0.001. We consider a specific lag effect significant if the confidence interval of either the ACF or the PACF is above or below zero. It is important to note that in some cases, the ACF and PACF values can fail to identify the need of an ARIMA model. However, including this criterion can make the algorithm significantly faster, when forecasting targets like price returns.

Algorithm 1: OFTER

Input: X (Feature Matrix), y (target), L_0 (Training Window Length), L (Lookback Window Length), L_D (Dynamic Selection Lookback Window), $c_{\min}$, c_{Original} (correlation thresholds), S^{kNN} , S^{GRNN} (parameter candidate sets), δ (proportion of variance to retain), p_{ADF} (p-value for ADF), LOSS (loss function)

Output: $\hat{y}$

- 1: Use the procedure described in Section 3.3.4, to arrive at the optimal ARIMA(p, r, q) model for y
 - 2: Use the optimal model to obtain fitted values y^{TS} . If ARIMA(0,0,0) is chosen set $y^{\text{TS}} = (0, \dots, 0)$
 - 3: Transform X so as to make $X_{1:L_0}$ a full column-rank (by QR decomposition) and standardized matrix
 - 4: Obtain the first p principal components of $X_{1:L_0}$, with p chosen via Equation (3.9), and set $U_p = (u_1, \dots, u_p)$ and $\Lambda_p = (\lambda_1, \dots, \lambda_p)$
 - 5: Obtain the $L_0 \times p$ matrix $\tilde{X} = XU_{1:L_0, 1:p}$
 - 6: Set $I^{\text{Original}} = \{i : \text{OSMC}(X_{1:L_0, i}, y_{1:L_0}) \geq c_{\text{Original}}\}$
 - 7: Replace $\tilde{X}$ by $(\tilde{X}, X_{1:T, I^{\text{Original}}})$
 - 8: Compute $R_{L_0}^{\text{GRNN}, s} \forall s \in S^{\text{GRNN}}$, $R_{L_0}^{\text{kNN}, k} \forall k \in S^{\text{kNN}}$, and $R_{L_0}^{\text{OLS}}$ using Equation (3.37)
 - 9: Compute $\eta_{L_0}^{\text{GRNN}, s} \forall s \in S^{\text{GRNN}}$, $\eta_{L_0}^{\text{kNN}, k} \forall k \in S^{\text{kNN}}$, and $\eta_{L_0}^{\text{OLS}}$ using Equation (3.41)
 - 10: **for** $j \leftarrow 1$ to $p + |I^{\text{Original}}|$ **do**
 - 11: Compute $c_j = \text{OSMC}(\tilde{X}_{1:L_0, j}, y_{1:L_0})$
 - 12: Set $v_j = \frac{c_j^2 1(|c_j| \geq c_{\min})}{\sum_{l=1}^{p+|I^{\text{Original}}|} c_l^2 1(|c_l| \geq c_{\min})}$
 - 13: **end for**
 - 14: Regress $y_{1:L_0}$ on $\tilde{X}_{1:L_0}$, using OLS regression, to obtain the coefficient vector β and the intercept β_0
 - 15: **for** $t \leftarrow L_0 + 1$ to D **do**
 - 16: Compute $\hat{y}_t$ using Algorithm 2
 - 17: Update U_p and Λ_p , using the secular equation method (detailed in Section 3.3.1)
 - 18: If t is divisible by 100, re-fit the OLS model using $\tilde{X}_{(t-L):t}$ and $y_{(t-L):t}$, to obtain β and β_0
 - 19: **end for**
 - 20: Update $\hat{y}$ to $\hat{y} + y_{(L_0+1):D}^{\text{TS}}$
-

Algorithm 2: Forecasting for time point t

Input: $\tilde{X}$ (Feature Matrix), y (Target), L (Lookback Window Length), v (distance vector), S^{kNN} , S^{GRNN} (parameter candidate sets)

Output: $\hat{y}_t$

- 1: Compute $d_{j,t} = \|\tilde{X}_t - \tilde{X}_j\|_v$ for $j = t - L, \dots, t - 1$
 - 2: **for** $s \in S^{\text{GRNN}}$ **do**
 - 3: Set $h_s = \frac{\text{median}(\{d_{j,t}\}_{j=t-L}^{t-1})}{s}$
 - 4: Compute $w_{j,t}^s = \frac{\exp\left(\frac{-d_{j,t}^2}{h_s}\right)}{\sum_{m=t-L}^{t-1} \exp\left(\frac{-d_{m,t}^2}{h_s}\right)} \forall j \in \{t - L, \dots, t - 1\}$
 - 5: Compute $\hat{y}^{\text{GRNN},s} = \sum_{j=t-L}^{t-1} w_{j,t}^s y_j$
 - 6: **end for**
 - 7: **for** $k \in S^{\text{kNN}}$ **do**
 - 8: Find the set $\text{NN}_k(\tilde{X}_t)$ of the k -nearest neighbors of $\tilde{X}_t$, in $\tilde{X}_{(t-L):(t-1)}$
 - 9: Compute $\hat{y}^{\text{kNN},k} = \frac{1}{k} \sum_{i \in \text{NN}_k(\tilde{X}_t)} y_i$
 - 10: **end for**
 - 11: Compute $\hat{y}^{\text{OLS}} = \beta_0 + \beta^T \tilde{X}_t$
 - 12: Estimate the fitted value $\hat{y}_t$ using Equation (3.40)
-

- OFTER-DR-FT (feature transformation and dimensionality reduction). This is the exact algorithm as given by Algorithm 1.

Unless stated otherwise we use $p_{\text{ADF}} = 0.05$. When performing dimensionality reduction, we consider $\delta = 0.9$. For the candidate sets we use $S^{\text{kNN}} = \{1, 2, 3, 5, 10, 15, 20, 30, 50\}$ and $S^{\text{GRNN}} = \{0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1, 5, 10, 50, 100\}$. In terms of features, we include those corresponding to the current period up to a lag of 3. Lastly, we set L_0 equal to $\lfloor 0.7T \rfloor$. Note that OFTER is generally robust for different choices of hyper-parameters, as evidenced by our robustness analysis reported in Appendix B.3.

3.4.1 Baseline Methods

We compare the performance of OFTER against that of several mainstream machine learning methods. We include the linear methods OLS, PCR, LASSO, but also the widely used Support Vector Machines, Random Forests and XGBoost. We also include some of the most widely used time series tailored deep learning methods, such as LSTM, Bi-LSTM, and GRU. Unlike OFTER, most of these methods are not naturally amenable to efficient implementation in an online setting and thus cannot be updated. For comparison purposes, we retrain each of these models every 100

steps. In the case of LSTM, Bi-LSTM, and GRU, we re-fit the model with only five epochs, starting from the previous weight matrix. For the sake of a fair comparison, instead of following the embedding updating step, we also re-compute the embedding of OFTER every 100 steps. We provide an overview of these methods, as well as details of their implementation for this task, in Appendix B.2. Note that the features are standardized for all baseline methods in the same manner as they are for OFTER.

3.4.2 Synthetic Data Sets

Before testing OFTER on real life data sets, we first check whether it can perform consistently well on some simple low dimensional toy models. For this purpose, the three five-dimensional models used by [189], which are simplistic, that are crafted to resemble electroencephalography signals, which is suitable as a data set for this forecasting task.

$$\text{M1: } \begin{cases} y_t^{(1)} = 0.2y_{t-1}^{(1)} - 0.4y_{t-1}^{(2)} + \epsilon_t^{(1)} \\ y_t^{(2)} = -0.5y_{t-1}^{(1)} + 0.15y_{t-1}^{(2)} + \epsilon_t^{(2)} \\ y_t^{(3)} = -0.14y_{t-1}^{(2)} + \epsilon_t^{(3)} \\ y_t^{(4)} = 0.5y_{t-1}^{(1)} - 0.25y_{t-1}^{(2)} + \epsilon_t^{(4)} \\ y_t^{(5)} = 0.15y_{t-1}^{(1)} + \epsilon_t^{(5)} \end{cases} \quad (3.46)$$

The first data set (M1) shown in Equation (3.46) is generated by a set of five linear equations.

$$\text{M2: } \begin{cases} y_t^{(1)} = 3.4y_{t-1}^{(1)} \left(1 - \exp \left(- \left(y_{t-1}^{(1)} \right)^2 \right) \right) \exp \left(-y_{t-1}^{(1)2} \right) + \epsilon_t^{(1)} \\ y_t^{(2)} = 3.4y_{t-1}^{(2)} \left(1 - \exp \left(- \left(y_{t-1}^{(2)} \right)^2 \right) \right) \exp \left(-y_{t-1}^{(2)2} \right) + 0.5y_{t-1}^{(1)}y_{t-1}^{(2)} + \epsilon_t^{(2)} \\ y_t^{(3)} = 3.4y_{t-1}^{(3)} \left(1 - \exp \left(- \left(y_{t-1}^{(3)} \right)^2 \right) \right) \exp \left(-y_{t-1}^{(3)2} \right) + 0.3y_{t-1}^{(2)} + 0.5y_{t-1}^{(1)2} + \epsilon_t^{(3)} \\ y_t^{(4)} = 0.5y_{t-1}^{(1)} - 0.25y_{t-1}^{(2)} + \epsilon_t^{(4)} \\ y_t^{(5)} = 0.15y_{t-1}^{(1)} + \epsilon_t^{(5)} \end{cases} \quad (3.47)$$

The second data set (M2) shown in Equation (3.47) is generated by a set of five equations, some of which are non-linear. [189] based these equations on a selection of models appearing in [181].

$$\text{M3: } \begin{cases} y_t^{(1)} = 0.1y_{t-1}^{(1)} - 0.6y_{t-3}^{(2)} + \epsilon_t^{(1)} \\ y_t^{(2)} = -0.15y_{t-3}^{(1)} + 0.8y_{t-3}^{(2)} + \epsilon_t^{(2)} \\ y_t^{(3)} = -0.45y_{t-3}^{(2)} + \epsilon_t^{(3)} \\ y_t^{(4)} = 0.45y_{t-3}^{(1)} - 0.85y_{t-3}^{(2)} + \epsilon_t^{(4)} \\ y_t^{(5)} = 0.95y_{t-2}^{(1)} + \epsilon_t^{(5)} \end{cases} \quad (3.48)$$

The third data set (M3) resembles M1, but is a time series with a higher order, making M3 a VAR model.

In each of the above cases $\epsilon_t^{(1)}, \dots, \epsilon_t^{(5)}$ are uncorrelated normally distributed random variables with unit variance. Note that a burn-in period of 100 periods is used with $y_1^{(1)}, \dots, y_1^{(5)}$ set equal to zero.

3.4.3 Data

ETF data set

We consider a daily data set for two forecasting applications, namely daily returns and volumes. We consider 10 ETFs (SPY, XLF, XLB, XLK, XLV, XLI, XLU, XLY, XLP, XLE). We forecast both daily returns and changes in daily volumes, which we define in the following paragraphs.

Daily Returns: Let $P_{i,d}$ denote the closing price for asset i at day d . We consider the close to close returns⁹ for asset i , defined as

$$R_{i,d}^k = \frac{P_{i,d+k} - P_{i,d}}{P_{i,d}}. \quad (3.49)$$

As a more representative quantity, grounded within financial applications, we consider the market excess returns, that is the difference between the returns of an instrument and that of the S&P-500 index, a widely used proxy for market returns.

$$\tilde{R}_{i,d}^k = R_{i,d}^k - R_{\text{SPY},d}^k. \quad (3.50)$$

Predicting daily returns is one of the most popular yet difficult problems in finance. Due to the low signal-to-noise ratio of this task, standard metrics like Mean Squared Error and correlation, are considered unsuitable. A more suitable economic-driven evaluation method is following a hypothetical transaction of the underlying instruments, based on the forecasted signs (with $+$ and $-$ corresponding to buy and sell respectively). For each instruments i and each day t , we denote the forecast as

⁹The returns considered are adjusted for splits and dividends.

$s_{i,t}$ and execute a hypothetical trade based on the sign of $s_{i,t}$. As this is simply for evaluation purposes, we ignore transaction costs and liquidity constraints.

Following this hypothetical strategy, we consider the (excess-returns-adjusted) Profit and Loss (P&L) metric defined by

$$\text{P\&L}_d^k = \frac{1}{N_d} \sum_{i=1}^{N_d} \tilde{R}_{i,t}^k \times \text{sign}(s_{i,t}), \quad (3.51)$$

The P&L is defined above for each day d , which yields a vector $\mathbf{P\&L} = (\text{P\&L}_1, \text{P\&L}_2, \dots, \text{P\&L}_{T-1})$ of daily P&L's, which measures the outcome of the hypothetical strategy over the entire time horizon under consideration. A measure of the long-term performance of a strategy is the annualized Sharpe Ratio (hereafter SR) [213], defined as

$$\text{SR} = \frac{\text{mean}(\mathbf{P\&L}) \times \sqrt{252}}{\text{sd}(\mathbf{P\&L})}. \quad (3.52)$$

Lastly, we define the P&L per dollar traded (PPD) as

$$\text{PPD} = \text{mean}(\mathbf{P\&L}), \quad (3.53)$$

which assumes that every day, the total allocated capital to the portfolio remains constant at 1 USD.

We employ the SR significance test by [17]. As returns are known to exhibit heavier tails than a normal distribution, this test is an extension of a traditional t -test, adjusted to account for the skewness and kurtosis of the distribution of the returns.

A higher magnitude forecast should naturally signify a higher confidence. For each trading day, we consider the quintiles of the forecast magnitudes $\{|s_{i,t}|\}_{i=1}^N$, in order to define five quintile portfolios, of decreasing size. In particular, we define Q1 to be the portfolio consisting of all assets, Q2 to be the portfolio consisting of the top 80% assets with the highest magnitude, and so on, up to Q5 which consists of the 20% highest magnitude assets.

As naive benchmarks, we follow two naive strategies. The first we call Buy_SPY, which includes buying and holding the SPY ETF. The other, we call Buy_all, which includes buying and holding all of the ETFs for the duration of the test sets.

Daily changes in trading volume: Denote the volume of instrument i as $V_{i,d}$, which is defined as the total number of shares traded during day d . Note that this

quantity is non-stationary and generally increases over time, across most instruments. To this end, we instead consider the 'log-volume-returns', defined as

$$\delta_{i,d}^k = \frac{\log(V_{i,d+k}) - \log(V_{i,d})}{\log(V_{i,d})}. \quad (3.54)$$

Features: The embedding built is common for the two forecasting tasks and is built using a feature matrix consisting of the following

- 1, 2 and 3-day returns for each of the instruments, i.e $R_{i,d}^1, R_{i,d}^2, R_{i,d}^3$
- 1, 2 and 3-day absolute returns for each of the instruments, i.e $|R_{i,d}^1|, |R_{i,d}^2|, |R_{i,d}^3|$
- 1, 2 and 3-day log-volume returns for each of the instruments, i.e $\delta_{i,d}^1, \delta_{i,d}^2, \delta_{i,d}^3$
- the three Fama & French daily market factors
- the daily returns of 49 different industry portfolios¹⁰
- daily returns of the VIX index

3.4.4 Evaluation Metrics

To evaluate the forecasting models in this section we make use of the following evaluation metrics. Denote the vector of predicted values $\hat{y}$ and the observed vector y .

Root Mean Squared Error (RMSE):

$$\text{RMSE} = \sqrt{\frac{1}{T} \sum_{t=1}^T (y_t - \hat{y}_t)^2}. \quad (3.55)$$

Mean Absolute Error (MAE):

$$\text{MAE} = \frac{1}{T} \sum_{t=1}^T |y_t - \hat{y}_t|. \quad (3.56)$$

¹⁰The industry portfolio returns and the daily market factors are taken directly from the Kenneth R. French library [1]. Based on a market portfolio, the three factors include small minus big (SMB) based on firm size, high minus low (HML) based on the book-to-market ratio, and the market excess return of the portfolio.

Spearman Correlation: The Spearman Correlation between $\hat{y}_t$ and y is the Pearson correlation between the ranks of the two vectors. This allows this metric to capture non-linear monotonic relations. The Pearson correlation between $\hat{y}$ and y is given by

$$C = \frac{\sum_{t=1}^T (y_t - \bar{y})(\hat{y}_t - \bar{\hat{y}})}{\sqrt{\sum_{t=1}^T (y_t - \bar{y})^2 (\hat{y}_t - \bar{\hat{y}})^2}}. \quad (3.57)$$

In addition, we use the **Diebold-Mariano (DM) test** to compare two different sets of forecasts, which we denote by $\{\hat{y}_t\}_{t=1}^T$ (for the model considered) and $\{n_t\}_{t=1}^T$ (for the naive forecasts). When y corresponds to a change (such as the daily return, or change in volumes) the naive forecast is set to $n_t = 0$. Otherwise, the naive forecast is set to $n_t = y_{t-1}$. The DM test is a popular variant of the paired t-test introduced in [76]. When using loss function g , the difference in the loss of the two forecasts at time point t is given by $d_t = g(n_t - y_t) - g(\hat{y}_t - y_t)$. The one-tailed version of DM that we use tests whether the considered forecasts $\{\hat{y}_t\}_{t=1}^T$ outperform the naive forecasts $\{n_t\}_{t=1}^T$. In particular it tests the null hypothesis

$$H_0 : \mathbb{E}(d_t) = 0 \quad (3.58)$$

versus the alternative hypothesis

$$H_1 : \mathbb{E}(d_t) \geq 0. \quad (3.59)$$

In our case we use the squared-error loss, i.e $g(e) = e^2$, and a significance level of 5%.

Lastly, we use an instance of the Theil-U statistic [202]. As pointed out by [35], it is recommended to use the second version of this statistic, which is defined as

$$U = \sqrt{\frac{\sum_t (y_t - \hat{y}_t)^2}{\sum_t (y_t - n_t)^2}}. \quad (3.60)$$

3.4.5 Numerical Results

We begin with a discussion of the running time of OFTER. In Table 3.1, we provide the average running time of the algorithm for each of the three considered data sets. We notice that all four configurations of OFTER are significantly faster than LSTM, Bi-LSTM and GRU, while at the same time significantly slower than the OLS, LASSO, PCR and SVM, as expected. It is also clear that among the OFTER configurations, OFTER-FT is the slowest as each feature is transformed separately without first reducing the dimensionality of the feature space. The time complexity

Data Set	T	d	LM	PCR	LASSO	Random Forests	SVM	LSTM	Bi-LSTM	GRU	OFTER	OFTER-DR	OFTER-FT	OFTER-DR-FT
Synthetic	3000	15	0.01	0.02	0.64	33.90	1.81	173.45	202.99	184.81	7.87	9.68	37.52	24.53
Financial	5022	154	0.37	0.48	5.38	3,144.02	0.17	442.86	503.78	549.30	17.58	22.25	222.87	153.06
Air Quality	1461	228	0.10	0.31	2.23	132.05	1.00	165.48	152.22	184.85	3.25	20.18	157.95	62.64

Table 3.1: For each forecasted quantity in each of the three multivariate time series the running time for each of the 12 mentioned methods is recorded. The average running time (in seconds) of each of the 12 methods across the different forecasted quantities is given, where T denotes the length of the time series, and d denotes the dimensionality of each data set. Note that the Air Quality data set mentioned here is explored in Appendix B.4.

of OFTER and OFTER-DR are more comparable, with the latter being slightly more computationally intensive.

For the numerical experiments concerning forecasting returns, we employ economic-driven metrics typical of this literature, as discussed in Section 3.4.3. For all the remaining experiments, we base the performance comparisons on RMSE and the other metrics defined in Section 3.4.4.

3.4.5.1 Synthetic data

For this task we use $L = 800$, $c_{\min} = 0.05$ and $c_{\text{Original}} = 0.05$, $L_D = 500$. The results are given in Table 3.2 with the RMSE scores for M1-1, M2-1 and M3-1 visualized in Figure 3.3. Importantly, in most cases OFTER has a p-value of zero for the DM test, and Theil-U statistics that are significantly lower than 1. Two exceptions are M1.3 (which is structurally a very low signal to ratio model), and M2.1 for which the OFTER configurations that do not use the feature transformation component score very poorly (similar to other linear methods).

For M1, the best performing models correspond to the linear methods. GRU and LSTM also outperform the OFTER configurations in a number of cases, with differences being small in most cases. On the other hand, OFTER generally outperforms Random Forests, XGBoost and SVM. In the case of M2, which is a non-linear multivariate non-linear model, the configurations of OFTER that employ the feature transformation component generally perform the best, with the exception of Random Forests, LSTM and GRU. In this case, OFTER outperforms the linear methods as well as XGBoost. Lastly, in the case of M3, which regards a vector autoregressive (VAR) model, OFTER is performing particularly well, outperforming all other models, with the exception of GRU, and having similar performances to the LM and LASSO.

M1	M1.1				M1.2				M1.3				M1.4				M1.5			
	MSE	MAE	TU	DM	MSE	MAE	TU	DM	MSE	MAE	TU	DM	MSE	MAE	TU	DM	MSE	MAE	TU	DM
LM	0.932	0.745	0.882	0.000	0.978	0.774	0.860	0.000	1.017	0.820	0.992	0.031	0.965	0.756	0.824	0.000	0.978	0.777	0.983	0.000
PCR	0.931	0.744	0.880	0.000	0.979	0.774	0.861	0.000	1.015	0.820	0.991	0.012	0.969	0.763	0.827	0.000	0.977	0.777	0.982	0.000
LASSO	0.928	0.740	0.877	0.000	0.977	0.772	0.859	0.000	1.014	0.818	0.989	0.027	0.965	0.759	0.824	0.000	0.976	0.776	0.981	0.000
Random Forests	0.965	0.766	0.913	0.000	0.979	0.781	0.860	0.000	1.034	0.832	1.009	0.200	0.972	0.765	0.829	0.000	0.990	0.784	0.995	0.010
SVM	0.955	0.764	0.903	0.000	1.029	0.812	0.904	0.000	1.061	0.854	1.035	0.131	1.013	0.804	0.865	0.000	1.001	0.796	1.006	0.029
XGBOOST	1.027	0.820	0.972	0.000	1.073	0.856	0.943	0.000	1.114	0.893	1.086	0.480	1.079	0.859	0.921	0.000	1.087	0.871	1.093	0.123
LSTM	0.923	0.728	1.000	0.091	0.946	0.755	0.982	0.002	1.010	0.821	1.001	0.866	0.975	0.772	1.003	0.135	0.980	0.787	0.998	0.039
Bi-LSTM	1.073	0.859	1.163	0.605	1.054	0.851	1.093	0.967	1.141	0.913	1.131	0.599	1.094	0.883	1.125	0.003	1.084	0.860	1.105	0.346
GRU	0.945	0.748	1.024	0.012	0.933	0.747	0.968	0.000	1.012	0.826	1.003	0.683	0.938	0.742	0.965	0.000	1.002	0.804	1.020	0.416
OFTER	0.943	0.753	0.892	0.000	0.990	0.786	0.869	0.000	1.024	0.828	0.998	0.300	0.971	0.763	0.829	0.000	0.986	0.783	0.991	0.002
OFTER-DR	0.941	0.750	0.890	0.000	0.987	0.782	0.867	0.000	1.019	0.826	0.995	0.060	0.970	0.762	0.828	0.000	0.980	0.779	0.985	0.003
OFTER-FT	0.937	0.747	0.886	0.000	0.981	0.779	0.861	0.000	1.020	0.825	0.995	0.073	0.965	0.756	0.824	0.000	0.977	0.774	0.982	0.000
OFTER-DR-FT	0.947	0.756	0.896	0.000	0.984	0.781	0.864	0.000	1.020	0.826	0.995	0.021	0.963	0.760	0.822	0.000	0.979	0.777	0.984	0.001

M2	M2.1				M2.2				M2.3				M2.4				M2.5			
	MSE	MAE	TU	DM	MSE	MAE	TU	DM	MSE	MAE	TU	DM	MSE	MAE	TU	DM	MSE	MAE	TU	DM
LM	1.157	0.918	1.004	0.590	1.374	1.113	0.999	0.261	1.520	1.141	0.879	0.000	1.025	0.822	0.815	0.000	1.036	0.827	0.991	0.001
PCR	1.157	0.917	1.004	0.542	1.376	1.113	1.001	0.248	1.553	1.172	0.898	0.000	1.052	0.843	0.837	0.000	1.036	0.827	0.991	0.001
LASSO	1.155	0.915	1.003	0.500	1.373	1.112	0.998	0.485	1.519	1.140	0.879	0.000	1.025	0.821	0.815	0.000	1.033	0.825	0.988	0.000
Random Forests	1.002	0.802	0.869	0.000	1.040	0.832	0.757	0.000	1.066	0.832	0.617	0.000	1.067	0.860	0.848	0.000	1.043	0.837	0.998	0.001
SVM	1.201	0.949	1.042	0.741	1.249	1.005	0.908	0.000	1.253	0.985	0.725	0.000	1.066	0.863	0.848	0.000	1.069	0.849	1.023	0.002
XGBOOST	1.068	0.853	0.927	0.000	1.153	0.920	0.840	0.000	1.125	0.886	0.651	0.000	1.152	0.920	0.916	0.000	1.125	0.904	1.076	0.069
LSTM	0.991	0.787	1.002	0.285	0.990	0.805	1.002	0.072	1.029	0.781	1.006	0.215	1.030	0.820	1.002	0.470	1.040	0.838	1.000	0.356
Bi-LSTM	1.078	0.864	1.090	0.601	1.104	0.877	1.116	0.323	1.130	0.921	1.104	0.170	1.130	0.903	1.098	0.277	1.145	0.921	1.100	0.108
GRU	0.998	0.793	1.010	0.568	1.002	0.814	1.013	0.011	1.021	0.778	0.998	0.010	1.013	0.808	0.985	0.000	1.048	0.842	1.008	0.932
OFTER	1.154	0.914	1.001	0.282	1.254	1.006	0.912	0.000	1.507	1.124	0.872	0.000	1.026	0.823	0.816	0.000	1.045	0.839	0.999	0.004
OFTER-DR	1.039	0.803	0.901	0.000	1.203	0.959	0.875	0.000	1.388	1.018	0.803	0.000	0.956	0.749	0.760	0.000	0.950	0.744	0.908	0.000
OFTER-FT	1.016	0.797	0.882	0.000	1.223	0.973	0.890	0.000	1.060	0.808	0.613	0.000	0.954	0.749	0.758	0.000	0.968	0.754	0.926	0.000
OFTER-DR-FT	0.994	0.781	0.862	0.000	1.261	0.998	0.917	0.000	1.053	0.805	0.609	0.000	0.971	0.763	0.772	0.000	0.965	0.757	0.923	0.000

M3	M3.1				M3.2				M3.3				M3.4				M3.5			
	MSE	MAE	TU	DM	MSE	MAE	TU	DM	MSE	MAE	TU	DM	MSE	MAE	TU	DM	MSE	MAE	TU	DM
LM	0.980	0.787	0.541	0.000	0.965	0.765	0.422	0.000	1.005	0.807	0.692	0.000	1.000	0.801	0.366	0.000	0.996	0.799	0.494	0.000
PCR	1.070	0.851	0.591	0.044	1.081	0.854	0.473	0.000	1.077	0.860	0.742	0.900	1.204	0.955	0.440	0.018	1.195	0.968	0.593	0.000
LASSO	0.981	0.787	0.542	0.000	0.965	0.764	0.422	0.000	1.004	0.806	0.691	0.000	1.001	0.803	0.366	0.000	0.995	0.799	0.494	0.000
Random Forests	1.018	0.820	0.562	0.000	1.021	0.806	0.446	0.000	1.038	0.826	0.714	0.000	1.041	0.833	0.381	0.000	1.031	0.833	0.512	0.000
SVM	1.056	0.840	0.583	0.000	1.108	0.874	0.484	0.000	1.056	0.843	0.727	0.000	1.148	0.888	0.420	0.000	1.076	0.856	0.534	0.000
XGBOOST	1.102	0.883	0.608	0.000	1.123	0.893	0.491	0.000	1.119	0.894	0.769	0.000	1.140	0.909	0.417	0.000	1.097	0.890	0.545	0.000
LSTM	1.070	0.852	1.021	0.538	1.080	0.859	1.026	0.404	0.978	0.787	0.932	0.202	1.105	0.889	1.050	0.602	1.087	0.875	1.019	0.004
Bi-LSTM	1.157	0.933	1.104	0.981	1.146	0.918	1.089	0.040	1.166	0.942	1.112	0.642	1.128	0.912	1.072	0.049	1.157	0.946	1.084	0.000
GRU	0.819	0.648	0.781	0.000	0.899	0.714	0.854	0.000	1.034	0.828	0.985	0.363	0.881	0.698	0.837	0.198	0.906	0.723	0.849	0.035
OFTER	0.981	0.790	0.542	0.000	0.965	0.763	0.422	0.000	1.012	0.811	0.697	0.000	0.995	0.797	0.364	0.000	0.998	0.804	0.495	0.000
OFTER-DR	0.981	0.791	0.542	0.000	0.961	0.760	0.420	0.000	1.009	0.807	0.694	0.000	0.995	0.800	0.364	0.000	0.995	0.803	0.494	0.000
OFTER-FT	0.983	0.791	0.543	0.000	0.972	0.769	0.425	0.000	1.008	0.804	0.694	0.000	0.992	0.802	0.363	0.000	0.995	0.800	0.493	0.000
OFTER-DR-FT	0.982	0.791	0.542	0.000	0.979	0.778	0.428	0.000	1.014	0.809	0.698	0.000	1.001	0.806	0.366	0.000	0.999	0.802	0.495	0.000

Table 3.2: This table summarizes the performance of the 13 methods for each of the three synthetic models described in Section 3.4.2. RMSE, MAE, TU and DM denote the Root Mean Squared Error, Mean Absolute Error, Theil-U statistic, and the p-value of the Diebold-Mariano test respectively.

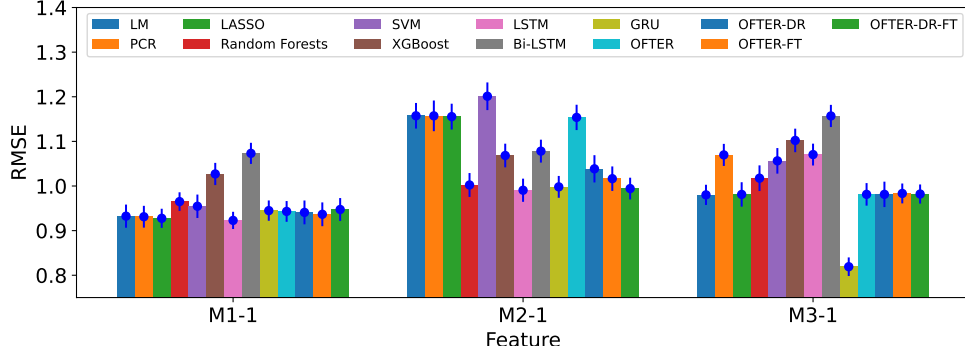


Figure 3.3: Grouped barplot showing a performance comparison of all methods for the synthetic data set. The metric of the comparison is the RMSE between the target values and the forecasts. The bootstrapped standard errors are also displayed.

3.4.5.2 ETF volumes

For both financial tasks, we use $L = 800$ and set $c_{\min} = c_{\text{Original}} = 0.01$ and $L_D = 300$, as typically, these are problems with a lower signal-to-noise ratio. For the task of predicting the delta-volumes of ETF instruments Equation (3.54), OFTER outperforms all other benchmark methods. As demonstrated in Table 3.3, the best performing method is an OFTER configuration in seven out of ten cases, with random forests being the best performing for XLB and XLP, and LASSO being the best performing for XLU. In addition, in all cases, for all OFTER configurations, the p-value corresponding to the DM test is equal to zero up to four decimal places.

ETF market excess returns

We simulate a daily strategy based on trading ten ETF instruments for the Q1-Q5 portfolios, and summarize the results in Table 3.4. For three out of four OFTER strategies, the SR is considered statistically significant.¹¹ As a visual demonstration of the results, we also plot the cumulative P&L for the Q1 strategy of each method in Figure 3.4, clearly demonstrating that OFTER-DR and OFTER-DR-FT are outperforming the other methods. With the exception of the Q5-SVM strategy, none of the other benchmark methods result in strategies with a statistically significant SR. Across most quantiles, OFTER-DR scores the highest performance, both in terms of SR and PPD. In Figure 3.4, we plot the cumulative P&L for the OFTER-DR

¹¹Using the Sharpe Ratio significance test of [17]. This is based on a generalization of the t-test, that accounts for the non-normality of returns, by taking into account the kurtosis and skewness of the P&L distribution.

	SPY				XLF				XLB				XLK				XLV			
	MSE	MAE	TU	DM	MSE	MAE	TU	DM	MSE	MAE	TU	DM	MSE	MAE	TU	DM	MSE	MAE	TU	DM
LM	0.016	0.013	0.860	0.000	0.020	0.015	0.938	0.000	0.025	0.020	0.947	0.000	0.022	0.017	0.881	0.000	0.023	0.019	0.967	0.000
PCR	0.016	0.013	0.861	0.000	0.019	0.014	0.883	0.000	0.024	0.019	0.903	0.000	0.022	0.017	0.857	0.000	0.022	0.018	0.920	0.000
LASSO	0.016	0.013	0.864	0.000	0.018	0.014	0.867	0.000	0.023	0.018	0.886	0.000	0.022	0.017	0.859	0.000	0.022	0.017	0.909	0.000
Random Forests	0.016	0.013	0.865	0.000	0.019	0.014	0.875	0.000	0.023	0.018	0.860	0.000	0.022	0.017	0.864	0.000	0.022	0.017	0.908	0.000
SVM	0.019	0.015	1.003	0.010	0.021	0.017	0.997	0.000	0.030	0.023	1.124	0.000	0.030	0.025	1.176	0.000	0.033	0.026	1.354	0.000
XGBOOST	0.017	0.014	0.915	0.000	0.021	0.016	0.969	0.000	0.026	0.020	0.973	0.000	0.024	0.019	0.963	0.000	0.026	0.020	1.058	0.000
LSTM	0.980	0.774	0.996	0.007	0.509	0.396	1.001	0.255	0.452	0.355	1.009	0.023	0.629	0.487	1.000	0.011	0.372	0.291	1.018	0.024
Bi-LSTM	1.100	0.882	1.117	0.669	0.704	0.586	1.384	0.008	0.665	0.560	1.483	0.078	0.797	0.653	1.267	0.069	0.632	0.550	1.729	0.835
GRU	0.962	0.760	0.978	0.000	0.520	0.409	1.023	0.000	0.470	0.364	1.050	0.014	0.640	0.497	1.018	0.528	0.391	0.308	1.069	0.016
OFTER	0.016	0.013	0.862	0.000	0.019	0.015	0.883	0.000	0.023	0.018	0.893	0.000	0.022	0.017	0.883	0.000	0.022	0.017	0.915	0.000
OFTER-DR	0.016	0.012	0.844	0.000	0.018	0.014	0.860	0.000	0.023	0.018	0.886	0.000	0.022	0.016	0.851	0.000	0.022	0.017	0.905	0.000
OFTER-FT	0.016	0.012	0.853	0.000	0.018	0.014	0.862	0.000	0.023	0.018	0.881	0.000	0.022	0.017	0.859	0.000	0.022	0.017	0.888	0.000
OFTER-DR-FT	0.016	0.012	0.856	0.000	0.018	0.014	0.858	0.000	0.023	0.018	0.888	0.000	0.022	0.017	0.860	0.000	0.022	0.017	0.915	0.000

	XLI				XLU				XLY				XLP				XLE			
	MSE	MAE	TU	DM	MSE	MAE	TU	DM	MSE	MAE	TU	DM	MSE	MAE	TU	DM	MSE	MAE	TU	DM
LM	0.025	0.019	1.046	0.000	0.021	0.016	0.958	0.000	0.025	0.020	0.956	0.000	0.023	0.018	0.902	0.000	0.020	0.015	0.993	0.000
PCR	0.022	0.017	0.937	0.000	0.020	0.015	0.902	0.000	0.023	0.018	0.901	0.000	0.022	0.018	0.866	0.000	0.018	0.015	0.925	0.000
LASSO	0.021	0.017	0.906	0.000	0.019	0.014	0.862	0.000	0.023	0.018	0.889	0.000	0.022	0.017	0.857	0.000	0.018	0.014	0.898	0.000
Random Forests	0.021	0.017	0.912	0.000	0.019	0.015	0.876	0.000	0.022	0.018	0.868	0.000	0.022	0.017	0.856	0.000	0.017	0.014	0.881	0.000
SVM	0.030	0.024	1.278	0.000	0.030	0.024	1.379	0.000	0.033	0.026	1.260	0.000	0.028	0.022	1.085	0.000	0.027	0.022	1.397	0.000
XGBOOST	0.026	0.020	1.115	0.000	0.022	0.018	1.039	0.000	0.027	0.021	1.032	0.000	0.026	0.020	1.004	0.000	0.020	0.016	1.012	0.000
LSTM	0.310	0.243	1.015	0.987	0.389	0.303	1.025	0.166	0.360	0.279	1.041	0.874	0.469	0.353	0.999	0.001	0.404	0.318	1.031	0.081
Bi-LSTM	0.604	0.537	1.976	0.844	0.628	0.543	1.652	0.686	0.616	0.539	1.779	0.168	0.687	0.585	1.462	0.001	0.639	0.550	1.632	0.797
GRU	0.323	0.255	1.057	0.000	0.406	0.319	1.068	0.058	0.370	0.288	1.068	0.002	0.498	0.381	1.060	0.288	0.410	0.325	1.048	0.003
OFTER	0.022	0.017	0.935	0.000	0.019	0.015	0.887	0.000	0.023	0.018	0.886	0.000	0.023	0.018	0.901	0.000	0.018	0.014	0.898	0.000
OFTER-DR	0.022	0.017	0.929	0.000	0.019	0.014	0.863	0.000	0.023	0.018	0.880	0.000	0.023	0.018	0.888	0.000	0.017	0.013	0.876	0.000
OFTER-FT	0.021	0.017	0.904	0.000	0.019	0.014	0.868	0.000	0.022	0.018	0.858	0.000	0.022	0.017	0.859	0.000	0.017	0.014	0.880	0.000
OFTER-DR-FT	0.022	0.017	0.932	0.000	0.019	0.014	0.871	0.000	0.022	0.018	0.866	0.000	0.023	0.018	0.885	0.000	0.017	0.014	0.876	0.000

Table 3.3: This table summarizes the performance of various regression methods for the delta volume forecasting task across each of the ETFs. RMSE, MAE, TU and DM denote the Root Mean Squared Error, Mean Absolute Error, Theil-U statistic, and the p-value of the Diebold-Mariano test respectively.

	Q1				Q2				Q3				Q4				Q5		
	SR	PPD	p-value		SR	PPD	p-value		SR	PPD	p-value		SR	PPD	p-value		SR	PPD	p-value
LM	-0.745	-1.211	0.967		-0.729	-1.270	0.964		-0.873	-1.783	0.984		-0.740	-1.689	0.965		-0.814	-2.664	0.977
PCR	-0.863	-1.324	0.981		-0.901	-1.523	0.985		-0.841	-1.737	0.979		-0.952	-2.234	0.989		-0.809	-2.728	0.976
LASSO	-0.890	-1.293	0.985		-0.920	-1.411	0.987		-1.052	-1.905	0.995		-1.192	-2.466	0.998		-0.995	-3.163	0.993
Random Forests	0.371	0.534	0.182		0.282	0.436	0.246		0.174	0.339	0.335		0.233	0.521	0.285		0.285	1.000	0.243
SVM	-0.370	-0.528	0.807		-0.314	-0.471	0.763		0.309	0.429	0.204		0.613	0.984	0.058		0.706	1.267	0.039
LSTM	0.308	0.433	0.227		0.261	0.396	0.262		0.421	0.771	0.152		0.403	0.818	0.163		0.059	0.176	0.442
Bi-LSTM	-0.810	-0.868	0.976		-0.824	-1.017	0.977		-0.388	-0.641	0.828		-0.238	-0.462	0.719		-0.456	-1.377	0.866
GRU	0.424	0.610	0.152		0.286	0.448	0.243		0.148	0.268	0.359		0.144	0.294	0.363		0.129	0.403	0.376
OFTER	0.534	0.811	0.089		0.653	1.082	0.052		0.616	1.203	0.064		0.290	0.646	0.233		0.444	1.463	0.140
OFTER-DR	1.616	2.346	0.000		1.704	2.607	0.000		1.229	2.194	0.001		1.258	2.441	0.001		1.261	3.765	0.001
OFTER-FT	1.143	1.737	0.002		1.344	2.197	0.000		1.261	2.476	0.001		0.988	2.167	0.007		0.491	1.545	0.110
OFTER-DR-FT	1.468	2.125	0.000		1.532	2.409	0.000		1.562	2.795	0.000		0.884	1.760	0.015		0.826	2.371	0.022

Table 3.4: This table summarizes the performance of various regression methods for the ETF price movement forecasting task. SR denotes the annualized Sharpe Ratio, PPD denotes the Profit Per Dollar, and p-value denotes the p-value obtained from the one-tailed SR significance test of [17]. Q1-Q5 correspond to the quantile portfolios as defined in Section 3.4.3.

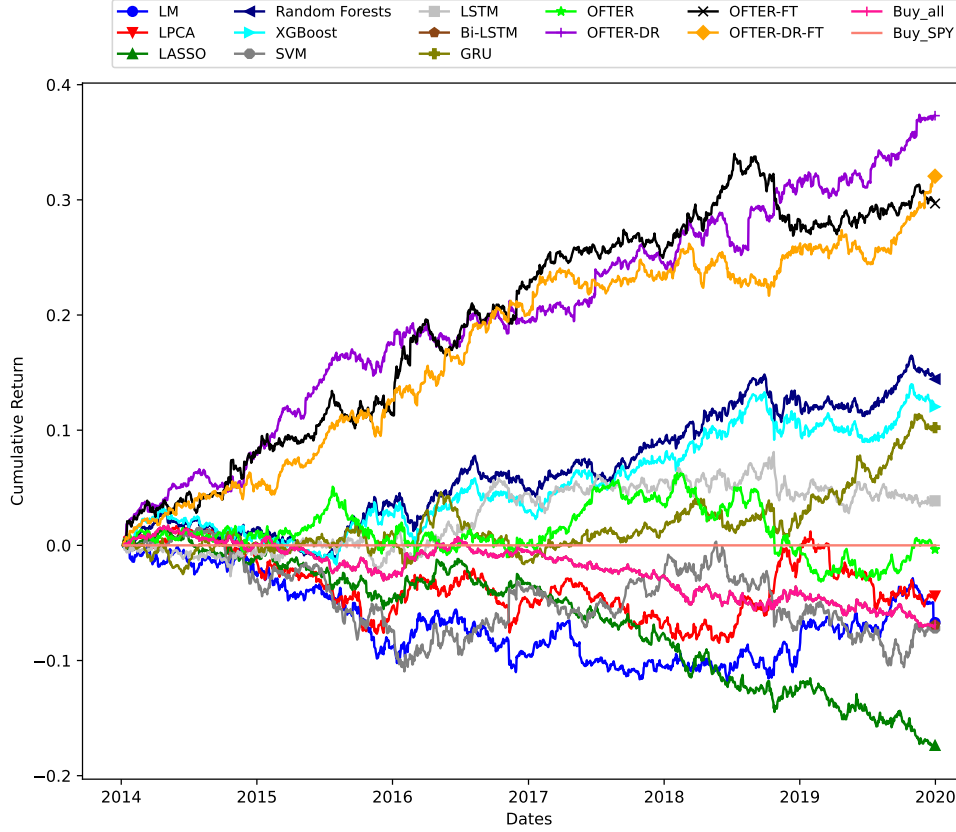


Figure 3.4: Cumulative P&L plot of the daily strategies corresponding to the 13 methods for the ETF price movement forecasting task. SR denotes the annualized Sharpe Ratio, PPD denotes the Profit Per Dollar, and p-value denotes the p-value as obtained from the one-tailed Sharpe Ratio significance test of [17]. The performance metrics for each method are given in Table 3.4.

strategy, clearly demonstrating a consistently well-performing set of strategies across different quantiles.

An application to an air pollution data set is given in Appendix B.4.

3.5 Applications to Feature Importance and Outlier Detection

In this section, we discuss additional applications of OFTER to two tasks of interest in the machine learning community. In particular, we consider a feature importance component that boosts the interpretative power of OFTER, and an application to outlier detection.

3.5.1 Feature Importance

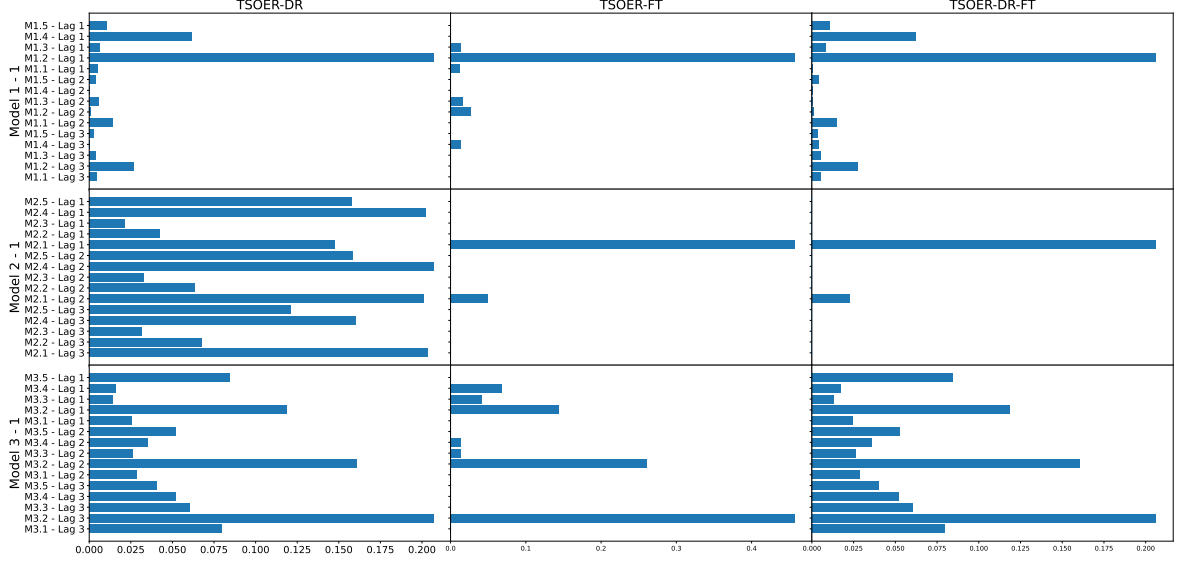


Figure 3.5: In this 3x3 plot grid, we show the feature importance weights, as obtained in Section 3.5.1. The first feature of each of M1, M2, and M3 is forecasted in the first, second and third row, respectively. In each case, there are 15 features corresponding to three lags of each of the five features. This is repeated for three of OFTER’s configurations, corresponding to the three columns of the grid.

In the OFTER pipeline, one can put forth a natural approach for estimating the contribution of each feature to the forecasting task. In particular, we aim to find a proxy for the importance of each of the input features. To this end, consider two points in the initial feature space x_1 and x_2 , and denote by $x'_1 = U^T x_1$ and $x'_2 = U^T x_2$ the transformed variables. Denote by $D' = \|x'_1 - x'_2\|_v$ the distance in the reduced space (used for calculating the nearest neighbors), and let $D_j = x_{1,j} - x_{2,j}$ such that $\|x_1 - x_2\|_2 = \sqrt{\sum_j D_j^2}$.

A metric for the importance of a variable in determining the relative distance of data points in the embedding space can be modeled as the sensitivity of the distance D' on the individual distance d_j , i.e $\left| \frac{\partial D'}{\partial D_j} \right|$. For our goal, it is natural to take the derivative at a point where $D_1 = \dots = D_d = 1$. We begin by writing

$$D'^2 = \|x'_1 - x'_2\|_v = \|U^T(x_1 - x_2)\|_v = \sum_{j=1}^p v_j \left(\sum_{k=1}^d U_{k,j}(x_1 - x_2)_k \right)^2 = \sum_{j=1}^p v_j \left(\sum_{k=1}^d U_{k,j} D_k \right)^2. \quad (3.61)$$

We then compute

$$\left. \frac{\partial D'^2}{\partial D_k} \right|_{d=e_d} = 2 \sum_{j=1}^p v_j \left(\sum_{k'=1}^d U_{k',j} D_{k'} \right) U_{k,j}, \quad (3.62)$$

which, using $\odot$ to denote the Hadamard product, can be written as

$$\nabla D'^2(e_d) = 2 \left| U \left[(U^T e_d) \odot v \right] \right|. \quad (3.63)$$

Keeping in mind Equation (3.42), we also augment the embedding space with the original features that are highly correlated with the target. Consider the vector v_2 containing the weight given to the original feature if used, and zero otherwise, i.e

$$v_{2,j} = \begin{cases} v_{I_k^{\text{Original}}} & j = I_k^{\text{Original}} \\ 0 & j \notin I^{\text{Original}} \end{cases}.$$

This now leads to

$$D'^2 = \sum_{j=1}^p v_j \left(\sum_{k=1}^d U_{k,j} D_k \right)^2 + \sum_{l=1}^d v_{2,l} D_l^2, \quad (3.64)$$

and thus the derivative is instead given by

$$\nabla D'^2(e_d) = 2 \left| U \left[(U^T e_d) \odot v \right] + v_2 \right|. \quad (3.65)$$

As an illustration, we consider the task of forecasting the first feature of each of the three synthetic data sets (M1-M3), using OFTER-DR, OFTER-FT, and OFTER-DR-FT. In Figure 3.5, we show a grid of the feature importance of each individual feature. As we ignore the unimportant features via the calculation of v in Equation (3.44), OFTER-FT attaches a non-zero importance weight only to a subset of the features. On the other hand, OFTER-DR assigns a non-zero weight to all of the original features, as they all contribute to the estimation of each principal component. This also holds true for OFTER-DR-FT, where dimensionality reduction is also performed. However, in the case of M2 where most of the predictability is present in a single feature (M2.1 - Lag1), OFTER-DR-FT assigns a zero importance weight to most of the features, similarly to OFTER-FT. Therefore, for the task of feature importance, we recommend the user to rely on either OFTER-FT or OFTER-DR-FT.

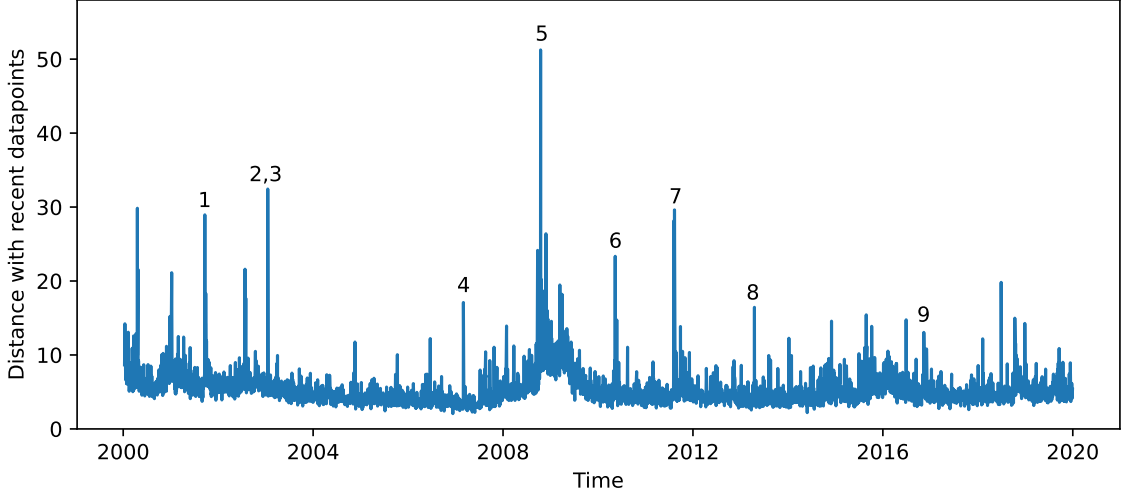


Figure 3.6: Plot showing the minimum distance $d_j^{\min}$ obtained using the suggested outlier detection algorithm for OFTER-DR, $L = 600$ and $\kappa = 5$. All numbered points are identified as outliers.

1. 18 September 2001 - Aftermath of the 9/11 Attack
2. 24 July 2002 - Stock market downturn of 2002 (12th consecutive day of falling prices)
3. 29 July 2002 - Stock market downturn of 2002 (3rd day of rising prices after sharp price rises following (2))
4. 28 February 2007 - Big drop in US and China stocks
5. 10 October 2008 - Market Crash across Europe and Asia
6. 7 May 2010 - Following the Flash Crash of May 6
7. 4 August 2011 - Preceding the Black Monday (8 August)
8. 15 April 2013 - One of the biggest drops in 2013 for DJIA, SPX
9. 8 November 2016 - US election

3.5.2 Outlier Detection

A byproduct of the OFTER pipeline is that it can facilitate the task of outlier detection, by considering the pairwise distance between time points in the embedding space. More specifically, points whose nearest neighbors are still quite “far” in the embedding space could be labeled as outliers, in the same spirit as the one-class SVM algorithm [207]. We demonstrate this using the following example.

Define the minimum distance between the current time point t and the time points in its recent history in the reduced space as

$$d_t^{\min} = \min(\{d_{t-L,t}, \dots, d_{t-1,t}\}). \quad (3.66)$$

Comparing $d_t^{\min}$ with the distribution of $(d_{t-L}^{\min}, \dots, d_{t-1}^{\min})$ may constitute input for any

outlier detection algorithm.¹² Denoting by $q_{t,1}^{\min}$ and $q_{t,3}^{\min}$ the first and third quantiles of $\{d_{t-L}^{\min}, \dots, d_{t-1}^{\min}\}$, we classify a time point t as an outlier if $d_j^{\min} > q_{t,3}^{\min} + \kappa (q_{t,3}^{\min} - q_{t,1}^{\min})$ for some $\kappa > 0$.

In Figure 3.6, we plot the minimum distance between each trading day and its $L = 600$ preceding days, by using $\kappa = 5$. From this figure, one could infer several different regimes, as there are periods where the levels of $d_{\min}$ are low, and periods where these are high, such as the periods around the 2008 financial crisis. The fact that almost all labeled outliers fall in the immediate vicinity of days significant in the financial history, does indeed indicate that the algorithm can successfully be used to uncover abnormality signals.

Finally, we remark that selecting a larger value of L would result in a large number of anomalies being highlighted. In practice, if this outlier detection approach is used in a trading application, a smaller κ would be desirable, as the outlier warning could possibly be qualitatively analyzed. Note that the above quantile-based outlier detection approach is just a sample algorithm to illustrate the potential of OFTER, as the pipeline is amenable to be incorporated into any other outlier detection algorithm.

3.6 Conclusion and Future Work

In this paper, we have introduced OFTER, a complete pipeline for online multivariate time series forecasting. By leveraging kNN and GRNN, both simple well-understood methods with trivial training complexities, OFTER manages to outperform a number of baselines, for mid-sized data sets. In particular, an extensive set of numerical experiments demonstrated that OFTER performed competitively with Random Forests and GRU in the synthetic data set, and outperformed all other methods in the real-world applications considered, with a running time much lower than those of deep learning methods. The shortcomings due to the curse of dimensionality are overcome via dimensionality reduction and the use of a weighted norm when comparing distances in the embedding space.

Furthermore, OFTER brings additional advantages, including its flexible nature, with the ability to use or omit many of its components. As a byproduct, OFTER can be integrated seamlessly into an outlier detection pipeline, in an online manner. Lastly, OFTER is naturally equipped to return the user a feature importance score for all the input variables.

¹²Instead of $d_t^{\min}$, one could employ the average distance $\bar{d}_t = \frac{\sum_{j=t-L}^{t-1} d_{j,t}}{t}$.

Future work. Future extensions of this work could include the utilization of other machine learning methods in the model selection step of the algorithm, for example, LASSO models with different regularization parameters. One disadvantage of OFTER is that it struggles to discover non-linear interaction effects between covariates, even though this is a general issue with models and data sets of this size. One potential extension would be to include a pre-screening component, where a number of potential interaction terms are tested in order to augment the feature space, either prior to or after the embedding step. Additionally, it would be interesting to see the performance of OFTER in other mid-sized data sets and assess whether other dimensionality reduction algorithms are able to attain improved performance, relative to the use of PCA (this was not the case for the data sets used in this project). Yet another interesting direction concerns an adaptation of OFTER for probabilistic forecasting. Probabilistic models have recently gained in popularity, being able to provide not only forecasts but also associated uncertainty estimates, which is of crucial importance in certain applications. In addition, multi-step probabilistic forecasting is another area which could provide a natural extension for the pipeline presented in this paper. Lastly, we note that the OSMC metric we introduced is of independent interest, both as an alternative to the maximal correlation but also as a computationally fast additional step for other simple regression methods.

Chapter 4

Options-driven volatility forecasting

4.1 Introduction

Forecasting volatility Volatility forecasting is one of the most well-studied problems in finance and is considered an essential tool for various market participants, such as market makers. Notable applications of volatility forecasting include [82], who show how volatility perception can affect risk management, [130], who incorporate volatility forecasts for bond yield forecasting, and [257], who use it for option pricing. Volatility forecasts can also be utilized directly as an alternative form of speculation, either via the use of option straddles or by other future strategies (see [127] for various examples).

Acknowledging the importance of volatility in financial markets, the Chicago Board Options Exchange (CBOE) introduced the volatility index (VIX) in 1993, which tracks the volatility of the S&P 500 index. With recent turbulent periods, interest in volatility has been as high as ever. Notably, VIX attained its highest value yet on 16 March 2020, with the long-term average generally considered to be below 20.

On one hand, the non-stationarity, non-linearity, and low signal-to-error nature of volatility render its forecasting task quite challenging. On the other hand, volatility is generally regarded as much easier to forecast compared with future price returns [10, 99]. This is partly due to volatility clustering, a phenomenon first noted by [162], where changes in price at a given time are likely to be followed by further price changes of a similar magnitude. Another important property of volatility is its path dependency, with [104] recently providing a simple yet high-performing model based on the weighted average of past returns and squared returns.

Historically, early predictive models for volatility were based on simple historical smoothing and averaging methods, albeit with rather limited success. An influential model was the Autoregressive Conditional Heteroskedasticity (ARCH) model by [85], which accounts for volatility clustering. This model has resulted in a stream of extensions and generalizations, starting with the Generalized Autoregressive Conditional Heteroskedasticity (GARCH) by [36].

The Heterogeneous Autoregressive (HAR) model, by [71], is a simple additive model that uses the daily, weekly and monthly volatilities as the only features of an OLS model. The motivation for HAR was twofold. Firstly, the Heterogeneous Market Hypothesis (HMH) states that the market is composed of different types of investors. In terms of their difference in trading frequency, traders can be approximately separated into three categories; intraday or high-frequency traders, investors who rebalance their portfolios weekly, and long-term investors who update their positions at a monthly frequency or longer. Investors with differing trading frequencies possibly also perceive and influence volatility at differing scales. Secondly, the HAR model appears to capture some properties of the return distributions as well, including the heavier-than-Gaussian tails and apparent long-memory structure in volatility.

An extension of the HAR model was proposed by [37], the HARQ model, including terms for the realized *quarticity* (fourth power variation), showing improvement in forecasting performance. Another important phenomenon is the leverage effect [93], whereby a negative return leads to increased volatility due to an increase in leverage. This results in the well-known negative correlation of returns and subsequent return volatility, which can further be insightful for forecasting volatility. Various extensions to ARCH and GARCH that account for this phenomenon have also been introduced, a review of which can be found in [32].

In recent decades, with the rapid rise of machine learning, there has been a turn away from parametric models for the volatility forecasting task, as many studies, such as [44, 61, 13, 14, 169, 157], have shown that black-box methods from the machine learning literature can match the performance or even outperform the classical volatility models. In particular, a plethora of deep learning methods have been proposed. Examples include [152] who use Long-Short Term Memory networks, and [55] who propose a Hybrid Deep Neural Network by utilizing a joint temporal embedding of volatility and price. [197] formulated HARnet, and [174] formulated DeepVol, both models with a dilated convolutional network structure.

Ensembles of various volatility estimators Combining multiple estimators is a popular way of reducing the effects of measurement errors. In the context of volatility, which itself is not an observable quantity, this is a widely studied topic. [180] use various volatility formulas related to the GARCH family of models to show an improvement in forecasting tail risk, [192] show that a weighted average of different volatility estimators outperforms any single estimator, when used as the input feature of a HAR model. Extreme range estimators have been shown to boost forecasting performance by [135].

Commonality in volatility Similar to the commonality in equity returns, volatilities can also be separated into market and idiosyncratic volatilities. A recent study is [248], who used the intraday commonality of volatilities to show improved forecasting performance. Earlier, [110] showed how stocks that are less correlated with the volatility of the market tend to outperform other stocks. Common volatility applications also include [90], who used VIX as a proxy for commonality.

Implied Volatility A well-researched topic is the role of implied volatility in forecasting of realized volatility. Various studies in the past have studied this question in detail [47, 138, 125]. While it is a common belief that implied volatility leads realized volatility, practice shows that when forecasting future volatility, predictability remains low, with [125] noting that implied volatility underpredicts future volatility. However, more recent and complex methodologies provide evidence in the opposite direction, such as [74], who show a significant Granger causality from option implied volatility to stock realized volatility and [196], who use a bivariate copula model to conclude that implied volatility is a better predictor of realized volatility than vice versa. [160] demonstrate the increased predictive power of implied volatility during the pandemic.

The main reason behind the failure of implied volatilities is the high dimensionality of such a data set, which naturally points to dimensionality reduction as the remedy. An early study on the dimensionality reduction of implied volatility surfaces is that of [16], who rely on Principal Component Analysis (PCA). [68] use a Karhunen–Loeve decomposition to show how IVS surfaces can be empirically modeled using only 2-3 functional components. [79] classify equity options in the U.S. markets based on their level of systemic and idiosyncratic risk, relying on the fluctuations of their implied volatility surfaces. They propose a dimensionality reduction model based on PCA

and random matrix theory¹. [15] construct an eigenportfolio based on the volatility returns, and show how the top eigenvalue is correlated to a weighting scheme based on Vega and open interest. Furthermore, they show that using a multilinear SVD on the order-4 tensor of implied volatility surfaces, is better able to track the open interest-Vega portfolio compared to the flattened PCA of returns.

Alternatively, custom features are extracted from the implied volatility surface, which usually depend on specific delta-maturity pairs. Relevant literature include the concepts of implied volatility skew, defined by [239] as the difference between the implied volatilities of an ATM call option with an OTM put option, both with 30-day maturities. Other popular features include the CIV and PIV. The CIV (PIV) corresponds to the implied volatility of an ATM call (put) option. Such features are used by [106] to predict jumps in equity returns.

A popular approach is proposed by [81], who use a deterministic volatility function approach to extract various features. More sophisticated approaches include [96] who extract the volatility skew and term structure using an expansion formula for the Black–Scholes volatility. Lastly, [6] extract a closed form for implied volatility surface features such as curvature and term-structure under a stochastic volatility model with jumps.

Volatility-volume interplay The interplay between volatility and volume has also been widely studied, with the mixture of distribution hypothesis by [69] and the sequential information arrival hypothesis by [63], giving two different interpretations of how information embedded in trading volume is informative for future changes in price. Recent studies into this interplay include [150] who use an empirical mode decomposition method on trading volume to show improvements in volatility forecasting. There are also a number of studies that investigate this interplay for the cryptocurrency market, such as [40] who look into the Granger causality from trading volume to volatility.

Main contributions In this paper, we tackle the task of forecasting daily volatility. Previously, [227] and [192] have both explored how to combine multiple volatility proxies into a single estimator, aiming to minimize the measurement error and maximize the predictive power of a HAR or GARCH type of model. We also base our model on the HAR model, but instead of substituting the realized volatility for a

¹In particular, the Marchenko-Pastur distribution is used to distinguish between informative and uninformative principal components.

single predictor, we augment the model with additional daily, weekly and monthly proxies of volatility. To that end, we make use of historical volatility proxies, which we calculate for the monthly and weekly horizons, as well as several proxies for daily volatility. We compare and contrast the different models to the original HAR.

As previous attempts at using the implied volatility to forecast realized volatility have mostly proved to be inconclusive, we consider a dimensionality reduction approach based on an implied volatility surface data set. Instead of a structurally agnostic method, we consider a simple decomposition based on the maturity and delta characteristics of the surface, which effectively reduces dimensionality, in a easily interpretable manner. Furthermore, we translate a similar idea to the implied volatility using the Heston model, whereby we use a calibration process to recover the underlying parameters of Heston and Bates models. One of those parameters, namely the current volatility, is a parallel to the implied volatility for the Black–Scholes model, albeit one that captures the price process more accurately (without the need of an implied volatility surface), providing an alternative way of utilizing the information embedded in the option prices. Lastly, similar to [248], we also consider a cross-sectional model, using information from different instruments as predictors of future volatility. We build this model based on the forecasts of the single assets models, which allow for a further boost in the predictive performance of the model.

Paper outline The rest of this paper is structured as follows; in Section 4.2, we define the realized volatility and various measures of historical volatility. In Section 4.3, we provide an overview of the implied volatility surface data and the dimensionality reduction methodology used to extract volatility estimators. In Section 4.4, we explain the methodology of extracting the option implied volatility using the Heston and Bates models. In Section 4.5, we set up the forecasting problem, and explain all candidate models. The empirical results are presented and analyzed in Section 4.6. Lastly, we draw conclusions in Section 4.7, and provide future avenues of research.

4.2 Volatility Estimators

4.2.1 Realized Volatility

Denote by $(S_t : t \geq 0)$, $(\mu_t : t \geq 0)$ and $(\sigma_t : t \geq 0)$ the price, mean, and instantaneous volatility processes corresponding to an underlying instrument. Under the stochastic

volatility family of models, the price process is driven by

$$\frac{dS_t}{S_t} = \mu_t dt + \sigma_t dB_t, \quad (4.1)$$

where $\{B_t : t \geq 0\}$ denotes a standard Brownian motion process.

The integrated variance is the theoretical value of the total variability occurred over the interval $[t_1, t_2]$, calculated as

$$\text{IV}_{(t_1, t_2)} = \int_{t_1}^{t_2} \sigma_t^2 dt. \quad (4.2)$$

As this is a theoretical quantity, there exist a number of different estimators to approximate it. By far the most common quantity used to approximately measure the integrated variance is the realized variance, which is given for the interval $[t_1, t_2]$ and a sample distance of δt , and is calculated as

$$\text{RV}_{(t_1, t_2, \delta t)} = \sum_{t=t_1+\delta t}^{t_2} \log \left(\frac{S_t}{S_{t-\delta t}} \right)^2. \quad (4.3)$$

It has been shown by [23], that assuming no drift ($\mu_t = 0$) the realized variance follows

$$\text{RV}_{(t_1, t_2, \delta t)} \sim N(V_{(t_1, t_2, \delta t)}, 2\text{IQ}_{(t_1, t_2)} \delta t), \quad (4.4)$$

where

$$\text{IQ}_{(t_1, t_2)} = \int_{t_1}^{t_2} \sigma_t^4 dt, \quad (4.5)$$

is the integrated quarticity. Therefore, realized variance converges in probability to integrated variance

$$\delta t \rightarrow 0 \text{RV}_{(t_1, t_2, \delta t)} = \text{IV}_{(t_1, t_2)}. \quad (4.6)$$

Another very popular estimator for integrated variance is the Bipower Variation (BPV) introduced by [24], and calculated as

$$\text{BPV}_{(t_1, t_2, \delta t)} = \sum_{t=t_1+\delta t}^{t_2-\delta t} \left| \log \left(\frac{S_t}{S_{t-\delta t}} \right) \right| \left| \log \left(\frac{S_{t+\delta t}}{S_t} \right) \right|. \quad (4.7)$$

Compared with realized variance, BPV is more robust to price jumps and is thus used as an additional feature as a proxy for integrated variance.

Our goal for this paper is to forecast the daily realized volatility (t_1 and t_2 corresponding to market open and close, respectively), based on 1-minute intraday returns ($\delta t = 1$ minute). Denote by $S_{d,1}, \dots, S_{d,391}$ the intraday mid-prices at 9:30, 9:31, $\dots$, 16:30 respectively for day d . The 1-minute realized variance is then calculated as

$$\text{RV}_d = \sum_{j=2}^{391} \log \left(\frac{S_{d,j}}{S_{d,j-1}} \right)^2. \quad (4.8)$$

4.2.2 Heterogeneous Autoregressive (HAR) Model

One of the most popular volatility forecasting models is HAR, an additive model that relies on the previous daily, weekly and monthly variances as predictors for the next day realized variance. Initially formulated by [71], it takes the following form

$$\text{RV}_d = \beta_0 + \beta_1 \text{RV}_{d-1} + \beta_2 \left(\frac{1}{5} \sum_{j=1}^5 \text{RV}_{d-j} \right) + \beta_3 \left(\frac{1}{22} \sum_{j=1}^{22} \text{RV}_{d-j} \right). \quad (4.9)$$

In contrast to other models capturing the long-memory property, HAR does not make any strong assumptions for the underlying price process. Despite its simplicity, HAR has matched or even outperformed more complex models, making it a standard benchmark for volatility forecasting.

The use of a simple autoregressive model for realized volatility, such as HAR, can entail a misspecification when assuming Equation (4.4), as it can cause an attenuation bias, that arises from the presence of measurement errors in volatility. Instead, [36] considers evolving coefficients which are linearly dependent on the square root of realized quarticity

$$\text{RQ}_{(t_1, t_2, \delta t)} = \frac{t_2 - t_1}{3\delta t} \sum_{t=t_1+\delta t}^{t_2} \log \left(\frac{S_t}{S_{t-\delta t}} \right)^4, \quad (4.10)$$

which is a consistent estimate of the integrated quarticity in Equation (4.2). In particular, the HARQ model of [37], is formulated as

$$\text{RV}_d = \beta_0 + \left(\beta_1 + \beta_{1,0} \sqrt{\text{RQ}_{d-1}} \right) \text{RV}_{d-1} + \beta_2 \left(\frac{1}{5} \sum_{j=1}^5 \text{RV}_{d-j} \right) + \beta_3 \left(\frac{1}{22} \sum_{j=1}^{22} \text{RV}_{d-j} \right), \quad (4.11)$$

and is shown to achieve improvements over the HAR model. While β_2 and β_3 can also be modeled as being non-constant and dependent on realized quarticity in the same manner as β_1 , [71] shows that the estimation errors from such misspecifications are on a much smaller magnitude than for β_1 , rendering Equation (4.11) a sufficiently good model. Our proposed models are extensions of the HAR model. To incorporate additional features representing the weekly and monthly variances, we consider a number of historical volatility estimators, as shown in Section 4.2.3. Furthermore, we incorporate additional variance estimates at the daily level, including the BPV, and volatility proxies calibrated from option data.

4.2.3 Historical Volatility Estimators

As features, we consider a number of additional volatility proxies, which are based on the historical volatility of the past L days. We denote the opening, closing, highest and lowest prices for day d as follows

$$\begin{aligned} O_d &= S_{d,1} \\ C_d &= S_{d,391} \\ H_d &= \max(S_{d,1}, \dots, S_{d,391}) \\ L_d &= \min(S_{d,1}, \dots, S_{d,391}) \end{aligned} \tag{4.12}$$

One of the most basic estimators used is the standard historical close-close volatility (HV) estimator, calculated as

$$\text{HV}_d^{(L)} = \sqrt{\frac{1}{L-1} \sum_{l=0}^{L-1} \log\left(\frac{C_{d-l}}{C_{d-1-l}}\right)^2}. \tag{4.13}$$

As this only takes into account the closing prices, and omits the intraday variation, volatility is severely underestimated. To this end, we additionally consider some of the so-called *advanced volatility estimators*, which also take into account the opening, high and low values within a day. One of the earliest advanced volatility estimators was the high-low estimator of Parkinson (HLP) formulated in [190], which incorporates the high and low prices as follows

$$\text{HLP}_d^{(L)} = \sqrt{\frac{\sum_{l=0}^{L-1} \left[\log\left(\frac{H_{d-l}}{L_{d-l}}\right) \right]^2}{4\log(2)L}}. \tag{4.14}$$

The Garman–Klass (GK) volatility estimator [97] was formulated as an extension of the Parkinson estimator that also includes the opening and closing prices. It is calculated as

$$\text{GK}_d^{(L)} = \sqrt{\frac{1}{L} \sum_{l=0}^{L-1} \left\{ \frac{1}{2} \left[\log\left(\frac{H_{d-l}}{L_{d-l}}\right) \right]^2 - (2\log(2) - 1) \left[\log\left(\frac{C_{d-l}}{O_{d-l}}\right) \right]^2 \right\}}. \tag{4.15}$$

Both HV, HLP and GK estimators all assume zero drift in the returns process. An extension which is able to accommodate for non-zero drifts is the volatility estimator of Rogers–Satchell (RS) [198], formulated as

$$\text{RS}_d^{(L)} = \sqrt{\frac{1}{L} \sum_{l=0}^{L-1} \left[\log\left(\frac{H_{d-l}}{C_{d-l}}\right) \log\left(\frac{H_{d-l}}{O_{d-l}}\right) + \log\left(\frac{L_{d-l}}{C_{d-l}}\right) \log\left(\frac{L_{d-l}}{O_{d-l}}\right) \right]}. \tag{4.16}$$

Note that none of the estimators discussed so far take into account the overnight jumps, which leads them to underestimate the real volatility. The Yang–Zhang extension of the GK model (GKYZ), does take into account the overnight jumps and takes the form

$$\text{GKYZ}_d^{(L)} = \sqrt{\frac{1}{L} \sum_{l=0}^{L-1} \left\{ \frac{1}{2} \left[\log \left(\frac{O_{d-l}}{C_{d-l-1}} \right) \right]^2 + \frac{1}{2} \left[\log \left(\frac{H_{d-l}}{L_{d-l}} \right) \right]^2 - (2\log(2) - 1) \left[\log \left(\frac{C_{d-l}}{O_{d-l}} \right) \right]^2 \right\}} \quad (4.17)$$

We also consider the Yang–Zhang estimator (YZ) which also includes a term for the overnight jumps by combining the RS estimator with the overnight and intraday volatilities.

$$\text{YZ}_d^{(L)} = \sqrt{(1 - k^{(L)})\text{RS}_d^{(L)} + k^{(L)}\sigma_c^2 + \sigma_o^2}, \quad (4.18)$$

where

$$k^{(L)} = \frac{\alpha - 1}{\alpha + \frac{L+1}{L-1}}, \text{ with } \alpha = 1.34, \quad (4.19)$$

$$\sigma_o^2 = \text{Var} \left(\left\{ \log \left(\frac{O_{d-(L-j)}}{C_{d-(L-1-j)}} \right) \right\}_{j=0}^{L-1} \right), \quad (4.20)$$

$$\sigma_c^2 = \text{Var} \left(\left\{ \log \left(\frac{C_{d-(L-j)}}{O_{d-(L-j)}} \right) \right\}_{j=0}^{L-1} \right). \quad (4.21)$$

Note that the above weights and coefficients are in most cases selected to maximize the metric of *efficiency*, computed as

$$\frac{\text{Volatility of Standard Estimator}}{\text{Volatility of Estimator}}, \quad (4.22)$$

where “Volatility of Standard Estimator” refers to the volatility of the standard close–close estimator. The maximum theoretical efficiencies are given in Table 4.1, where the Yang–Zhang estimator attains the highest efficiency. The maximum theoretical efficiency reflects the reduction in the number of data points required to achieve a level of volatility under an idealized distribution.

The last feature we consider is VIX-Fix, a volatility proxy developed by Larry Williams [234]. This proxy has been shown to successfully replicate the VIX time series when constructed for SPX, and can be easily constructed from historical data, for any desired instrument, including other indices and stocks. It is calculated as

$$\text{WVF}_d = 100 \frac{\max(H_d, H_{d-1}, \dots, H_{d-21}) - L_d}{\max(H_d, H_{d-1}, \dots, H_{d-21})}. \quad (4.23)$$

Estimator	Efficiency
Standard close-close volatility (HV)	1
Parkinson (HLP)	5.2
Garman–Klass (GK)	7.4
Rogers–Satchell (RS)	8
Yang–Zhang extension of Garman–Klass (GKYZ)	8
Yang–Zhang (YZ)	14

Table 4.1: This table shows the efficiencies as always by Equation (4.22), for each of the volatility estimators. These values are taken by [242] and [60].

4.2.4 Path Dependent Volatility

Contrary to stochastic volatility (SV) models which have volatility as a separate stochastic process to returns, recent empirical evidence shows how volatility depends on the path of the price return process [255, 59]. To account for these effects, we incorporate the two features from [104]

$$R_{1,d} = \sum_{d_i \leq d} \text{Ker}(d - d_i; h_1, \alpha_1) r_{d_i}, \quad (4.24)$$

$$R_{2,d} = \sqrt{\sum_{d_i \leq d} \text{Ker}(d - d_i; h_2, \alpha_2) r_{d_i}^2}, \quad (4.25)$$

where $r_{d_i} = \log \left(\frac{S_{d_i,391}}{S_{d_i-1,391}} \right)$ denotes the daily close-close return for day d_i . The kernel Ker is selected to be a time shifted power law,² i.e $\text{Ker}(\tau; h, \alpha) = \frac{(\tau+h)^{-\alpha}}{\int_0^\infty (\tau+h)^{-\alpha} d\tau}$, where the delay helps ensure that the weights do not blow up for vanishing τ . A simple linear model using Equations (4.24) and (4.25) can be shown to achieve great predictive power for forecasting VIX [104].

4.3 Extracting Volatility Estimators from Implied Volatility Surface Data

4.3.1 The Black–Scholes model

The Black–Scholes model is a simple framework for option pricing, and is often used to extract the implied volatility from the observed prices. Under this model, the volatility of returns is assumed to be constant, denoted as σ_{BS} . Denote by r the

²The particular values of parameters we use, is the optimal configuration for VIX as given by [104] for the volatility of SPX-500. In particular, we use $\alpha_1 = 1.06, h_1 = 0.020, \alpha_2 = 1.60, h_2 = 0.052$ as well as a look-back window of 1000 days.

interest rate and q the dividend yield of the underlying. Under the risk-neutral probability measure $\mathbb{Q}$, the underlying return process is then driven by the diffusion

$$dS_t = (r - q)S_t dt + \sigma_{BS} dB_t. \quad (4.26)$$

Consider a vanilla European call option at time t^3 with a strike price of K , that expires at time T . The payoff is given by

$$C_{K,T,\sigma_{BS}}^{\text{BS-Call}}(t, S) = e^{-r(T-t)} \mathbb{E}^{\mathbb{Q}} [(S_T - K)^+ | S_t = S]. \quad (4.27)$$

In a Black-Scholes world where the price process is driven by Equation (4.27), its value is

$$C_{K,T,\sigma_{BS}}^{\text{BS-Call}}(t, S_t) = S_t e^{-q(T-t)} \Phi(d_1) - K e^{-r(T-t)} \Phi(d_2), \quad (4.28)$$

where

$$d_1 = \frac{\log\left(\frac{S_t}{K}\right) + \left(r - q + \frac{\sigma_{BS}^2}{2}\right)T}{\sigma_{BS}\sqrt{T-t}}, \quad d_2 = d_1 - \sigma_{BS}\sqrt{T-t},$$

and Φ denotes the cumulative distribution function of a standard normal variable. Using the call-put parity, it can be shown that a put option with the same maturity and strike price is given by

$$C_{K,T,\sigma_{BS}}^{\text{BS-Put}}(t, S_t) = K e^{-r(T-t)} \Phi(-d_2) - S_t e^{-q(T-t)} \Phi(-d_1). \quad (4.29)$$

The only unobserved variable is the volatility σ_{BS} . A value for this can be obtained by solving Equation (4.28) or Equation (4.29) using a simple binary search algorithm.

4.3.2 Implied Volatility Surface

The cost for the simplicity of the BS model is its inability to capture a time-evolving variance. Furthermore, a commonly observed phenomenon is the volatility smile, in which the implied volatility differs for different options products, both in terms of $\Delta = \partial C_{K,T}^{\text{BS-Call}} / \partial S$ and maturity $T - t$. By considering a fixed set of 10 maturity values ($\{30, 60, 91, 122, 152, 182, 273, 365, 547, 730\}$ days) and 28 values for Δ ($\{-0.75, -0.70, \dots, -0.15, -0.10, 0.10, 0.15, \dots, 0.70, 0.75\}$), a three dimensional object denoted as the *Implied Volatility Surface* (IVS) is formed, often visualized as a surface plot.⁴

³Note that the value of a European call option is equal to the value of an American call option for a stock that does not pay dividends. As we work with stocks that do not pay dividends, we are able to use the formula for European options.

⁴Note that the natural characteristics of an option product are maturity and strike price. However, Δ is used in place of the strike price as it is a normalized value, taking values in $[-1, 1]$, in contrast to the unconstrained strike price. In other words, Δ is a bounded proxy for *moneyness*, S_t/K .

An early method of constructing volatility surfaces is that of [91], who used a spline interpolation method. We instead use the data set by OptionMetrics [183], which is built using a kernel interpolation technique, which we summarize in the next paragraph.

We denote the implied volatility for day d , for the option contract with Δ equal to δ and maturity equal to m as $\hat{\sigma}_{\text{BS},m,\delta}^{(d)}$. As such options are not guaranteed to exist, the implied volatilities are computed for each delta-maturity pair using an interpolation method based on a kernel smoothing algorithm. Denote by N_t the number of different option products traded for day t , with implied volatilities $(\sigma_1, \dots, \sigma_{N_d})$, Vega's $(\mathcal{V}_1, \dots, \mathcal{V}_{N_d})$,⁵ expiration times $(T_1, \dots, T_{N_d})$ and Δ 's $(\Delta_1, \dots, \Delta_{N_d})$. We then calculate

$$\hat{\sigma}_{\text{BS},m,\Delta}^{(d)} = \frac{\sum_{j=1}^{N_t} \mathcal{V}_j \sigma_j K(\Delta_j - \Delta, \log(\frac{m_j}{m}), 1_{\text{sign}(\Delta)=\text{sign}(\Delta_j)})}{\sum_{i=1}^{N_t} \mathcal{V}_i K(\Delta_i - \Delta, \log(\frac{m_i}{m}), 1_{\text{sign}(\Delta)=\text{sign}(\Delta_i)})}, \quad (4.30)$$

where

$$K(x, y, z) = \frac{1}{2\pi} \exp\left(-\left[\frac{x^2}{2h_1} + \frac{y^2}{2h_2} + \frac{z^2}{2h_3}\right]\right),$$

and the choice of $h_1 = 0.05, h_2 = 0.005, h_3 = 0.001$.

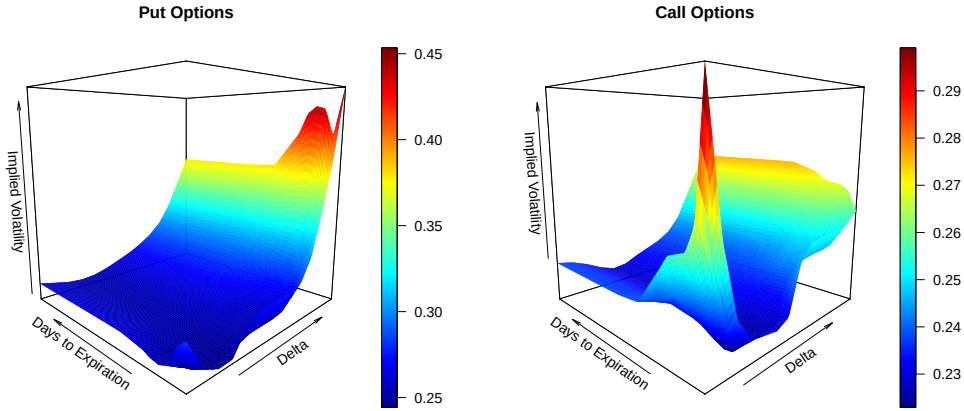


Figure 4.1: Example of Implied Volatility Surface for Triumph Group Inc (TGI) - 23-January-2013. We separate the plots for call and put options.

In Figure 4.1, we plot one example of an implied volatility surface (IVS), for the symbol Triumph Group Inc (TGI) for 23-January-2013. We choose this particular instrument-date pair as it features a number of common patterns observed for IVS. Firstly, put options have a higher implied volatility compared with the corresponding call options, as is the case in this example, where implied volatilities range from 0.25

⁵The Greek Vega captures the sensitivity of the option to volatility and is computed as $\mathcal{V} = \frac{\partial C_{K,T,\sigma_{\text{BS}}}^{\text{BS-Call}}}{\partial \sigma_{\text{BS}}}$.

to 0.45 for put options, and from 0.23 to 0.29 for call options. This stems from the behavior of investors shielding themselves against market crashes with put options. Secondly, the lowest implied volatilities are observed for ATM options (that is with $\Delta = 0.5$ or -0.5). Typically, IVS take the usual U-shape form, while in the cases shown in Figure 4.1, the implied volatility for put options is much higher for Δ close to 0 (OTM) and implied volatility for call options is much higher for Δ close to 1 (ITM). Lastly, there is a more subtle effect, where option products with shorter maturities have a higher implied volatility.

4.3.3 Dimensionality reduction for implied volatility surfaces

The dimensionality of implied volatility surfaces is quite high at 280. An early investigation for the relationship between implied and realized volatility is [47], who tests each of the implied volatility values of the surface separately, without finding consistent predictability. More recent research focused on the implied volatility of indices, includes [74], who showed a significant Granger causality between stock market implied volatility and stock volatility, and [231], who showed a connection between the implied volatility for the gold and silver markets to the corresponding futures in the Chinese market. To this end, Principal Component Analysis (PCA) and other dimensionality reduction methods have been used to reduce the dimensionality of the surfaces (see, e.g., [79]). Disadvantages of using a black-box dimensionality algorithm include that they are structure-agnostic and that any mispricing or faulty entries can have a disruptive effect on many of the principal components (PC). In this section, we consider an alternative method of reducing the dimensionality, by decomposing implied volatility using a simple decomposition. In particular, for each day d , we wish to extract the average level of the volatility $\hat{\sigma}_0^{(d)}$, a maturity effect for maturity m denoted as $\sigma_{M=m}^{(d)}$, and a delta effect for δ denoted as $\sigma_{\Delta=\delta}^{(d)}$. Therefore, this decomposition takes the form

$$\hat{\sigma}_{\text{BS},m,\delta}^{(d)} = \underbrace{\hat{\sigma}_0^{(d)} + \sigma_{M=m}^{(d)} + \sigma_{\Delta=\delta}^{(d)}}_{\text{Structural Component}} + \underbrace{\epsilon_{m,\delta}^{(d)}}_{\text{Residual Component}}. \quad (4.31)$$

Using this decomposition, we reduce the dimensionality from 280 down to $39 + p$. We select the parameters by minimizing

$$\sum_{\delta=1}^{|\Delta|} \sum_{m=1}^{|M|} \left(\epsilon_{m,\delta}^{(d)} \right)^2. \quad (4.32)$$

Equation (4.32) has a unique solution, when we set the natural constraints

$$\sum_{\delta=1}^{|\Delta|} \sigma_{\Delta=\delta}^{(d)} = 0, \sum_{m=1}^{|M|} \sigma_{M=m}^{(d)} = 0. \quad (4.33)$$

We solve Equation (4.32) by considering the Lagrangian

$$\mathcal{L} = \sum_{\delta=1}^{|\Delta|} \sum_{m=1}^{|M|} \left(\hat{\sigma}_{m,\delta}^{(d)} - \hat{\sigma}_0^{(d)} - \sigma_{\Delta=\delta}^{(d)} - \sigma_{M=m}^{(d)} \right)^2 + \nu_1 \sum_{m \in M} \sigma_{M=m}^{(d)} + \nu_2 \sum_{\delta \in \Delta} \sigma_{\Delta=\delta}^{(d)}. \quad (4.34)$$

The solution obtained via Equation (4.34) is given by

$$\hat{\sigma}_0^{(d)} = \frac{\sum_{m \in M} \sum_{\delta \in \Delta} \hat{\sigma}_{m,\delta}^{(d)}}{|\Delta||M|}, \quad (4.35)$$

$$\sigma_{\Delta=\delta}^{(d)} = \frac{\sum_{m \in M} \hat{\sigma}_{m,\delta}^{(d)}}{|M|} - \hat{\sigma}_0^{(d)}, \quad (4.36)$$

$$\sigma_{M=m}^{(d)} = \frac{\sum_{\delta \in \Delta} \hat{\sigma}_{m,\delta}^{(d)}}{|\Delta|} - \hat{\sigma}_0^{(d)}. \quad (4.37)$$

This decomposition of implied volatility provides a single value of the implied volatility $\hat{\sigma}_0^{(d)}$, with the rest of the features being directly interpretable, being the mean increase for each level of maturity or Δ .

4.4 Extracting Volatility Estimators under Stochastic Volatility Models

In this section, we describe how to extract a volatility estimator using the Heston model, via calibrating its parameters. As we are only interested in tracking the current volatility and its changes, the other parameters can be fixed. We first describe the Heston model, discuss the sensitivity of option prices to its parameters, and briefly explain the calibration process. We then justify the decision to fix some of its parameters, discuss an extension to the Bates model, and lastly provide some implementation details.

4.4.1 Heston Model

Unlike the Black–Scholes model, volatility under the Heston model is not assumed to be constant. Instead, volatility is a separate random process σ_t^2 , driven by a diffusion

process which is correlated with the price process S_t . In particular, the dynamics are described by the diffusions

$$dS_t = \mu S_t dt + S_t \sigma_t dB_t, \quad (4.38)$$

$$d\sigma_t^2 = \kappa(\bar{V} - \sigma_t^2)dt + \sigma_V \sigma_t dB_t^1, \quad (4.39)$$

where B , B^1 are standard Brownian motion processes with correlation ρ , i.e. $\text{Cov}(dB_t, dB_t^1) = \rho$. Overall, the Heston model is governed by the following five parameters: the mean reversion speed of volatility (κ); long-term volatility average ($\bar{V}$); volatility of volatility (σ_V); correlation between the two diffusion processes (ρ); the current/initial volatility at time t ($\sigma_{t,0}$).

4.4.2 Sensitivity analysis of the Heston model parameters

In later sections, we explain that precise calibration of all parameters in the Heston model isn't necessary for deriving a meaningful volatility estimator. Certain parameters influence the model's behavior more than others. To identify these key drivers, we initially assess the sensitivity of option prices to each of the five parameters.

We consider the influence of each parameter for four different times to expiration ($T-t \in \{5, 15, 90, 500\}$ days) and four different strike prices ($K \in \{0.75, 0.9, 1.1, 1.25\}$) for a theoretical instrument with current spot price $S = 1^6$, and setting both $r = 0$ and $q = 0$. We consider the parameter sets $\Theta_\kappa, \Theta_{\sigma_{t,0}^2}, \Theta_{\bar{V}}, \Theta_{\sigma_V}, \Theta_\rho$, each containing five candidate parameter values.

We calculate the set of sensitivities for κ as

$$I_\kappa(T, K) = \{|C_{T,K,(\kappa_1, \sigma_{t,0}, \bar{V}, \sigma_V, \rho)}^{\text{Heston-Call}}(t, S_t) - C_{T,K,(\kappa_2, \sigma_{t,0}, \bar{V}, \sigma_V, \rho)}^{\text{Heston-Call}}(t, S_t)| : \\ (\kappa_1, \kappa_2, \sigma_{t,0}, \bar{V}, \sigma_V, \rho) \in \Theta_\kappa \times \Theta_{\sigma_{t,0}^2} \times \Theta_{\bar{V}} \times \Theta_{\sigma_V} \times \Theta_\rho\}, \quad (4.40)$$

and define $I_{\sigma_{t,0}}(T, K)$, $I_{\bar{V}}(T, K)$, $I_{\sigma_V}(T, K)$, $I_\rho(T, K)$ in a similar manner.

In Figure 4.2, we plot a grid of box and whisker plots showing the set of sensitivities for each of the parameters. For low maturity options, the current volatility $\sigma_{t,0}$ is the most influential parameter, while the long term volatility $\bar{V}$ and mean reversion speed κ have a reduced significance. For higher maturities, the long term volatility $\bar{V}$ becomes the most influential parameter instead. For a maturity period of 500 days,

⁶Note that the implied volatility surface data used is constructed based on maturity and Delta, while the grid created for the sensitivity analysis is based on maturity and strike price, which is more suitable for demonstration purposes. As the price is fixed at $S = 1$, there is a 1-1 relationship between strike prices and deltas.

κ also becomes a more influential parameter than $\sigma_{t,0}$. $\sigma_{t,0}$ also appears to have more influence for options that are closer to being ATM, than options with $K = 0.75$ or 1.25 . Crucially, it is evident that the price primarily reacts to only three parameters: $\sigma_{t,0}$, $\bar{V}$, and κ . ρ and σ_V exhibit noticeably lower influence across all maturities and strike prices, with median sensitivities being close to zero for all maturity/strike price combinations.

4.4.3 Full Calibration of the Heston Model

Unlike the Black–Scholes model, which can be calibrated with one market price and a simple binary search, the Heston model’s parameter space is five-dimensional. In particular, at any time t , the parameter vector $\theta = (\kappa, \sigma_{t,0}, \bar{V}, \sigma_V, \rho)$ needs to be determined. A semi-closed solution can be obtained for vanilla option prices [111], with the price of a European call option being given by

$$C_{T,K,\theta}^{\text{Heston-Call}}(t, S_t) = S_t P_1(t, S_t, T, K, \theta) - e^{-r(T-t)} K P_2(t, S_t, T, K, \theta), \quad (4.41)$$

where

$$P_1(t, S_t, T, K, \theta) = \frac{1}{2} + \frac{1}{\pi} \int_0^\infty \text{Re} \left(\frac{e^{-iu \log K}}{S_t i u e^{r(T-t)}} \phi^{(\text{Heston})}(\theta; u - i, T - t) \right) du, \quad (4.42)$$

$$P_2(t, S_t, T, K, \theta) = \frac{1}{2} + \frac{1}{\pi} \int_0^\infty \text{Re} \left(\frac{e^{-iu \log K}}{iu} \phi^{(\text{Heston})}(\theta; u, T - t) \right) du. \quad (4.43)$$

In the above, i corresponds to the imaginary unit and ϕ refers to the characteristic function of $\log(S_t)$. There are multiple formulations for ϕ , but we choose to follow [208], whose formulation gives numerically stable results:

$$\begin{aligned} \phi^{(\text{Heston})}(\theta; u, T - t) = \exp \left\{ iu (\log(S_t) + r(T - t)) + \frac{\kappa \bar{V}}{\sigma_V^2} [(\xi - g)(T - t) - \right. \\ \left. 2 \log \left(\frac{1 - g_2 e^{-g(T-t)}}{1 - g_2} \right) \right. \\ \left. + \frac{\sigma_{t,0}^2}{\sigma_V^2} (\xi - g) \frac{1 - e^{-g(T-t)}}{1 - g_2 e^{-g(T-t)}} \right\}, \quad (4.44) \end{aligned}$$

where

$$\xi = \kappa - \sigma_V \rho i u, \quad (4.45)$$

$$g = \sqrt{\xi^2 + \sigma_V^2 (u^2 + i u)}, \quad (4.46)$$

$$g_1 = \frac{\xi + g}{\xi - g}, g_2 = \frac{\xi - g}{\xi + g}. \quad (4.47)$$

Denote by N_d the number of different option products traded for day d , and denote their maturity times by $(T_1, \dots, T_{N_d})$, strike prices by $(K_1, \dots, K_{N_d})$, mid-quote prices by $(C_1, \dots, C_{N_d})$, traded volumes by $(\text{Volume}_1, \dots, \text{Volume}_{N_d})$ and quote spreads by $(\text{Spread}_1, \dots, \text{Spread}_{N_d})$. The vector of the residuals between theoretical and observed prices is given by

$$e = (e_1, \dots, e_{N_d}) = \left(C_{T_1, K_1, \theta}^{\text{Heston-Call}}(t, S_t) - C_1, \dots, C_{T_{N_d}, K_{N_d}, \theta}^{\text{Heston-Call}}(t, S_t) - C_{N_d} \right). \quad (4.48)$$

By choosing a suitable weighting scheme $\{w_i\}_{i=1}^{N_d}$ for the different option products (as discussed in the next section), we can then calibrate the Heston model using the following objective function

$$\min_{\theta=(\kappa, \sigma_{d,0}, \sigma_V, \bar{V}, \rho)} \sum_{i=1}^{N_d} w_i e_i = \min_{\theta=(\kappa, \sigma_{d,0}, \sigma_V, \bar{V}, \rho)} \sum_{i=1}^{N_d} w_i [C_i - C_{T_i, K_i, \theta}^{\text{Heston-Call}}(t, S_t)]^2. \quad (4.49)$$

A full description of how to explicitly compute the gradient of the right-hand side of Equation (4.49) is given by [72]. Alternatively, a difference method can be used to calculate each partial derivative. A solution can then be obtained using a Levenberg–Marquardt approach [173]. We carry out the full calibration using a combination of a gradient method and Particle Swarm Optimization. We provide further details in Appendix C.1.

4.4.4 Partial Calibration of the Heston Model

In the previous section, we provided details on how to fully calibrate the Heston model. However, this is generally a hard problem for a number of reasons. Firstly, the estimation of Equation (4.27) for the Heston model involves the estimation of the integrals in Equations (4.42) and (4.43), making it non-negligible in terms of computing complexity. Secondly, due to the non-convexity of the problem, there is no guarantee that a gradient descent approach will converge to the global minimum. In cases where the number of products is low, multiple parameter sets can yield equally good fits, especially when there are less than five option products. Finally, the success of gradient methods for the Heston model is known to be quite sensitive to the initial set of parameters [178].

The aforementioned reasons render the calibration process a difficult task. While we borrow methodology from the calibration literature, our goal here is inherently different. In our case, an exact calibration of the problem is not required, as we are only interested in obtaining a proxy of the volatility based on option data; in

other words, we only wish to track the evolution of $\sigma_{d,0}$. Therefore, we are content with keeping the other four parameters fixed, deviating from the standard calibration procedure.

Initially we run a full calibration algorithm as described in Appendix C.1. For each stock, we fix the values of ρ and σ_V to be equal to the median values of the calibrated parameters as obtained from the full calibration of each stock. We consider the ranges of values that κ and σ_V take. Between these extreme values we take a total of 5 values equally spaced on a logarithmic scale (as volatilities have highly skewed distributions). This creates a 5x5 grid of (κ, σ_V) pairs. For each of the pairs, we calibrate with respect to $\sigma_{d,0}$, while keeping the other four parameters fixed, by finding

$$\sigma_{\text{Heston}}^{(d)}(\kappa, \sigma_V) = \min_{\sigma_0} \sum_i w_i \left[C_i - C_{T_i, K_i, (\kappa, \sigma_{d,0}, \rho, \sigma_V, \bar{V})}^{\text{Heston-Call}}(d, S_d) \right]^2. \quad (4.50)$$

Once we calculate this grid of values for pairs $(\kappa, \sigma_V) \in \Theta_\kappa \times \Theta_{\sigma_V}$, we can then proceed to find the pair that minimizes the error function, and thus the value of volatility that corresponds to the minimum value of the loss function.

$$(\kappa^*, \sigma_V^*) = \arg \min_{\kappa \in \Theta_\kappa, \sigma_V \in \Theta_{\sigma_V}} \sum_i w_i \left[C_i - C_{T_i, K_i, (\kappa, \sigma_{\text{Heston}}^{(d)}(\kappa, \sigma_V), \rho, \sigma_V, \bar{V})}^{\text{Heston-Call}}(d, S_d) \right]^2, \quad (4.51)$$

where

$$\sigma_{\text{Heston}}^{(d)} = \sigma_{\text{Heston}}^{(d)}(\kappa^*, \sigma_V^*). \quad (4.52)$$

Secondly, we carry out PCA on a vectorized form of the 5x5 grid of volatility values, after subtracting $\sigma_{\text{Heston}}^{(d)}$ by all values. For all instruments investigated, we find that two principal components suffice to capture 95% of variability.

4.4.5 Partial Calibration of the Bates Model

We also consider the Bates model [26], which enhances the Heston framework by incorporating a jump-diffusion component, allowing for differentiation between high volatility and ad-hoc abrupt price movements. The Bates model is a combination of the Heston and Merton models. The dynamics of the volatility process under the Bates model is still described by Equation (4.39), similar to the Heston model. The price process is similar to Equation (4.38), but incorporates a jump component

$$dS_t = \mu S_t dt + S_t \sqrt{V_t} dB_t^1 + S_t (e^{J_t} - 1) dN_t, \quad (4.53)$$

where N_t is a Poisson process with intensity λ , and $J_t \sim N(\mu_J, \sigma_J)$. Therefore, in addition to the five parameters present in the Heston model, there are three additional parameters to be calibrated.

Once again, as we are only interested in tracking changes in volatility, an accurate calibration of the model is not required. In fact, it is natural to decouple the jumps from the rest of the process, by detecting the price jumps, prior to the calibration. While the field of intraday jump detection is vast and evolving, detection of daily jumps is a significantly easier task, as it is almost equivalent to the task of outlier detection. Thus, we use the following simple sliding window algorithm to determine whether day d is a jump point. Based on the L previous daily returns, we calculate the mean $\text{HM}_d^{(L)} = \sum_{l=0}^{L-1} \log \left(\frac{C_{d-l}}{C_{d-(l-1)}} \right)$ and standard deviation $\text{HV}_d^{(L)}$ (Equation (4.13)). If the z-score $\frac{|\log \left(\frac{C_{d-L}}{C_{d-(L-1)}} \right) - \text{HM}_d^{(L)}|}{\text{HV}_d^{(L)}}$ is greater than some pre-determined threshold, then the daily price process of day d is classified as having a jump.

Once all the jumps have been identified, we estimate λ, μ_J and σ_J for day d , based on the jumps detected in days $d - L', \dots, d - 1$, using the Maximum Likelihood estimates for the Poisson and Normal distributions, respectively. We then follow the same methodology as in Section 4.4.4, by using Equation (4.41). With the extended parameter set $\theta = (\kappa, \sigma_{d,0}, \sigma_V, \bar{V}, \rho, \lambda, \mu_J, \sigma_J)$, and the pre-calculated values of λ, μ_J and σ_J , we replace the characteristic function in Equation (4.44) with

$$\phi^{(\text{Bates})}(\theta; u, t) = \phi^{(\text{Heston})}(\theta; u, t) \exp \left\{ t\lambda \left((1 + \mu_J)^{iu} e^{\frac{iu}{2}(iu-1)\sigma_J^2} - 1 \right) \right\}. \quad (4.54)$$

4.4.6 Choice of Weighting Scheme

A common choice of weighting scheme is taking into account the option spread, by setting

$$w_j \propto \frac{1}{F(\text{Spread}_j)}, \quad (4.55)$$

where F is a monotonically increasing function, usually one of $\sqrt{(\cdot)}$, $|(\cdot)|$, or $(\cdot)^2$. It is shown by [177] that each of these weighting schemes achieves a lower error than a uniform scheme or schemes based on Vega. As we are dealing with daily data and our purpose is not an accurate calibration of the model, we take a slightly different approach. In particular, we use the following weighting scheme:

$$w_j \propto \exp \left(\frac{-|K_j - S_d|}{w_{\text{Moneyess}}} \right) \exp \left(\frac{-T_j}{w_{\text{Maturity}}} \right) \log(\text{Volume}_j). \quad (4.56)$$

This weighting is motivated by the sensitivity of the Heston parameters, as shown in Figure 4.2. The impact of $\sigma_{d,0}$ during the calibration is greater for products which are

close to expiration and which are close to being ATM. The parameters $w_{\text{Moneyiness}}$ and w_{Maturity} are empirically chosen to be 1 and 10 respectively, so that the normalized distributions of $\exp\left(\frac{-|K_j - S_d|}{w_{\text{Moneyiness}}}\right)$, $\exp\left(\frac{-(T_j - d)}{w_{\text{Maturity}}}\right)$ and $\log(\text{Volume}_j)$ have similar values for kurtosis.

4.5 Forecasting Setup

We aim to test whether augmenting a HAR-based baseline model with additional estimators of the daily, weekly and monthly variance can increase the forecasting performance. Note that as the distribution of daily realized variance has a high positive skew, we focus on forecasting $\log(\text{RV})$ instead of RV . For each instrument, we fit a distinct model, with features built using data from its own past historical price and option data, i.e of the form

$$\log(\text{RV}_{i,d}) = \mathcal{F}_i \left([X_{i,d-1}^{\text{Base}}, X_{i,d-1}^{\text{Augm}}] \right),$$

where $\text{RV}_{i,d}$ denotes the realized variance, and $X_{i,d}^{\text{Base}}(X_{i,d}^{\text{Augm}})$ denotes the baseline (additional) matrix of predictors used to forecast it, for the i^{th} instrument on the d^{th} day. In Figure 4.4, we show the different augmented feature sets we investigate in this section. Other than the natural extension of the HAR model, which is the HARQ model Equation (4.11), this includes the additional daily, weekly and monthly volatility estimators, as well as additional predictors which are commonly used in the literature.

In this chapter we setup the forecasting task, starting with a short note in Section 4.5.1 on the data set used. We then describe the evaluation metrics used in Section 4.5.2, and the fitting methods used in Section 4.5.3. In Section 4.5.4, we focus on how to utilize the implied volatility features described in Section 4.3. Lastly, in Section 4.5.5, we describe all the different augmented feature sets.

4.5.1 Data Set

In this study, we utilize data from OptionMetrics covering the period 2013–2019. In particular, we use the implied volatility surfaces, option prices, and option volume data sets from OptionMetrics. We also use intraday minutely data for equities, extracted from LOBSTER (add ref), which is used to calculate daily volatility estimators and various intraday features. Lastly, we use daily summary data for equity volumes from CRSP (add reference WRDS). In total, we consider a selection of 66

NYSE stocks that do not pay dividends and for which the described data is available for the specified period.

4.5.2 Evaluation Metrics

We use the following evaluation metrics to benchmark the forecasting models under consideration. Denote the vector of forecasted values by $\hat{y}$, and the observed vector by y , which in the context of this paper mainly refers to realized volatility. We consider the following suite of standard evaluation metrics.

Root Mean Squared Error (RMSE):

$$\text{RMSE} = \sqrt{\frac{1}{D} \sum_{d=1}^D (y_d - \hat{y}_d)^2}. \quad (4.57)$$

Mean Absolute Error (MAE):

$$\text{MAE} = \frac{1}{D} \sum_{d=1}^D |y_d - \hat{y}_d|. \quad (4.58)$$

Spearman Correlation: The Spearman Correlation between $\hat{y}$ and y is the Pearson correlation between the ranks of the two vectors. This allows this metric to capture non-linear monotonic relations. The Pearson correlation between $\hat{y}$ and y is given by

$$C = \frac{\sum_{d=1}^N (y_d - \bar{y})(\hat{y}_d - \bar{\hat{y}})}{\sqrt{\sum_{d=1}^N (y_d - \bar{y})^2 (\hat{y}_d - \bar{\hat{y}})^2}}. \quad (4.59)$$

Quasi-Likelihood (QLIKE):

$$\text{QLIKE} = \frac{1}{D} \sum_{d=1}^D \frac{\exp y_d}{\exp \hat{y}_d} - (y_d - \hat{y}_d) - 1, \quad (4.60)$$

We also directly compare models in a hypothesis testing framework. Consider $\hat{y}^{(i)}$, the forecasts generated by model i . We compute the loss differential for day d between models i and j as

$$\text{LD}_d(i, j) = \text{LOSS}(\hat{y}_d^{(i)}, y_d) - \text{LOSS}(\hat{y}_d^{(j)}, y_d). \quad (4.61)$$

To test whether model j outperforms the baseline model i , we use the hypothesis

$$H_0 : \mathbb{E}[\text{LD}_d(i, j)] = 0, \quad (4.62)$$

For that purpose, we use the one-tailed **Diebold–Mariano (DM) test**, a popular variant of the paired t-test introduced in [76], which tests the null hypothesis against the alternative

$$H_1 : \mathbb{E} [\text{LD}_d(i, j)] > 0. \quad (4.63)$$

We also utilize the **Model Confidence Set (MCS)** procedure by [107], which is used to identify a subset $\mathcal{M}^*$ of models from a candidate set $\mathcal{M}_0 = \{M1, M2, \dots\}$ that are statistically indistinguishable from the best model. In other words, we are looking for the set $\mathcal{M}^* \subset \mathcal{M}_0$ defined as

$$\mathcal{M}^* = \{i \in \mathcal{M}_0 : \mathbb{E} [\text{LD}_d(i, j)] \leq 0 \quad \forall j \in \mathcal{M}_0\}. \quad (4.64)$$

For each pair of models, we calculate

$$T_{i,j} = \frac{\text{LD}_d(i, j)}{\text{Var}(\{\text{LD}_d(i, j)\}_{d=1}^D)}. \quad (4.65)$$

Once we estimate $T_{i,j}$ for all pairs of models, the one-tailed test statistic for each model is given by

$$T_i = \max_{j \neq i} T_{i,j}. \quad (4.66)$$

MCS is based on sequential testing, where models that perform significantly worse than others are eliminated. In particular, at step n , a p-value p_i is computed for each model by comparing T_i to a bootstrap sample. If some models have p-values smaller than a critical value, the model with the greatest test statistic is removed by defining a new candidate set

$$\mathcal{M}_n = \mathcal{M}_{n-1} / \left\{ \underset{j \in \mathcal{M}_{n-1}}{\text{argmax}} T_j \right\}. \quad (4.67)$$

Once no further models can be removed from the candidate set, the remaining set constitutes the Model Confidence Set, which includes models that are not significantly different from the best model.⁷

4.5.3 Regression Methods

As the number of features we use is relatively large and the features are highly correlated, OLS regression would be inadequate. Deep learning methods are also inadequate, as the number of available data points is quite low (as this is a daily forecasting problem). We consider three widely used models: Elastic Nets, LASSO,

⁷Note that for both MCS and the DM test, we use the QLIKE function, with a significance level of 5%.

and XGBoost, which we shortly describe. Considering three different methods, acts as a robustness check for the claim that augmenting the HAR model with additional features results in an increase in performance.

The Elastic Net (EN) is a linear regression method, which alters the objective function of Ordinary Least Squares regression to simultaneously include both ℓ_1 and ℓ_2 norms, to penalize overfitting as follows

$$\hat{\beta} = \operatorname{argmin}_{\beta} \|y - X\beta\|^2 + \lambda \left(\alpha \|\beta\|_1 + (1 - \alpha) \frac{\|\beta\|_2^2}{2} \right), \quad (4.68)$$

where β is the parameter vector and X is the feature matrix. Here, λ and α are hyperparameters chosen using cross-validation.⁸ LASSO (Least Absolute Shrinkage and Selection Operator) is a special case of the EN with α set to 1 (so that only the ℓ_1 norm appears in the penalty term).

Boosting is a class of ensemble models where the weak learners usually take the form of decision trees. At each iteration, a new weak learner is fitted, whose training phase focuses on the previously misclassified data points. The eXtreme Gradient Boosting (XGBoost) is one of the most widely used variants of boosting.

4.5.4 Incorporating IVS features

While implied volatility is clearly related to realized volatility, using it in practice can be challenging and often fails to be a good predictor for realized volatility. In particular, a different set of option products may be traded each day, with a multitude of expiration dates and strike prices. Therefore, using implied volatility surfaces creates a set of 280 features which are directly comparable across days. However, this leads to a significant increase in dimensionality, with many features being highly correlated. Furthermore, in many instances, the IVS are calculated using only a handful of option products which were traded on a particular day, thus rendering the IVS a data set of lower intrinsic dimensionality.

A natural solution is to perform dimensionality reduction, such as Principal Component Analysis (PCA).⁹ In Section 4.3.3, we explored a decomposition that effectively reduces dimensionality from 280 to 39. We set up a total of five feature sets

⁸We use a 60-10-30 split rule for training, validation, and testing.

⁹Denote by $C(X)$ the covariance matrix of X , and its eigen-decomposition by $C(X) = U\Lambda U^T$, where Λ is a diagonal matrix capturing the $\lambda_1 \geq \dots \geq \lambda_{\operatorname{col}(X)} \geq 0$ eigenvalues of $C(X)$, and the columns of U are given by the $u_1, \dots, u_{\operatorname{col}(X)}$ orthonormal eigenvectors (in the same order as their corresponding eigenvalues), where $\operatorname{col}(X)$ denotes the number of columns of matrix X . The principal component scores are the columns of XU , with each new component's variance given by its corresponding eigenvalue (component i has variance λ_i). Selecting the first p components retains a proportion of the total variance in the feature matrix equal to $\sum_{i=1}^p \lambda_i / \sum_{i=1}^{\operatorname{col}(X)} \lambda_i$. Given the

that use either of the two methods or a combination of both. We describe each feature set by its additional features X^{Augm} .

- Complete (IVS-1): This feature set includes all 280 features of the implied volatility surface in vectorized form, i.e with no regard to the IVS structure.
- Complete-PCA (IVS-2): Here we perform PCA on the IVS matrix of dimensionality 280. We choose a dimensionality of $p = 7$ as it is enough to retain 90% of the variance for 94% of instruments.
- Decomposed (IVS-3): This set composes of all of the fixed maturity and delta effects obtained using the methodology explained in Section 4.3.3, which reduces dimensionality down to $p = 39$.
- Decomposed-PCA (IVS-4): For this set, we perform PCA to reduce the dimensionality the matrix containing all of the maturity and delta effects $\sigma_{M=m_1}^{(d)}, \dots, \sigma_{M=m_{10}}^{(d)}, \sigma_{\Delta=\delta_1}^{(d)}, \dots, \sigma_{\Delta=\delta_{28}}^{(d)}$. With just six components, 90% of variance is retained for 96% of instruments. With the addition of $\sigma_0^{(d)}$, dimensionality is reduced to $p = 7$.
- Decomposed-Separate PCA (IVS-5): Unlike IVS-4, we use PCA on each maturity and delta effects separately. In particular, we use the first two principal components built using the maturity effects $\sigma_{M=m_1}^{(d)}, \dots, \sigma_{M=m_{10}}^{(d)}$, the first four principal components built using the delta effects $\sigma_{\Delta=\delta_1}^{(d)}, \dots, \sigma_{\Delta=\delta_{28}}^{(d)}$, and the IV level $\sigma_0^{(d)}$. Using these features is sufficient for retaining 90% of the variance for 91% of the instruments.

For each set, we compute the reconstruction error. Note that for IVS-1, it is equal to zero by default. For IVS-2, this is calculated as the reconstruction error for PCA.¹⁰ For IVS-3 we calculate the reconstruction error as

$$\sqrt{\sum_{d=1}^D \sum_{m \in M} \sum_{\delta \in \Delta} \left(\hat{\sigma}_{m,\delta}^{(d)} - \hat{\sigma}_0^{(d)} - \sigma_{\Delta=\delta}^{(d)} - \sigma_{M=m}^{(d)} \right)^2}. \quad (4.70)$$

proportion of variance to retain $\delta_{\text{prop}} \in (0, 1)$, we set

$$p = \min \left\{ i : \frac{\sum_{k=1}^i \lambda_k}{\sum_{k=1}^d \lambda_k} \geq \delta_{\text{prop}} \right\}. \quad (4.69)$$

¹⁰We can reconstruct X using the first p eigenvectors as $\tilde{X} = XU U^T$. The reconstruction error can then be calculated as $\|X - \tilde{X}\|_F$, where $\|\cdot\|_F$ denotes the Frobenius norm.

	Dimension (p)	Reconstruction Error			
		LQ	Median	Mean	UQ
IVS-1	280	0	0	0	0
IVS-2	7	9.524	13.995	16.954	21.303
IVS-3	39	4.253	7.100	7.834	10.478
IVS-4	7	5.483	8.109	8.967	11.272
IVS-5	7	5.263	7.745	8.799	11.543

Table 4.2: Comparison of dimensionality reduction methods by reconstruction error. The reconstruction error is calculated as the Frobenius norm of the difference between the reconstructed surface and the original surface. The median, mean, lower quantile (LQ) and upper quantile (UQ) across instruments are given.

	MSE		MAE		R^2		QLIKE		DM test	MCS
	Mean	% of improv.	Mean	% of improv.	Mean	% of improv.	Mean	% of improv.	% of rejections	
IVS-1	0.7123	43	0.4905	43	0.2321	50	0.7656	57	16	
IVS-2	0.6995	57	0.4788	52	0.2487	55	0.7621	57	21	
IVS-3	0.7093	55	0.4852	54	0.2416	59	0.7635	57	18	
IVS-4	0.6989	62	0.4723	63	0.2520	64	0.7693	61	24	*
IVS-5	0.6996	57	0.4796	59	0.2496	59	0.7619	61	21	*

Table 4.3: Comparison of the performance of five different models making use of IVS data. All models are fitted using EN. For each model, we compare the performance in terms of the out-of-sample R^2 , Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). For each performance metric, we show the mean and the percentage of cases where the null model is outperformed. In addition, for models IVS-2, IVS-3, IVS-4, and IVS-5 we perform a Diebold–Mariano test comparing the residuals of each model with the residuals of the HAR model, which provides (in the last column) the number of instruments for which the null hypothesis is rejected (out of 66).

For IVS-4 and IVS-5 the calculation is done by first reconstructing $\{\sigma_{\Delta=\delta_i}^{(d)}\}_{i=1}^{28}$ and $\{\sigma_{M=m_i}^{(d)}\}_{i=1}^{10}$ and then calculating the error using Equation (4.70).

In Table 4.2, we show the reconstruction error for each of these models. Using PCA after the decomposition shown in Equation (4.1) performs better than directly applying PCA, as both IVS-4 and IVS-5 clearly outperform IVS-2. On the other hand, the performance of IVS-4 is only marginally better than IVS-5. While the reconstruction errors for both these sets are higher than that of IVS-3, the differences are rather small. We opt to use IVS-5, as it has the added benefit of its features being easier to interpret.

In Table 4.3, we directly compare the performance of five models based on each feature set, fitted using EN, compared to the baseline model described later in Section 4.5.5. Note that all of the models incorporate the features of a HAR model (daily,

weekly and monthly volatility) for this task. IVS-1 (using all 280 features of the IVS) is outperformed by the other sets across all metrics, thus confirming the benefit of performing a dimensionality reduction step. Furthermore, IVS-4 and IVS-5 achieve the highest performance, and significantly outperform the baseline model for more than 21 instruments. Furthermore, both appear in the MCS. On the other hand, IVS-4 and IVS-5 have very similar performances. Therefore, taking into account that the factors of IVS-5 are readily interpretable, renders IVS-5 as our feature set of choice.

Based on IVS-5, we plot the first two principal component loadings for maturity, and the first four principal loadings for Δ in Figure 4.5 for Adobe Inc. Note that these effects are mostly similar across most instruments, up to a difference in sign.

- *The first PC for maturity* assigns positive weights to low maturities (30, 60 and 91) and negative weights for high maturities, thus acting as the signed difference in volatility between short and long term options.
- In contrast to the first PC, *the second PC for maturity* has a mostly upwards trend, assigning positive weights to high maturities, and negative weights for low maturities. However, a very high positive weight is placed on the maturity of 30 days, thus substantially differentiating the importance of the short term (30-day) expiration options.
- *The 1st PC for Δ* assigns mostly positive weights to put options ($\Delta < 0$) and negative weights to call options ($\Delta > 0$), with the highest magnitudes being given for Δ 's close to 0. This essentially yields a signed difference in volatility between call and put options, with a focus on OTM options.
- In contrast to the first PC, *the 2nd PC for Δ* assigns mostly negative weights to put options ($\Delta < 0$) and positive weights to call options ($\Delta > 0$), with the highest magnitudes being assigned to Δ 's close to -1 or 1. This essentially gives a signed difference in volatility between call and put options with a focus on ITM options.
- *The 3rd PC for Δ* assigns mostly positive weights to options with Δ close to 0.5 or -0.5, and negative weights for ITM or OTM options. This effect is quite clear for put options, and less so for call options, where weights are smaller and the effect for $\Delta = 0.75$ is relatively high, which distorts the overall trend of the graph.

- The 4th PC for Δ assigns mostly positive weights to options with Δ close to 0 and negative weights to options with Δ close to -1 or 1. But unlike the first three PCs, this effect appears to have multiple turning points.

4.5.5 Description of Feature Sets

In this paper, we have described several categories of features in Sections 4.2.3, 4.3 and 4.4. We consider a number of different models built using one of the three methods described in Section 4.5.3 and one of the feature sets described in this subsection. In particular, we define models using the same features as HAR, baseline models (on for each method) based on HAR with some additional features, or a number of different features sets used to augment the baseline feature set. In this subsection, we define a number of feature sets used to build models to determine whether each set of features (X^{Augm}) provides additional forecasting power compared to a HAR-based baseline model.

M1: It consists of all the features included in a HAR model Equation (4.9), which is widely used as benchmark for volatility forecasting problems.

M2: M2 is used as the baseline model for each method, against which the other models are compared. It consists of all the features included in a HAR model Equation (4.9), as well as percentage change of realized volatility defined as $\frac{RV_d - RV_{d-1}}{RV_{d-1}}$. Furthermore, we include the value of the Volatility index (VIX) and percentage change (return) of VIX, as it was by shown [128] that it offers additional forecasting power (when forecasting the realized volatility of the stock market).

The following feature sets (M3-M13) include all of the features included in the baseline set (M2) as well as the set-specific features described in the subsequent paragraphs.

M3: As discussed, the natural extension of the HAR model is the HARQ model in Equation (4.11), i.e M3 also includes the realized quadricity feature.

M4: In Section 4.2, we have discussed additional daily estimators that we incorporate in a model denoted as $M3$. Importantly, this includes the Bipower Variation Equation (4.7). We also make use of the daily Parkinson (Equation (4.14)) and Garman–Klass (Equation (4.15)) estimators in the same manner as [32], i.e we set $L = 1$. Lastly, we apply the concept of realized semivariance [22]. In particular, the positive and negative semivariances are calculated as

$$RV_d^+ = \sum_{j=2}^{391} \log \left(\frac{S_{d,j}}{S_{d,j-1}} \right)^2 1(S_{d,j} \geq S_{d,j-1}). \quad (4.71)$$

$$RV_d^- = \sum_{j=2}^{391} \log \left(\frac{S_{d,j}}{S_{d,j-1}} \right)^2 1(S_{d,j} \leq S_{d,j-1}). \quad (4.72)$$

M5: In Section 4.2.3, we have discussed various historical volatility estimators. In *M5*, we include all weekly estimators including $HV^{(5)}$, $HLP^{(5)}$, $GK^{(5)}$, $RS^{(5)}$, $GKYZ^{(5)}$, $YZ^{(5)}$ as defined in Equations (4.14) to (4.18), with $L = 5$.

M6: In *M6*, we include all monthly estimators including $HV^{(21)}$, $HLP^{(21)}$, $GK^{(21)}$, $RS^{(21)}$, $GKYZ^{(21)}$, $YZ^{(21)}$ as defined in Equations (4.14) to (4.18), with $L = 21$. We also include VIX-Fix (Equation (4.23)).

M7: In Section 4.2.4, we have discussed the path dependent estimators used by [104]. Thus, we include these two features in feature set *M7*.

We also consider volume-based features, which have been previously reported in the literature to be associated with volatility, as discussed in Section 4.1. Denote by $Volume_d$ and OI_d the traded volume and open interest for day d respectively. Note that trading volumes are not stationary and tend to increase over time. To render the features stationary, we consider the percentage change of daily volume $\frac{Volume_d - Volume_{d-1}}{Volume_{d-1}}$, as well as the daily turnover $\frac{Volume_d}{OI_d}$.

M8: In *M7* we include both the percentage change of daily volume and the daily turnover those features for each corresponding equity instrument.

M9: For *M9* we consider the daily turnover and percentage change in volume for options. We also consider those quantities restricted to call and put options respectively.

M10: Feature set *M10* includes intraday features. In particular, we calculate the intraday returns and Realized Variance using only the last $l = 5, 10, 30, 60$ and 90 minutes prior to closing. That is $\log \left(\frac{S_{d,j}}{S_{d,j-l}} \right)$ and $\sum_{j=391-l}^{391} \log \left(\frac{S_{d,j}}{S_{d,j-1}} \right)^2$ for $l = 5, 10, 30, 60$ and 90 respectively, similar to [248]. We also include the sum of the positive minutely returns and the sum of the negative minutely returns respectively

$$\sum_{j=2}^{391} \max \left(\log \left(\frac{S_{d,j}}{S_{d,j-1}} \right), 0 \right) \text{ and } \sum_{d=2}^{391} \max \left(-\log \left(\frac{S_{d,j}}{S_{d,j-1}} \right), 0 \right). \quad (4.73)$$

A major focus of the paper is on formulating volatility features using option prices. In sets M11-M13 we consider the volatility estimators discussed in the paper.

M11: In Section 4.5.4, we explored a number of feature sets using the features built using the IVS data set. M11 incorporates the features of IVS-5.

M12: The features based on the Heston model described in Section 4.4, are included in *M12*, as well as the percentage differences for each feature.

M13: The features based on the Bates model described in Section 4.4, are included in *M13*, as well as the percentage differences for each feature.

M14 and **M15:** To compare the effect of features built using option prices we formulate *M14* to include all of the discussed features excluding those of *M11-M13*, and *M15* to include all features discussed (including *M11-M13*).

4.6 Results

In Table 4.4, we present the performance results of models based on M1-M15 and the three different methods. For each model, we report the mean RMSE, MAE, R^2 , and QLIKE. Using M2 as a baseline, we also provide the percentage of instruments for which the DM hypothesis is rejected, as well as the percentage of instruments that show an increase in performance for each metric.

For each instrument, we classify days into normal and high-volatility periods, where a high-volatility day is defined as having observed RV higher than at least 80% of the other days (i.e it is in the top quintile). In Table 4.5, we report the performance of each model separately for days with high and normal observed RV. We also report the classification accuracy, which is the accuracy in correctly forecasting whether volatility will be high or normal.

In the following paragraphs, we discuss the results for each of the considered data sets shown in Tables 4.4 and 4.5. In each paragraph, the effect of augmenting the HAR-based baseline model with each set of features is discussed.

Effect of Additional Daily Volatility Estimators (M3 and M4) In general, the evidence that HARQ is a significant improvement over the baseline models appears to be weak. For a large number of metrics, the percentage of improvements is close to or even below 50%, while the mean performance is below that of the baseline. This is quite different for M4, which is included in the MCS for both LASSO and EN. Many improvements can be seen across metrics, and in fact, for EN, M4 was an improvement over M2 for all instruments considered in terms of QLIKE. Augmenting the baseline model with additional volatility estimators thus appears to improve forecasting performance.

Effect of Additional Historical Volatility Estimators (M5, M6, and M7) The inclusion of weekly historical volatility estimators shows mixed results across the three models. For EN and LASSO, M5 fails to outperform the baseline in terms of

EN	RMSE		MAE		R^2		QLIKE		DM Test	MCS
	Mean	% of improv.	Mean	% of improv.	Mean	% of improv.	Mean	% of improv.	% of rejections	Included
Model 1 (M1) - HAR	0.6236	7	0.4787	9	0.2810	5	0.7713	61	46	
Model 2 (M2) - Baseline Model	0.5986	NA	0.4577	NA	0.3225	NA	0.7759	NA	NA	
Model 3 (M3) - HARQ	0.6022	53	0.4592	53	0.3164	58	0.7754	53	31	
Model 4 (M4) - Additional Daily Volatility Estimators	1.8763	30	1.8129	32	0.3005	41	0.7726	66	49	*
Model 5 (M5) - Historical Volatility Estimators (Weekly)	0.6362	41	0.4730	43	0.3013	46	0.7743	54	39	
Model 6 (M6) - Historical Volatility Estimators (Monthly)	0.9349	29	0.6741	34	0.2944	25	0.7697	46	37	
Model 7 (M7) - Path Dependence Features	0.6066	41	0.4619	46	0.3122	57	0.7747	54	43	
Model 8 (M8) - Equity Volume Features	0.5990	50	0.4577	52	0.3245	66	0.7753	57	41	
Model 9 (M9) - Option Volume Features	0.5989	59	0.4602	50	0.3284	73	0.7759	50	35	
Model 10 (M10) - Intraday Features	0.6006	66	0.4606	64	0.3359	73	0.7750	52	43	
Model 11 (M11) - IVS Volatility Features	0.6062	47	0.4695	44	0.3195	49	0.7750	60	50	*
Model 12 (M12) - Heston Model Volatility Features	0.5921	70	0.4542	43	0.3298	64	0.7772	43	32	
Model 13 (M13) - Bates Model Volatility Features	0.6203	54	0.4792	45	0.3207	55	0.7834	43	39	
Model 14 (M14) - All Features (excluding Option Features)	0.6204	55	0.4688	50	0.3256	62	0.7727	50	44	
Model 15 (M15) - All Features	0.6182	68	0.4769	64	0.3304	70	0.7703	66	52	*

LASSO	RMSE		MAE		R^2		QLIKE		DM Test	MCS
	Mean	% of improv.	Mean	% of improv.	Mean	% of improv.	Mean	% of improv.	% of rejections	Included
Model 1 (M1) - HAR	0.6237	7	0.4790	9	0.2810	5	0.7711	59	48	
Model 2 (M2) - Baseline Model	0.5990	NA	0.4581	NA	0.3224	NA	0.7758	NA	NA	
Model 3 (M3) - HARQ	0.6021	67	0.4588	64	0.3163	53	0.7755	51	33	
Model 4 (M4) - Additional Daily Volatility Estimators	0.9632	32	2.2640	34	0.3019	39	0.7730	62	46	*
Model 5 (M5) - Historical Volatility Estimators (Weekly)	0.6494	41	0.4754	46	0.2976	45	0.7744	55	43	
Model 6 (M6) - Historical Volatility Estimators (Monthly)	0.9449	29	0.6807	34	0.2931	25	0.7696	48	37	
Model 7 (M7) - Path Dependence Features	0.6068	45	0.4626	54	0.3119	50	0.7745	57	46	
Model 8 (M8) - Equity Volume Features	0.5978	60	0.4565	58	0.3247	70	0.7754	60	46	
Model 9 (M9) - Option Volume Features	0.5983	59	0.4594	46	0.3284	75	0.7760	41	31	
Model 10 (M10) - Intraday Features	0.5978	70	0.4583	64	0.3370	79	0.7756	45	36	
Model 11 (M11) - IVS Volatility Features	0.6062	46	0.4693	46	0.3195	52	0.7750	57	47	*
Model 12 (M12) - Heston Model Volatility Features	0.5925	70	0.4546	52	0.3298	66	0.7770	41	32	
Model 13 (M13) - Bates Model Volatility Features	0.6215	50	0.4803	43	0.3201	54	0.7835	46	40	
Model 14 (M14) - All Features (excluding Option Features)	0.6877	59	0.4831	57	0.3231	64	0.7716	52	39	
Model 15 (M15) - All Features	0.6119	68	0.4721	68	0.3297	71	0.7707	61	51	*

XGBoost	RMSE		MAE		R^2		QLIKE		DM test	MCS
	Mean	% of improv.	Mean	% of improv.	Mean	% of improv.	Mean	% of improv.	% of rejections	Included
Model 1 (M1) - HAR	0.6178	25	0.4596	23	0.1990	12	0.7214	73	44	
Model 2 (M2) - Baseline Model	0.6029	NA	0.4390	NA	0.2619	NA	0.7262	NA	NA	
Model 3 (M3) - HARQ	0.5840	59	0.4289	62	0.2628	54	0.7265	52	24	
Model 4 (M4) - Additional Daily Volatility Estimators	0.5962	59	0.4269	66	0.2795	70	0.7289	45	22	
Model 5 (M5) - Historical Volatility Estimators (Weekly)	0.5943	50	0.4379	43	0.2568	52	0.7282	43	18	
Model 6 (M6) - Historical Volatility Estimators (Monthly)	0.6228	18	0.4700	23	0.2364	30	0.7300	29	21	
Model 7 (M7) - Path Dependence Features	0.6098	36	0.4599	48	0.2528	45	0.7294	50	30	
Model 8 (M8) - Equity Volume Features	0.5993	48	0.4393	43	0.2617	57	0.7252	76	48	
Model 9 (M9) - Option Volume Features	0.5961	68	0.4348	59	0.2752	70	0.7268	55	40	
Model 10 (M10) - Intraday Features	0.5775	71	0.4276	66	0.2826	79	0.7310	30	18	
Model 11 (M11) - IVS Volatility Features	0.6094	37	0.4562	35	0.2382	41	0.7256	57	40	
Model 12 (M12) - Heston Model Volatility Features	0.6017	52	0.4413	45	0.2709	61	0.7300	41	21	
Model 13 (M13) - Bates Model Volatility Features	0.6011	61	0.4539	48	0.2758	62	0.7324	30	26	
Model 14 (M14) - All Features (excluding Option Features)	0.5891	61	0.4323	57	0.2874	75	0.7312	38	27	
Model 15 (M15) - All Features	0.5824	64	0.4201	61	0.3124	79	0.7213	69	43	

Table 4.4: Performance comparison for M1-M15. For each model, we compare the performance in terms of out-of-sample R^2 , Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Quasi Likelihood (QLIKE). For each model, the mean of each metric is given as well as the percentage of cases where that model outperforms M2. Lastly, we provide the number of instruments (out of the full universe of 66 instruments) that the null hypothesis of the DM was rejected, signifying an improvement. In red (blue) we highlight the highest (second highest) performing method.

EN	RMSE		MAE		R^2		QLIKE		Classification	MCS	
	Low	High	Low	High	Low	High	Low	High	Accuracy	Low	High
Model 1 (M1) - HAR	0.6078	1.5408	0.3918	0.9488	0.1635	0.0717	0.9098	2.3629	0.8054		
Model 2 (M2) - Baseline model	0.6072	1.4713	0.3910	0.8691	0.1997	0.0725	0.9440	2.9783	0.8256		
Model 3 (M3) - HARQ	0.5973	1.4693	0.3899	0.8726	0.1999	0.0687	0.9228	2.5686	0.8238		
Model 4 (M4) - Additional Daily Volatility Estimators	0.8157	2.5707	0.3990	2.9457	0.1904	0.0693	0.9223	2.7471	0.8183		
Model 5 (M5) - Historical Volatility Estimators (Weekly)	0.6084	1.5452	0.3925	0.9057	0.1962	0.0628	0.9221	2.5742	0.8209		
Model 6 (M6) - Historical Volatility Estimators (Monthly)	0.6130	1.9199	0.4005	1.2302	0.1872	0.0669	0.9234	2.5734	0.8103		
Model 7 (M7) - Path Dependence Features	0.5984	1.4794	0.3902	0.8775	0.1992	0.065	0.9328	2.7721	0.8207		*
Model 8 (M8) - Equity Volume Features	0.5977	1.4685	0.3900	0.8691	0.2017	0.0704	0.9329	2.7739	0.8267	*	
Model 9 (M9) - Option Volume Features	0.5984	2.5644	0.3913	1.3256	0.1984	0.0762	0.9395	2.9625	0.825		
Model 10 (M10) - Intraday Features	0.5962	1.4647	0.3872	0.8723	0.2148	0.0759	0.9530	3.1772	0.8235		
Model 11 (M11) - IVS Volatility Features	0.6073	1.4825	0.4014	0.8829	0.1906	0.0755	0.9238	2.5752	0.8271		*
Model 12 (M12) - Heston Model Volatility Features	0.6058	1.4529	0.3928	0.8609	0.1999	0.077	0.9658	3.3805	0.8277		
Model 13 (M13) - Bates Model Volatility Features	0.6378	1.4356	0.4197	0.8483	0.1867	0.0825	0.9638	3.1812	0.8163	*	
Model 14 (M14) - All Features (excluding Option Features)	0.5877	1.6079	0.3844	1.0018	0.2149	0.0744	0.9211	2.5590	0.8123		
Model 15 (M15) - All Features	0.6049	1.4835	0.3941	0.8897	0.2053	0.0795	0.9449	2.3570	0.8134		*

LASSO	RMSE		MAE		R^2		QLIKE		Classification	MCS	
	Low	High	Low	High	Low	High	Low	High	Accuracy	Low	High
Model 1 (M1) - HAR	0.6022	1.5415	0.3920	0.9495	0.1633	0.0718	0.9098	2.5698	0.8049		
Model 2 (M2) - Baseline model	0.6011	1.4722	0.3910	0.8698	0.1997	0.0719	0.9439	2.9781	0.8252		
Model 3 (M3) - HARQ	0.5977	1.4684	0.3901	0.8720	0.1998	0.0690	0.9228	2.5689	0.8232		
Model 4 (M4) - Additional Daily Volatility Estimators	0.8372	2.7903	0.4025	9.6149	0.1884	0.0713	0.9229	2.7483	0.8199		
Model 5 (M5) - Historical Volatility Estimators (Weekly)	0.6109	1.5496	0.3932	0.9068	0.1959	0.0626	0.9222	2.5741	0.8211		
Model 6 (M6) - Historical Volatility Estimators (Monthly)	0.6117	1.8778	0.3990	1.1940	0.1865	0.0660	0.9229	2.5727	0.8135		
Model 7 (M7) - Path Dependence Features	0.5987	1.4803	0.3902	0.8790	0.1986	0.0663	0.9328	2.7715	0.8207		*
Model 8 (M8) - Equity Volume Features	0.5981	1.4690	0.3903	0.8695	0.2012	0.0705	0.9331	2.7736	0.8263	*	
Model 9 (M9) - Option Volume Features	0.5992	2.5642	0.3920	1.3255	0.1978	0.0761	0.9398	2.9624	0.8247		
Model 10 (M10) - Intraday Features	0.5952	1.4675	0.3866	0.8757	0.2143	0.0753	0.9527	3.1755	0.8206		
Model 11 (M11) - IVS Volatility Features	0.6063	1.4820	0.4004	0.8818	0.1917	0.0755	0.9235	2.5752	0.8272		*
Model 12 (M12) - Heston Model Volatility Features	0.6063	1.4528	0.3931	0.8601	0.2000	0.0756	0.9659	3.3807	0.8272		
Model 13 (M13) - Bates Model Volatility Features	0.6383	1.4361	0.4203	0.8494	0.1869	0.0827	0.9639	3.1807	0.8177	*	
Model 14 (M14) - All Features (excluding Option Features)	0.7399	1.7555	0.3932	1.0370	0.2111	0.0723	0.9206	2.5588	0.8134		
Model 15 (M15) - All Features	0.6043	1.4632	0.3946	0.8676	0.2071	0.0774	0.9455	2.5042	0.8175		*

XGBoost	RMSE		MAE		R^2		QLIKE		Classification	MCS	
	Low	High	Low	High	Low	High	Low	High	Accuracy	Low	High
Model 1 (M1) - HAR	0.6077	1.5604	0.3959	0.9508	0.1663	0.0424	0.9114	2.3599	0.7968		
Model 2 (M2) - Baseline model	0.6113	1.5025	0.3982	0.8941	0.1806	0.0473	0.9444	2.9709	0.8080		
Model 3 (M3) - HARQ	0.6049	1.4828	0.3951	0.8822	0.1878	0.0488	0.9227	2.5641	0.8155	*	
Model 4 (M4) - Additional Daily Volatility Estimators	0.5973	1.9199	0.3922	3.0381	0.1959	0.0523	0.9241	2.7530	0.8142		
Model 5 (M5) - Historical Volatility Estimators (Weekly)	0.6063	1.5002	0.3967	0.8868	0.1841	0.0509	0.9236	2.5715	0.8168		
Model 6 (M6) - Historical Volatility Estimators (Monthly)	0.6350	1.5208	0.4188	0.8956	0.1597	0.0503	0.9298	2.5783	0.8049	*	
Model 7 (M7) - Path Dependence Features	0.6239	1.5073	0.4104	0.9011	0.1754	0.0527	0.9372	2.7656	0.7952		
Model 8 (M8) - Equity Volume Features	0.6066	1.5047	0.3963	0.9030	0.1814	0.0484	0.9322	2.7647	0.8063		
Model 9 (M9) - Option Volume Features	0.6065	6.5928	0.3965	1.3475	0.1804	0.0545	0.9385	2.9554	0.8079		
Model 10 (M10) - Intraday Features	0.6173	1.4524	0.4046	0.8504	0.1985	0.0601	0.9572	3.1784	0.8215		*
Model 11 (M11) - IVS Volatility Features	0.6246	1.5074	0.4097	0.9002	0.1612	0.0514	0.9244	2.5719	0.8061		*
Model 12 (M12) - Heston Model Volatility Features	0.6168	1.4877	0.4008	0.8871	0.1838	0.0544	0.9652	3.3740	0.8058		*
Model 13 (M13) - Bates Model Volatility Features	0.6320	1.4517	0.4124	0.8557	0.1698	0.0615	0.9588	3.1767	0.8070		
Model 14 (M14) - All Features (excluding Option Features)	0.6335	1.4595	0.4222	0.8598	0.1842	0.0602	0.9296	2.5695	0.8126		
Model 15 (M15) - All Features	0.6426	1.4552	0.4274	0.8527	0.1782	0.0699	0.9521	2.7808	0.8141		*

Table 4.5: Performance comparison for M1-M15, differentiating between normal and high days. For each model, we compare the performance in terms of the out of sample R^2 , Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Quasi Likelihood (QLIKE). The days in the test data set are separated into 'High' and 'Low' volatility and the average performance across each type of day is given for each of the metrics. Lastly, we give the average classification accuracy of each model. In red (blue) we highlight the highest (second highest) performing method per product.

RMSE and MAE but shows improved performance in terms of QLIKE, also being included in the MCS for EN. A similar pattern holds for monthly historical volatility estimators (M6), which performs significantly worse than the baseline for most metrics, with the exception of QLIKE. Notably, M6 significantly outperforms the baseline in terms of QLIKE and scores the best in terms of mean QLIKE. Adding path dependence features (M7) does not seem to result in significant improvements. The only exception is QLIKE for the LASSO model, where the mean is slightly lower than the baseline, with a total of 46 rejections of the DM test.

Effect of Volume Features (M8 and M9) The impact of equity volume is significant in many cases. For LASSO, M8 outperforms the baseline across all metrics, being the second-best model for both RMSE and MAE, as well as improving the R^2 for 70% of the instruments and achieving 46 rejections for the DM test for QLIKE. For XGBoost, it outperforms the baseline in 76% of the cases in terms of QLIKE, with a total of 48 rejections for the DM test. For option volumes (M9), results are mixed, with the most significant improvement appearing in terms of R^2 , as all three models show an increase of at least 70%.

Effect of Intraday Features (M10) An increase in performance is observed across most models and metrics, with the exception of QLIKE, where improvement rates are close to or below 50%. M10 appears to be the best or second-best performing model in several cases. It is the best-performing model in terms of RMSE, MSE, and R^2 for normal volatility days and in terms of classification accuracy when using a LASSO model.

Effect of Option Features (M11, M12, M13 and M15) The IVS features of M11 appear to only outperform the baseline in terms of QLIKE, being included in the MCS for both the LASSO and EN models, with 50 and 47 rejections of the DM test, respectively. In terms of the other metrics, it is outperformed by the baseline. This contrasts with the features from the Heston model (M12), where significant improvements can be seen across metrics, making M12 the best or second-best performing model. However, it does not outperform the baseline in terms of QLIKE. M13, which incorporates features from the Bates model, performs poorly for LASSO and EN, but shows good performance in RMSE and R^2 for XGBoost.

When comparing M14 and M15, it becomes evident that augmenting with option features significantly enhances forecasting performance. In fact, across all comparisons between the two, M14 outperforms M15 only in terms of mean MAE for EN. M15 is also the model that most consistently outperforms the baseline, with the percentage of instruments showing improvement above 60% in all cases, and in many instances, it has the highest percentage of improvements. In terms of mean QLIKE, M15 outperforms the baseline and is either the best or second-best performing model for each method. Finally, for both LASSO and EN, M15 is included in the MCS and has the highest number of rejections for the DM test.

Performance assessment in high and normal volatility regimes In terms of QLIKE, most models outperform M2 in both normal and high-volatility regimes. For normal-volatility regimes, the MCS includes M8 and M13 for LASSO and EN, and M3, M4, and M6 for XGBoost. For high-volatility regimes, M7, M11, and M15 are included in the MCS for all methods, with M10 and M12 also in the MCS for XGBoost. The highest classification accuracy is achieved by M11 and M12 for LASSO and EN. When comparing M14 and M15, the latter improves on high-volatility regime performance and classification accuracy, while being outperformed in normal-volatility regimes.

M15 is an improvement over the baseline for more instruments than M14, and it has better mean performance for most metrics across methods. Additionally, M15 is part of the MCS in several cases. This improvement seems to result from M15's significantly better performance in high-volatility regimes, which is generally of higher importance to practitioners. This demonstrates that augmenting the HAR-based baseline with features from the option market enhances predictive power.

Feature Importance In Figure 4.6, we plot the feature importance by calculating the median absolute coefficient of each feature when using a LASSO method on the complete data set (M15). We further augment the complete data set (M15) with ten random features, whose entries all independently follow a standard Gaussian distribution. It can be clearly seen that a large number of features have a median coefficient higher than all of the randomly generated features. Notably, all of the different data sets used contain at least one feature scoring a higher feature importance than the ten random features.

The 15 features with the highest feature importance include the three HAR features: the previous daily, weekly, and monthly realized volatilities, as well as the

realized quadraticity. The top 15 features also include VIX, the sum of positive and negative price jumps (Equation (4.73)), as well as positive and negative semivariances. The daily and monthly versions of HLP (with the weekly HLP in the top 30 features) and the Yang–Zhang monthly volatility are included.

When it comes to features derived from option pricing, the average volatility derived from the IVS data is the second most important feature, with the volatility from the Bates model being the sixth most important feature (the volatility from the Heston model is ranked 19th in the feature importance ranking). Lastly, the MSE feature of the Bates model is also among the 15 features with the highest importance.

In Figure 4.7, we plot the correlation of each of the daily volatility estimators with realized volatility. We observe that the Heston volatility as well as the average IVS implied volatility are the estimators that are the least correlated with realized volatility. This is natural, as these are the only estimators that are not calculated using historical data. This observation verifies that the usage of such estimators is beneficial, as they can bring new information into the augmented HAR model.

In Figure 4.7, we plot the correlation of the daily volatility estimators with realized volatility. We observe that the Heston, Bates and IVS implied volatility are the estimators least correlated with realized volatility. This is expected, as these are the only estimators not calculated using historical data. This observation confirms that using such estimators is beneficial, as they introduce new information into the augmented HAR model.

4.6.1 Cross-Asset Model

In this section, we explore an additional volatility model based on the factors used for M15: a *cross-asset* model similar to the one used in [248]. This model provides a different perspectives on volatility, based on the concepts of market volatility and volatility spillover.

Cross-asset models have been widely used for financial problems, such as in [67, 155, 43]. High volatility in one instrument often leads to increased volatility in other instruments, a phenomenon known as volatility spillover. For the cross-asset model, we use multiple features from multiple instruments. For each asset whose volatility we predict, the model is defined as follows

$$\log(\text{RV}_{i,d}) = \mathcal{F}_i \left([X_{1,d-1}^{\text{Base}}, X_{1,d-1}^{\text{Augm}}], \dots, [X_{i,d-1}^{\text{Base}}, X_{i,d-1}^{\text{Augm}}], \dots, [X_{N,d-1}^{\text{Base}}, X_{N,d-1}^{\text{Augm}}] \right).$$

As we are dealing with daily volatility and our sample size is not large enough, we are unable to use the full feature set employed in M15 for each of the instruments.

	RMSE		MAE		R^2		QLIKE		DM Test
	Mean	% of improv.	Mean	% of improv.	Mean	% of improv.	Mean	% of improv.	% of rejections
Model 2 (M2) - Baseline model	0.6029	NA	0.4390	NA	0.2619	NA	0.7262	NA	NA
Model 15 (M15) - All Features	0.5824	64	0.4201	61	0.3124	79	0.7213	69	43
Cross Asset Model	0.5821	67	0.4111	58	0.3109	79	0.7200	72	43

Table 4.6: Comparison of the performances of different models for the forecasting of realized variance. For each model, we compare the performance in terms of the out-of-sample R^2 , Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE), and Quasi Likelihood (QLIKE). For each model, the average of these quantities is shown, as well as the percentage of cases where that model outperforms M2. Lastly, we provide the number of instruments out of 66 that the null hypothesis of the DM was rejected, signifying an improvement over M2.

Instead, we use the features from M15 for the asset’s own features and then augment the model by adding two additional features for each instrument: the previous day’s realized volatility and its forecast based on the M15 model.

Forecasting VIX While cross-asset models do not significantly improve on our ability to forecast the realized variance, we utilize them for a different task. Instead of forecasting realized variance for individual stocks as have done so far, we shortly turn our attention to forecasting VIX, a theoretical quantity based on the volatility of SPX, the index of the S&P market. This renders VIX a natural forecasting task for the cross-asset model. To this end, we compare three different models, M2 the baseline model, M15 the full model and the cross-asset model. The predictors are calculated based on data for the SPX index, as VIX is intrinsically linked to its variance. However, note that as SPX is not directly tradable, the traded volume features are excluded from M15 and the cross-asset model.

In Table 4.7, we show the performance of the three models for VIX. We note that unlike RV forecasting, where the increase in performance between M2 and M15 was marginal, the difference in this case is substantial, with RMSE dropping from 2.415 down to 2.004. The cross-asset model further increases performance, as it outperforms M15 across all metrics.

Economic Application As an economic application, we consider the VIX Mid-Term Futures ETF (VIXM). Instead of using VIX as the target, we instead forecast the next day open-close returns of VIXM, using the forecasts from M15 for stocks as predictors. We formulate a simple theoretical trading strategy, whereby we open a position in VIXM depending on the sign of the forecast on the open-close returns of VIXM of the next trading day, which we close on the closing of next day.

	RMSE	MAE	R^2	DM test (p-value)
Model 2 (M2) - Baseline	2.415	1.603	0.691	NA
Model 15 (M15) - Complete	2.004	1.193	0.787	0.000
Cross-Asset Model	1.923	1.173	0.794	0.000

Table 4.7: Comparison of performance of M2, M15 and the cross-asset model, for forecasting VIX. For each model, we compare the performance in terms of the out-of-sample R^2 , Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE).

In Figure 4.8, we plot the cumulative P&L. It can be clearly seen that following a cross-asset model based on the forecasts of the M15 model, results in a profitable strategy, which highly outperforms the baseline (longing and holding the VIXM). This comes in contrast to a strategy based on forecasts from M14, which scores very poorly in terms of both Sharpe Ratio and Profit per Day (eg Profit & Loss returns). In fact, both M2 and M14 are outperformed by a strategy, where VIXM is shorted. This confirms that not only are the features included in M15 beneficial for volatility forecasting, but they can also bring further economic benefits when employed for forecasting VIX.

4.7 Conclusion

We have augmented the established HAR model with additional estimators of the daily, weekly, and monthly volatility. We have shown significant improvements in performance, with each considered augmentation adding predictive power to the HAR-based baseline model.

We have provided novel methodology for extracting volatility information from the options market by assuming different models. Unlike the majority of the other daily volatility proxies and trackers, the features extracted from the options market are weakly correlated to their contemporaneous realized volatility, indicating how they augment existing models in a meaningful manner, as information that does not originate from historical equity prices. Models M11-M13 (which include the option extracted features) show a significant increase in performance with respect to most metrics. Furthermore, we observe that the extracted volatility estimators from both the IVS and the calibrated current volatility assuming non stochastic volatility models, are both shown to have high feature importance. The first maturity effect, as well as the first two delta effects extracted from IVS data also score highly in terms of feature importance.

We demonstrated how the individual forecasts received from all the individual equity volatilities can be employed as predictors for the forecasting of the different instruments. These forecasts are also effective in boosting the forecasting performance of the market volatility, as they are able to offer predictive power for VIX.

Future work could extend the methodology of extracting volatility from the Heston model by incorporating jumps or considering rough volatility models. In addition, this study has avoided equities that pay dividends, as the calibration of American options is a significantly more challenging problem, and the resulting calibrations were unstable. Some natural extensions could include the further enhancement of our model using news sentiment data, as well as the usage of more complex models such as [174], which was used for high frequency volatility forecasting.

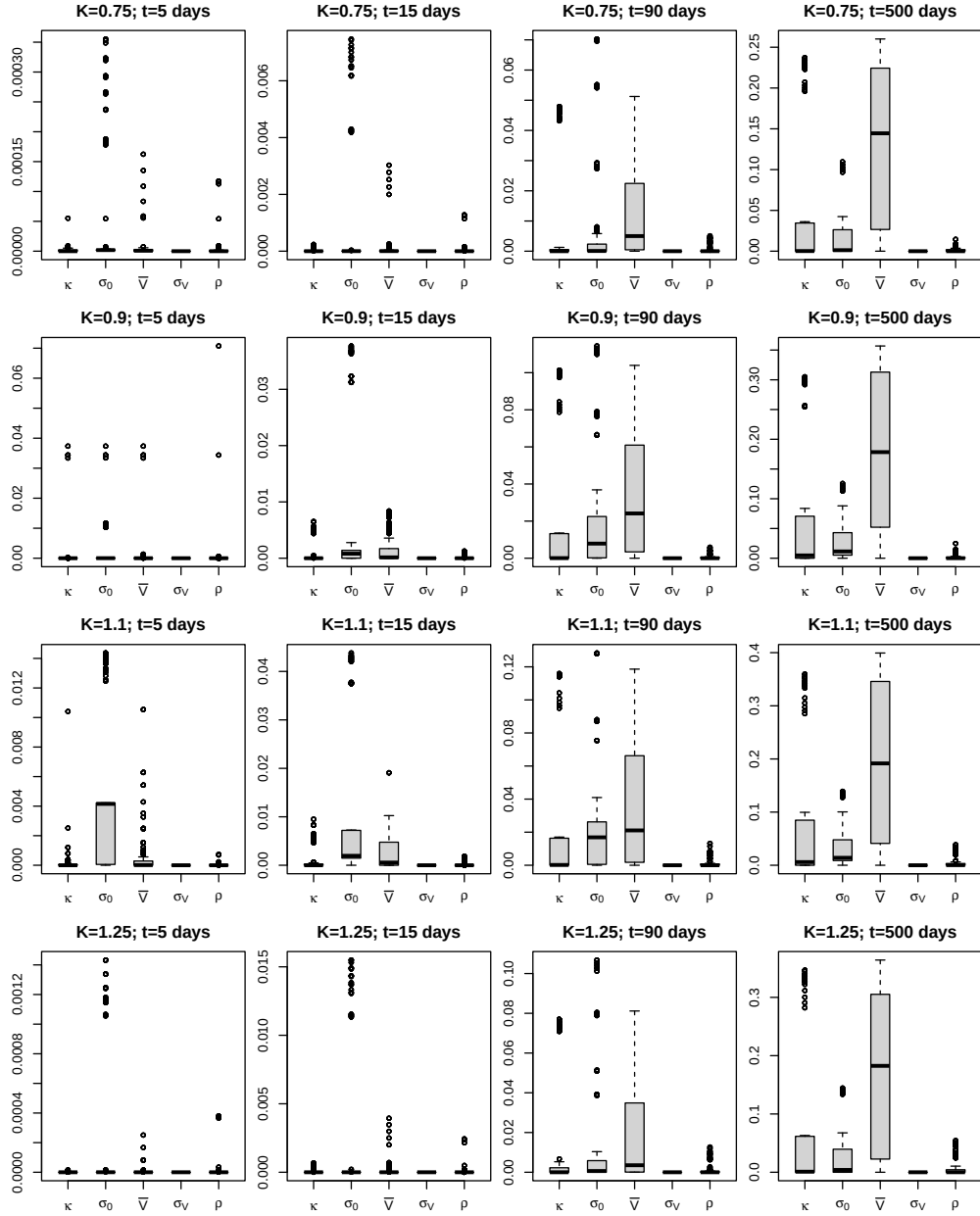


Figure 4.2: Sensitivity of the call option price to each of the Heston model's parameters, for a total of 16 (strike price, time to expiration) combinations. Each plot takes the form of a Box-Whisker plot, demonstrating the distribution of sensitivities, which are calculated as explained in Equation (4.40).

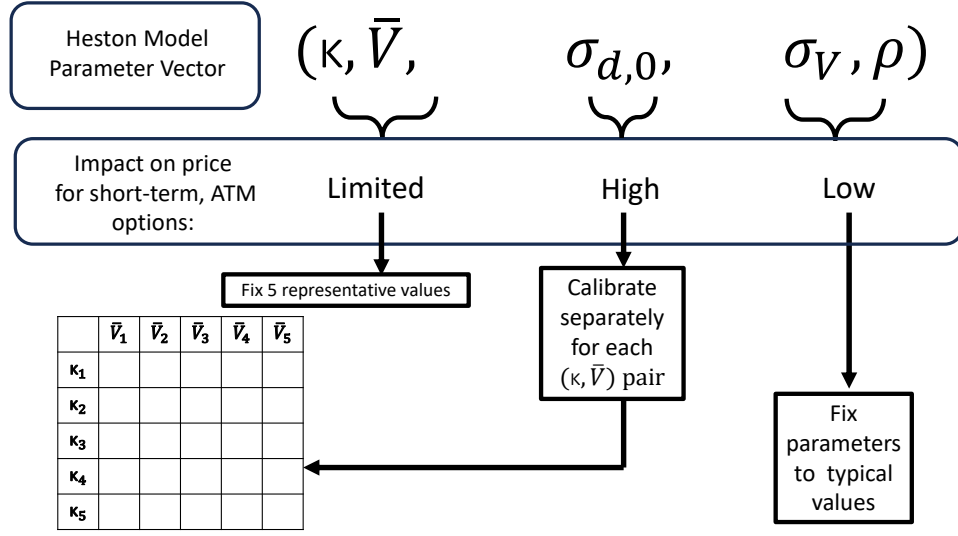


Figure 4.3: Pictorial illustration for the process of the extraction of the volatility estimator $\sigma_{\text{Heston}}^{(d)}$, described in Section 4.4.4. For each instrument, the parameters (σ_V, ρ) are fixed to a singular value, while the parameters $(\kappa, \bar{V})$ take one of five predetermined values, creating 2-dimensional grid. For each combination $(\kappa, \bar{V})$ and the fixed values of (σ_V, ρ) , the model is then calibrated for the optimal volatility $\sigma_{d,0}$, based on minimizing Equation (4.49).

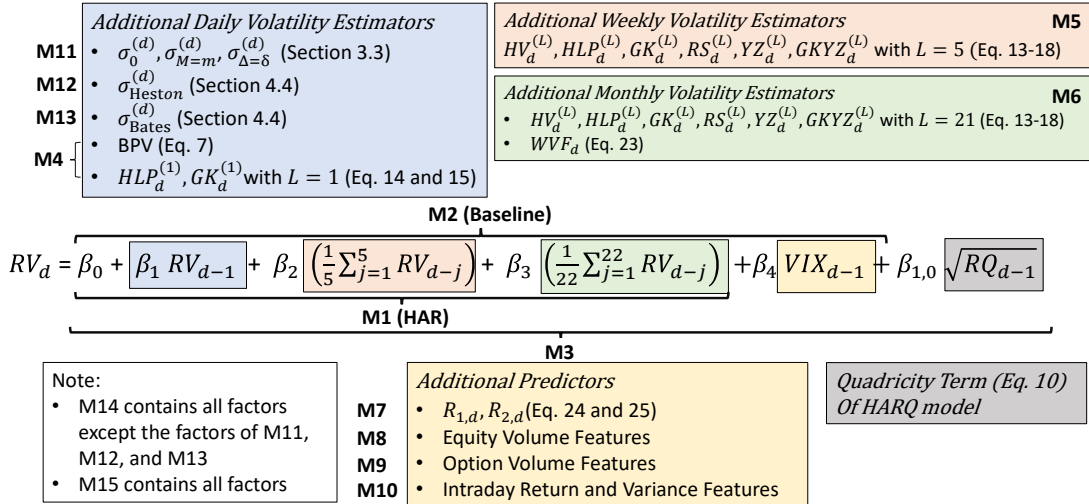


Figure 4.4: Pictorial illustration for the investigated features. M1 (resp., M2) corresponds to the HAR (resp., the baseline) model. The additional predictors for daily, weekly and monthly predictors are shown correspondingly, as well as a list of additional predictors for volatility.

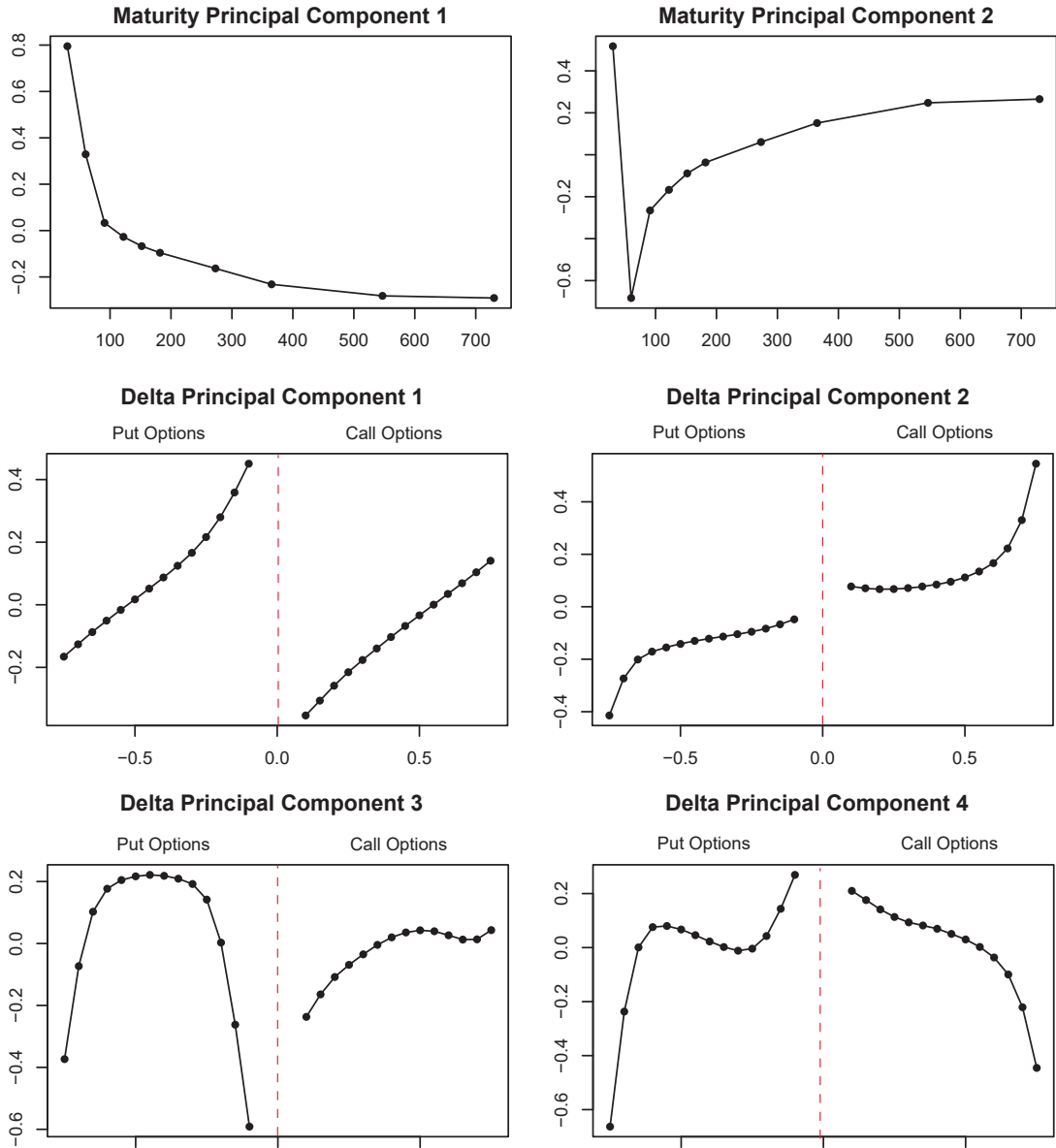


Figure 4.5: Plots of the first two principal components for the maturity effects, and the first four principal components for the delta effects, corresponding to model IVS-5.

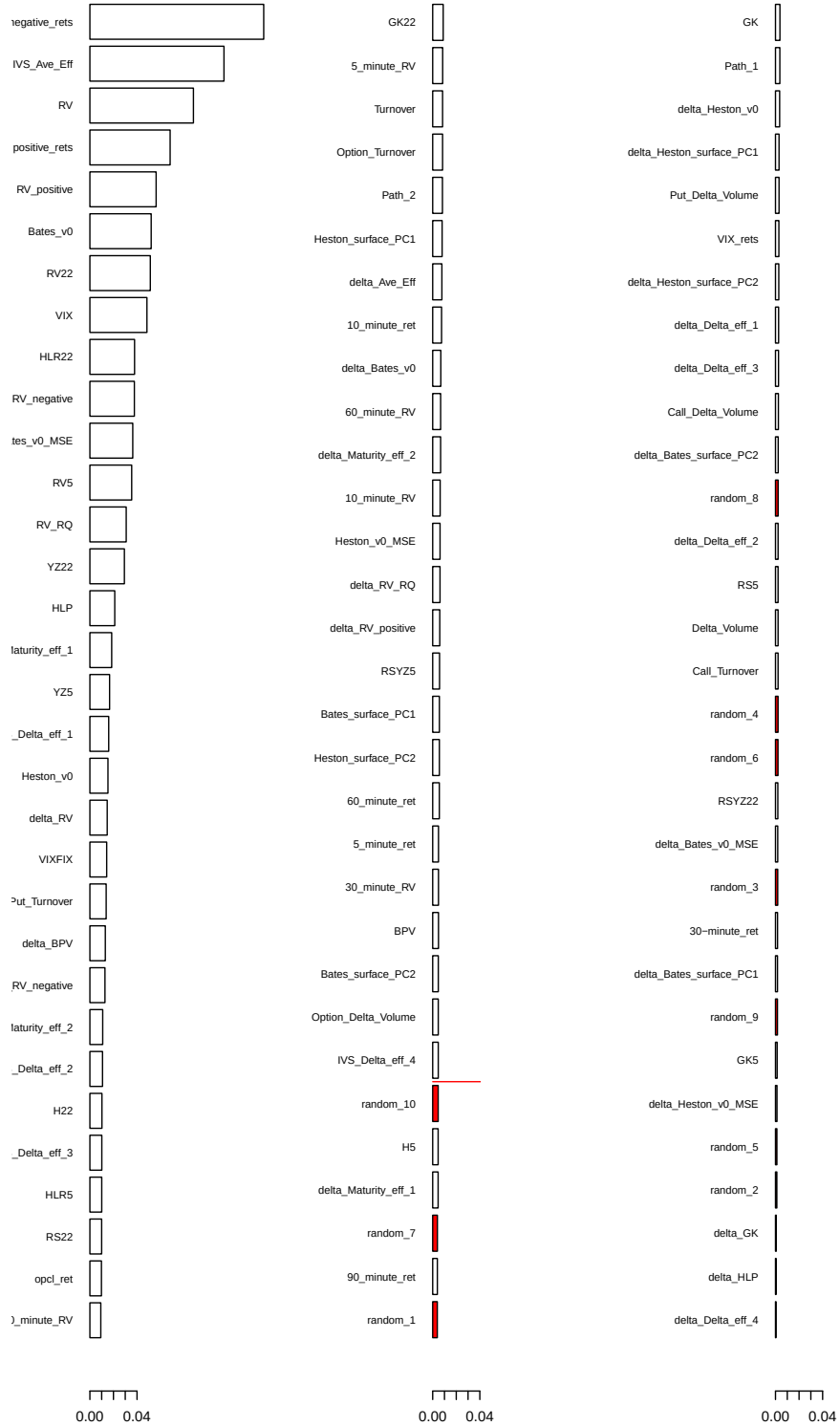


Figure 4.6: Barplot of the median absolute value of each feature's coefficient using an EN model, fitted with the full data set (M15). In addition, we use 10 features whose entries are all independent identically distributed $N(0,1)$, which are shown in red color.

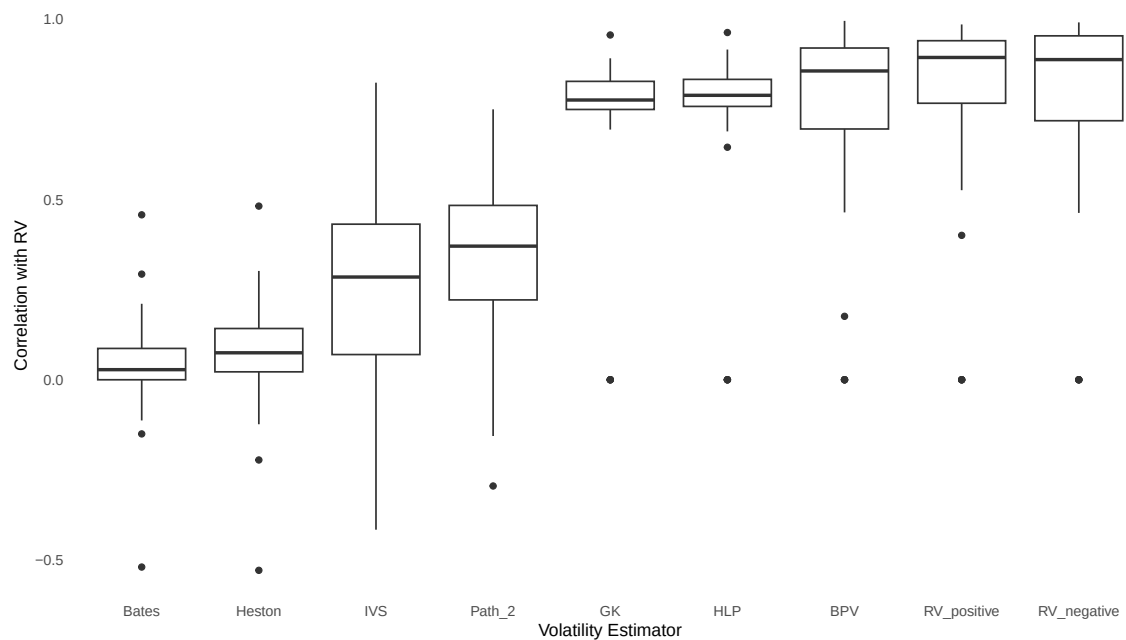


Figure 4.7: Box and Whisker plot of the median correlation of the different daily volatility estimators, with the corresponding value of the contemporaneous daily realized volatility.

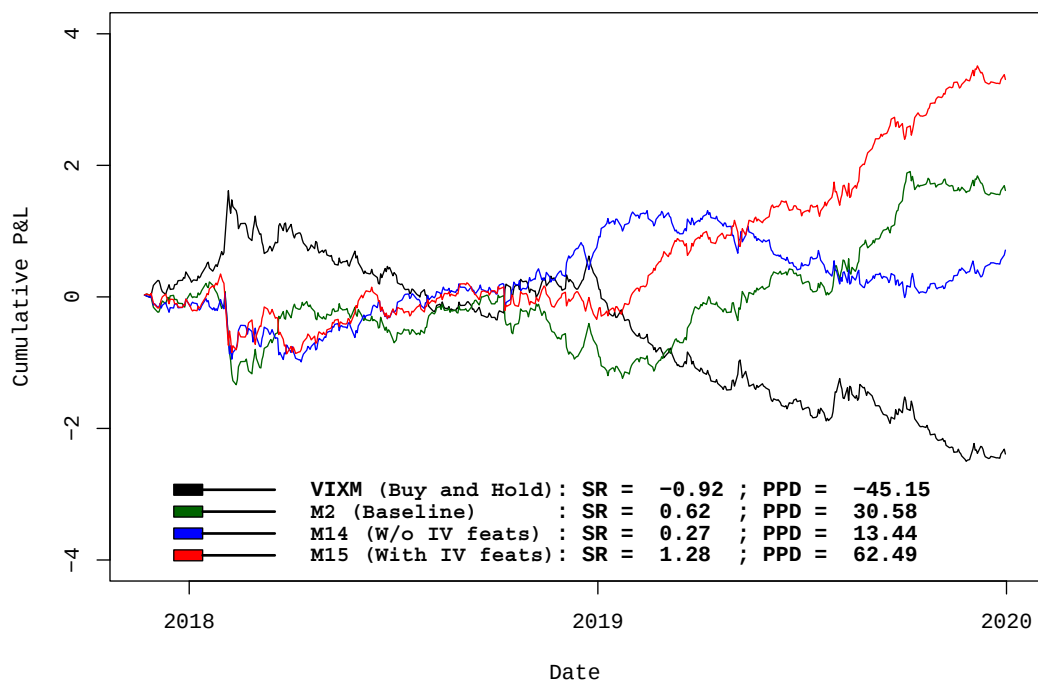


Figure 4.8: Cumulative P&L plot for VIXM. SR denotes the Sharpe Ratio and PPD the P&L return per Day, which are defined in Appendix C.2.

Chapter 5

A GCN-LSTM Approach for ES-mini and VX Futures Forecasting

5.1 Introduction and Related Literature

5.1.1 Futures

Futures are highly traded financial instruments used for both risk management and speculation. Compared to directly trading the underlying instruments, many futures allow for 24-hour trading, providing traders with more flexibility. More importantly, as derivative products, they can offer higher leverage, amplifying potential returns and minimizing the cost of hedging. Additionally, they make feasible short-selling a contract, even when the underlying asset is not owned.

In this paper, we focus on the interaction of futures contracts with different expirations. The volatility term structure of VIX futures has been studied by [156], who provide an approximation for the price of VIX futures. A later alternative is [117], who offer an exogenous model based on specific characteristics. Our goal is to follow a data-driven approach for forecasting the price movement, volatility, and traded volume of these futures by employing data from various products across different expiration periods.

5.1.2 SPX and VIX

We focus on two types of futures: E-mini S&P 500 futures (ES), with a quarterly cycle, and CBOE Volatility Index futures (VX), with a monthly cycle. Both futures are widely used in describing and tracking the market.

The SPX is a stock market index that measures the performance of 500 of the largest publicly traded companies in the United States, across different industries, making it one of the most widely used gauges of the overall market trend. The VIX represents the market's expectations for volatility over the coming 30 days and reflects investor sentiment on market risk. It is derived from the prices (and, thence, the implied volatilities) of SPX options, as explained in [58]. High values of VIX indicate increased market uncertainty, while low values suggest a more stable market environment.

The ES futures, traded on the Chicago Mercantile Exchange (CME), are a more accessible alternative to standard S&P 500 futures, being only one-fifth of their notional size.¹ An important question concerns what is the market in which the S&P 500 price discovery is taking place. While the SPX ETF is generally accepted to lead its derivative products, some studies (e.g., [137]) on price discovery have concluded that E-mini futures have a higher informational share than their ETF counterparts during periods with high volatility.

There has also been significant academic interest in the dynamics of the VIX market. [251] suggested a model for the pricing of VIX futures. [218] formulated a simple trading strategy based on whether the basis is in contango (VIX futures are shorted) or in backwardation (VIX futures are longed).²

This relationship has been shown to be more relevant after major news announcements [56]. Other profitable trading strategies have been suggested, including [221], who used six readily available predictors. [184] have made use of option pricing data and recurrent neural networks. A causal lead-lag relationship was also established with VIX-tracking Exchange-Traded Products [182].

5.1.3 Relations between Returns, Volatility and Trading Volume

Price movements of products tied to the same underlying index are typically highly correlated. While the ES and VX futures track different indices, they can still be informative for each other, as the two indices are naturally intertwined. Specifically, it has been established that returns for VIX and SPX are negatively correlated (see e.g. [233]), which is a natural consequence of returns being negatively correlated with

¹One E-mini futures contract corresponds to 50 times the value of the SPX index, as opposed to 250 times for the standard S&P 500 futures.

²Contango (Backwardation) occurs when futures prices are higher (lower) than the risk-neutral price for futures.

volatility. For example, [175] showed this for the German and French stock markets. [18] analyzed the lead-lag relationship of futures corresponding to the VIX and SPX indices, showing that under a high-volatility regime, VIX futures lead SPX futures, and the correlation is highly negative.

The relationship between volatility and trading volume has also been widely studied. There is a popular saying that “*it takes volume to move prices*”, which directly explains the positive correlation between the two quantities, with [4] showing how volume is an informative predictor in a volatility model. Theoretical frameworks for this interaction include the mixture of distribution hypothesis by [69] and the sequential information arrival hypothesis by [63], both modeling how information embedded in trading volume is informative for future changes in price.

There have also been studies on the relationship between returns and volume. For example, [153] showed that the relationship between returns and volume depends on the prevalence of informed trading in the market. In Chapter 2, we showed how volumes from the option market can be used to forecast future returns for the underlying equities.

5.1.4 Graph Neural Networks

In this paper, we use the methodology of Graph Neural Networks (GNN) to analyze the relationships between different products. GNNs were first introduced by [205] as a type of neural network capable of dealing with graphical data. They gained rapid popularity after the introduction of Graph Convolutional Networks (GCN) by [134]. These networks and their extensions have been used for many applications, including link prediction [252], node classification [200], and outlier detection [237]. Much research has also been conducted on time-series forecasting, including [49], who introduced the spectral temporal GNN. Traffic forecasting has been a major area of application, with [145] and [78] being major examples of urban and maritime traffic forecasting, respectively.

There have been numerous applications of GNNs in finance. Examples include [164], who used a knowledge graph to predict stock prices in the Nikkei 225 market, and [144], who used a combination of LSTM and GCN with news data to classify stocks into upwards and downward trending. More recently, [246] combined GNN with Heterogeneous Auto-regressive Regression models to forecast the volatility and covariance matrix of different instruments. A detailed overview of other GNN applications in finance has been covered in [229].

It is important to stress that GNNs encompass quite a diverse class of methods, with many tailored to the specific problem at hand. A multi-layered version of GNN is proposed by [102], who extended GNNs to a multi-layered version by analyzing the intra-layer and inter-layer interactions separately. Furthermore, the time-series nature of a dataset can also be considered using one of the various graph neural network methods tailored for time series forecasting [123]. Focusing on research based on stock market forecasting, [253] also provided a shorter summary of relevant research.

GCNs have also been used by [143] to enhance the predictability of chart similarity approaches. The improved graph convolutional network (IGCN) is suggested by [54], which captures both stock market and individual-level features. [244] used an architecture with a GCN and gated recurrent unit (GRU) to show promising performance for stock movement prediction.

5.1.5 Main contributions of this paper

The focus of this paper is the forecasting of three major financial quantities monitored by practitioners—financial returns, volatility, and volume—for two of the major indices (SPX and VIX) that represent the state of the market. The change in the overall market state is captured by the return of SPX, the current and expected volatilities are captured by the volatility of SPX and the value of VIX, while the “second order volatility” is captured by VIX’s volatility. The volumes of instruments corresponding to these indices are deemed to act as a proxy for the market trading activity. A disproportionately high trading activity in VIX could indicate high hedging activity or even a pessimistic view of the market.

While much research has been conducted on the interactions between these quantities, a less explored angle is a data-driven study of the interplay in the futures market with various products of differing expirations. The multitude of products can naturally be represented as a network, where each node of the network represents a different product. This network can act as a visualization of the correlation structure between the different quantities discussed.

More importantly, these networks may be used as input to a graph neural network (GNN). As all of these quantities are highly correlated, a joint forecasting and subsequent propagation via a GNN could prove promising. Such a layer can be used on top of a Long Short-Term Memory (LSTM) network [112], which is a staple in time series forecasting. In later sections, we compare and contrast the base LSTM network as well as other baselines with the GCN-enhanced LSTM network (GCN-LSTM). The

version of GCN-LSTM we formulate is suitable for highly correlated graphs. By utilizing features generated by products of differing term structures, we augment and improve the forecasting performance for all three main financial quantities discussed in this paper.

As we are considering different quantities, multiple types of interactions are considered, which can result in different graphs. In particular, a different network is constructed for each pair of variables. These graphs are subsequently used in a multi-channel GCN network to provide parallel propagation attempts at smoothing and improving the base forecasts.

Lastly, as there has been a shift towards loss functions with economic significance, we utilize the Sharpe Ratio in the loss function for financial returns, as well as a quasi-likelihood loss function for volatility.

Paper Structure The rest of this paper is structured as follows; In Section 5.2, we give a high-level overview of the elements of the model that we use. In Section 5.3, we describe the data set used as well as the graphs that are used by the GCN-LSTM model. In Section 5.4, we consider some practical issues, including the evaluation metrics, predictive features, baselines and implementation details. In Section 5.5, we show the results for the forecasting tasks of returns, volatility and volumes. Finally, in Section 5.6 we conclude and discuss future directions.

5.2 Model

The model we use for forecasting returns, volatility, and trading volume is an iteration of the GCN-LSTM model. This approach involves combining an LSTM network layer, which is widely used for data sets with a temporal structure, such as the financial data in our study. For each individual product, a distinct module operates in parallel, as illustrated in Figure 5.1.

Although this increases the computational resources required and reduces the number of available trainable data points, it enables the model to assign differing weights to each product, thereby enhancing its expressive power. As shown in Figure 5.2, the architecture incorporates a GCN layer that integrates multiple graphs, allowing the individual forecasts for different products to be used in conjunction. Finally, the GCN layer is followed by a dense layer, which includes skip connections from the last dense layer before the GCN layer. In the remainder of this section, we provide an overview of the different layers employed.

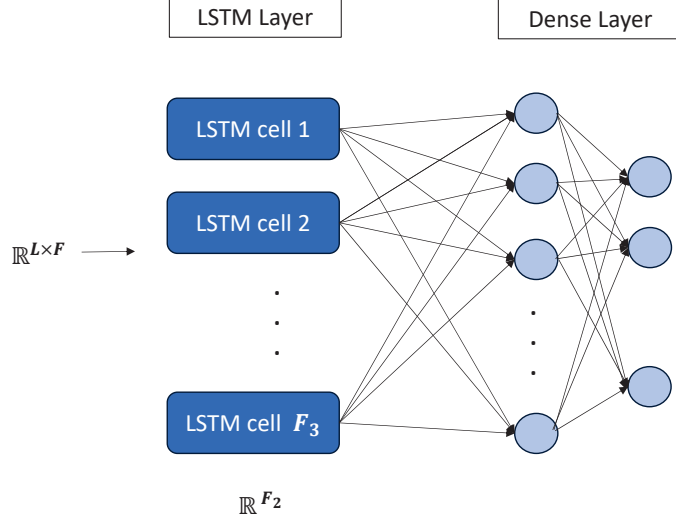


Figure 5.1: The module used for each distinct product, comprising an LSTM and a series of dense layers.

5.2.1 Graph Convolutional Networks

Graph Neural Networks (GNNs) represent a class of deep learning methods designed for network data, based on the concept of neural message passing. Following the notation of [105], we consider a set of nodes $V = \{v_1, \dots, v_N\}$. We denote by $\mathcal{N}(v)$ the set of nodes that are connected with node v .

At each layer k , the message of node v , represented as a vector of length F_k , is denoted by h_v^k and is calculated as

$$h_v^k = \text{UPDATE} \left(h_v^{(k-1)}, m_{\mathcal{N}(v)} \right), \quad (5.1)$$

where

$$m_{\mathcal{N}(v)} = \text{AGGREGATE} \left(\{h_u^{(k-1)}, \forall u \in \mathcal{N}(v)\} \right). \quad (5.2)$$

We choose the AGGREGATE operation to be the mean function, i.e the mean message coming from a node's neighbors. For single-layered networks, this structure allows each node to be promptly updated using information from its neighbors. Like other neural network architectures, multiple such layers can be stacked on top of each other for higher-order propagation. The result is that deeper networks can incorporate information from nodes that are at a higher geodesic distance from the current node.

One of the most widely used iterations of GNN is the Graph Convolutional Network (GCN), initially proposed by [134]. In its simplest form, a single layer of GCN

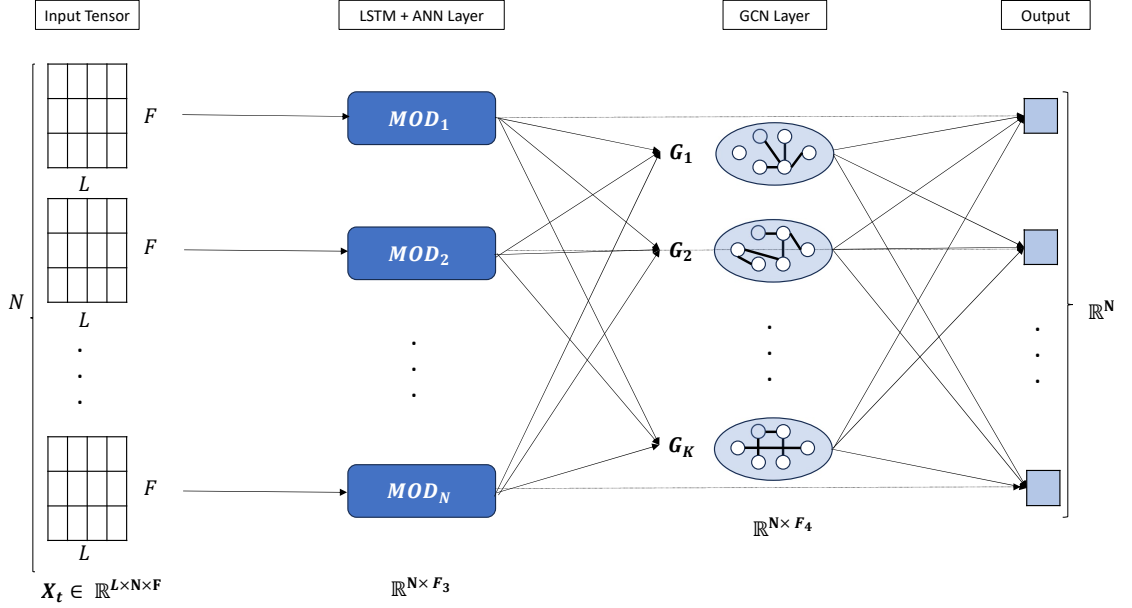


Figure 5.2: Overview of the GCN-LSTM model used. Each node/instrument corresponds to a separate trainable sub-network, consisting of LSTM and dense layers as shown in Figure 5.1. The outputs are then pooled in the GCN layer, which consists of K pre-calculated graphs. A final dense layer is then added to produce the overall output; this also includes skip connections from the individual sub-networks.

can be expressed as

$$h_v = \sigma \left(W \sum_{z \in \mathcal{N}(v) \cup \{v\}} \frac{h_z^0}{\sqrt{|\mathcal{N}(v)| |\mathcal{N}(u)|}} \right), \quad (5.3)$$

where h_z^0 is the input vector for node z , W is the $F_1 \times F$ weight matrix, and σ is an activation function (we use the tanh function). Given that we are dealing with a network of just 14 nodes, a single-layer architecture using an UPDATE function like Equation (5.3) is sufficient for effectively propagating forecasts across different products, particularly one that allows for weighted and signed graphs.

In our case, we are dealing with multiple distinct graphs, which necessitates training different weight matrices along with the rest of the network. Denote by $A^{(i)}$ the adjacency matrix of the i^{th} graph, by $D^{(i)}$ the $N \times N$ diagonal matrix whose i^{th} entry is equal to $|\mathcal{N}(v_i)|$, and by I_N the $N \times N$ identity matrix. Also, denote by $H^{(0)}$ the $N \times F$ matrix whose i^{th} column is equal to h_z^0 . The output of this layer is thus an order-3 tensor $(H^{(1)}, \dots, H^{(K)})$ computed as

$$H^{(i)} = \text{ReLU} \left((D^{(i)})^{-1/2} (A^{(i)} - I_N) (D^{(i)})^{-1/2} H^{(0)} W^{(i)} \right), \quad i = 1, \dots, K, \quad (5.4)$$

where $W^{(i)}$ are the $F_1 \times F$ matrices to be optimized.

5.2.2 Long-Short Term Memory Networks (LSTM)

Recurrent Neural Networks (RNNs) form a class of neural networks specifically designed for data sets with temporal structures. They achieve this by retaining information from the processing of previous time steps, which is used in conjunction with the input of the current time steps. The architecture and parameters are shared across time steps.

A common issue with simple RNNs is that gradients have a tendency to either vanish or explode during training. Long Short-Term Memory (LSTM) networks [112] are an RNN variant specifically designed to mitigate these issues by introducing a cell state. This cell state contains a summary of the previous cells and is regulated by several gates that improve gradient flow.

Denote by x_t , c_t , and h_t the input, cell state, and output of the module corresponding to time point t , respectively. The formula of a basic LSTM unit takes the form:

$$\begin{aligned}
\text{Forget Gate:} \quad & f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f), \\
\text{Output Gate:} \quad & o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o), \\
\text{Input Gate:} \quad & i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i), \\
\text{Candidate Cell State:} \quad & \hat{c}_t = \tanh(W_c x_t + U_c h_{t-1} + b_c), \\
\text{Updated Cell State:} \quad & c_t = f_t \odot c_{t-1} + i_t \odot \hat{c}_t, \\
\text{Hidden State:} \quad & h_t = c_t \odot \tanh(c_t).
\end{aligned} \tag{5.5}$$

where $W_f, U_f, b_f, W_i, U_i, b_i, W_o, U_o, b_o, W_c, U_c$ and b_c are all trainable parameters, and $\odot$ is the Hadamard product.

5.3 Data and Graph Construction

5.3.1 Data

We use Level-1 data provided at tick-level for E-mini *S&P* 500 and VX futures. At each point, we consider the eight VX futures and four ES futures closest to expiration, respectively. We also include the SPX and VIX indices in our analysis. We denote by VX_1 (ES_1) the current VX (ES) futures contract that is closest to expiration, by VX_2 (ES_2) the VX (ES) futures contract that is second-closest to expiration, and so on. The corresponding indices are denoted as VIX and SPX.

The time period considered spans 10 years, specifically from January 2014 to December 2023. We remove any entries that correspond to zero liquidity on either the bid or ask side. Our focus is on the hourly level, forecasting hourly returns, realized

volatility, and a proxy for trading volume. Specifically, as we do not have access to trading volume data, we approximate the number of trades by counting the number of updates to the top of the book data caused by trades (excluding cancellations and new orders).³

A plot of Spearman correlations of hourly returns is shown in Figure 5.3. ES futures clearly exhibit high positive correlations with each other, as well as with the SPX index, a pattern also observed for VX futures products. Naturally, products with similar expirations tend to have higher correlations. We also observe a negative correlation between ES and VX futures, with near-month ES (VX) contracts generally having higher correlations with VX (ES) futures compared to contracts with longer expiration periods.

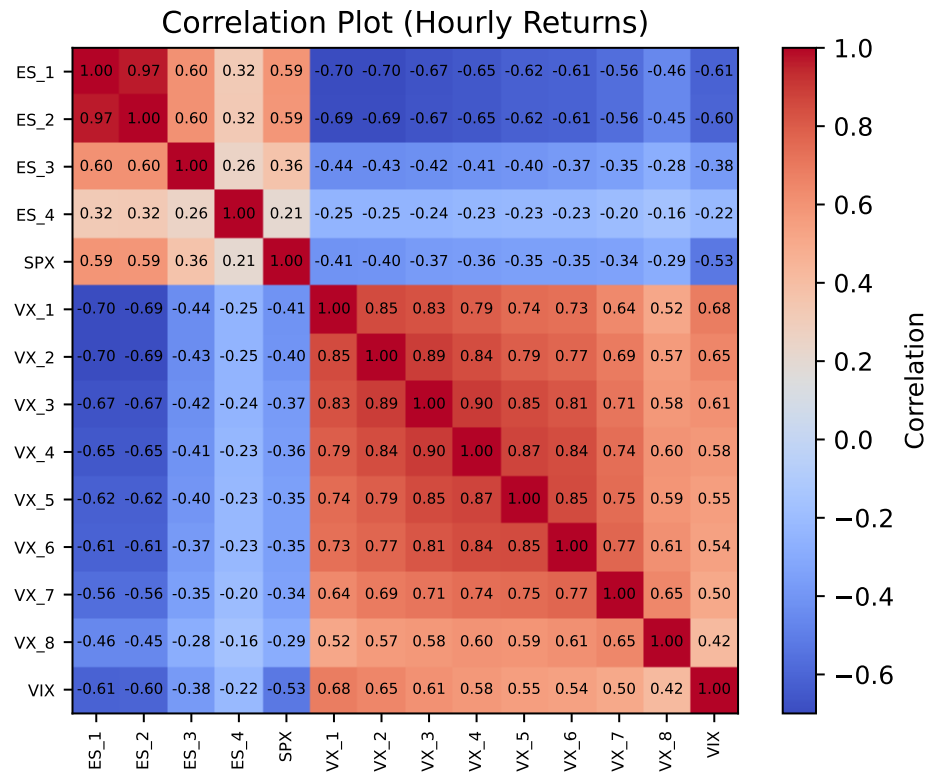


Figure 5.3: (Spearman) Correlation plot based on the hourly returns of ES and VX futures.

³While different from volume, the proxy we use is closely related to volume. This is because a very large proportion of such trades appear to correspond to trading volumes of one contract.

5.3.2 Graph Construction

Similar to [215], we use multiple graphs as input for the GCN layer. The constructed graphs are signed (i.e in the sense that the edges may take either positive or negative weights) and based on different ordered pairs of the three main quantities: returns, volatility, and trading volume. For each ordered pair of variables, we build four different adjacency matrices using Spearman Correlation; Contemporaneous Weighted, Contemporaneous Unweighted, Lagged Weighted and Lagged Unweighted.

We calculate $C_{x,y}$ as the Spearman correlation between vectors x and y . When constructing one of the **Contemporaneous** graphs, we use vectors x and y that are aligned (i.e elements at the same index correspond to the same time period). The lagged correlation, with a lag of one, is calculated when constructing one of the **Lagged** graphs, i.e we use $(x_1, \dots, x_{T-1})$ and $(y_2, \dots, y_T)$ instead of x and y .

Since we are dealing with a small number of nodes whose attributes are all highly correlated, we build the adjacency matrices to avoid working with a complete graph. A natural solution is to use a weighted adjacency matrix based on the correlation matrix. Except for the diagonal entries, which are set to 1, all positive and negative entries are normalized separately column-wise, as shown below.

$$A_{v_i, v_j}^{(\text{Weighted})} = 1_{(i=j)} + 1_{(i \neq j, C_{v_i, v_j} > 0)} \frac{C_{v_i, v_j}}{\sum_{j \neq i} C_{v_i, v_j} 1_{(C_{v_i, v_j} > 0)}} + 1_{(i \neq j, C_{v_i, v_j} < 0)} \frac{C_{v_i, v_j}}{\sum_{j \neq i} (-C_{v_i, v_j}) 1_{(C_{v_i, v_j} < 0)}}. \quad (5.6)$$

We also consider an alternative method to produce an unweighted graph. Given the high correlation among most instruments, the resulting graph would nearly be a complete graph. To address this, we limit the number of edges by including only a fixed number of connections, specifically the K -nearest neighbors to each node.

$$A_{v_i, v_j}^{(\text{Unweighted})} = 1_{(i=j)} + 1_{(i \neq j, i \in \text{NN}_K(j))} \text{sign}(C_{v_i, v_j}), \quad (5.7)$$

where

$$\text{NN}_K(k) = \left\{ j : \left| \{ i \neq k : |C_{i,k}| \geq |C_{j,k}| \} \right| \leq K \right\}. \quad (5.8)$$

5.3.3 Discussion

In Figure 5.4, we show a selection of the Contemporaneous Unweighted graphs (as discussed in Section 5.5, Contemporaneous graphs tend to outperform Lagged graphs). For brevity, we only include one graph for each pair, as the graphs corresponding to the two ordered pairs of a pair are based on the same correlation matrix (but

transposed). It is also important to note that since indices do not have corresponding trading volumes, there are no outgoing (incoming) edges for graphs where the first (second) pair variable is volume. In Figure 5.5, we present a bar plot of the out-degrees for these graphs, while in Table 5.1, we display some important summary statistics regarding these graphs.

The structure of most of these graphs is generally determined by the two product clusters (SPX and VIX). A few graphs, such as the Return-Return and Volatility-Volatility graphs, are highly clustered with only a few connections between the two clusters, most of which go from the VIX cluster to the SPX cluster and not vice versa. This is partly due to the fact that multiple VIX futures products have high liquidity, in contrast to SPX futures, where most liquidity is concentrated in ES_1.

For the graphs where return is one of the pair variables, there are multiple negative edges across nodes, while no such edges exist for the other graphs. This is largely due to the negative correlation between the market's return and volatility, which is evident from the large number of negative edges between the VIX and SPX clusters in the Return-Return graph. This phenomenon is also observed in the Return-Volatility graphs, where all edges between SPX nodes are negative. Interestingly, SPX nodes are mostly connected to other SPX nodes, despite VIX returns and SPX volatility being highly correlated. In the Volatility-Return graph, there are no edges originating from SPX nodes, with edges between VIX nodes being positive and those from VIX to SPX being negative. Lastly, in the Volume-Return graph, the edges from VIX to SPX are also negative, while the edges from SPX to VIX are positive. Interestingly, ES_3 has positive edges toward all VIX-related nodes.

It is noteworthy that in certain types of edges within the SPX cluster, and some cross-cluster edges, higher connectivity is shown by more liquid products (ES_1, ES_2, VX_1, and VX_2), while in other cases, products with very limited liquidity are more connected (ES_3, ES_4, VX_7, and VX_8). The Return-Return, Volatility-Volatility, and Volume-Volume graphs fall into the former category, as they are all based on high-correlation matrices. Another such network is Volatility-Volume, possibly due to the fact that the prices of these high-liquidity products are less prone to outliers, making large fluctuations possible only when the overall market trading volume is high.

More interesting are the cases where low-liquidity products play a bigger role. As can be seen in Figure 5.5, for Returns-Volume, Volume>Returns, and Volume-Volatility, the low-liquidity nodes such as ES_3, VX_7, and VX_8 have very high out-degrees. In particular, for Returns-Volume, VX_8 seems to have positive edges

	Number of Edges		Highest Out Degree		Number of Edges between clusters				Quantiles (0.25,50,75,100)
	Positive	Negative	Positive	Negative	SPX to SPX	SPX to VIX	VIX to SPX	VIX to VIX	
Return - Return	35	7	7 (VX_4)	3 (VX_1)	9	1	6	26	(-0.69,-0.40,0.22,0.66,0.96)
Return - Volatility	27	15	8 (VX_2)	5 (ES_2)	10	4	5	23	(-0.10,-0.03,0.01,0.04,0.13)
Return - Volume	31	5	8 (VX_5)	4 (ES_4)	3	6	9	18	(-0.04,0.00,0.00,0.01,0.03)
Volatility - Return	27	15	6 (VX_3)	4 (VX_4)	0	0	15	27	(-0.10,-0.03,0.01,0.04,0.13)
Volatility - Volatility	42	0	7 (VX_4)	NA	9	0	6	27	(0.01,0.11,0.30,0.45,0.88)
Volatility - Volume	36	0	11 (ES_1)	NA	7	16	5	8	(0.00,0.05,0.09,0.17,0.31)
Volume - Return	30	12	9 (ES_3)	4 (VX_8)	5	12	10	15	(-0.04,0.00,0.00,0.01,0.03)
Volume - Volatility	42	0	10 (VX_7)	NA	4	7	11	20	(0.00,0.05,0.09,0.17,0.31)
Volume - Volume	36	0	6 (VX_3)	NA	4	2	8	22	(0.00,0.00,0.61,0.85,0.98)

Table 5.1: Summary of the different networks for the nine different graphs for the unweighted contemporaneous graphs.

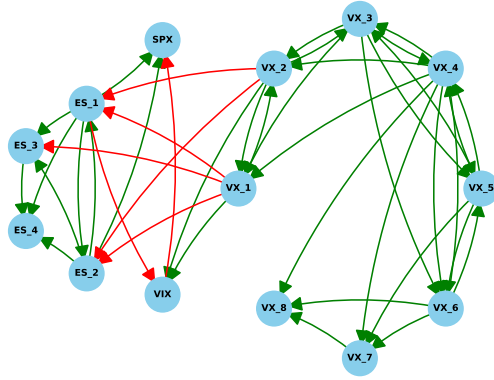
towards most SPX products, while ES_4 appears to have negative edges towards many VIX nodes.⁴ For Volume-Volatility, most connections from VIX to SPX occur with VX_7 and VX_8.

Note that a different pattern holds for edges connecting two VIX products, where products with expirations that are close tend to be connected. This results in a more uniform distribution of out-degrees for this part of the network, with VX_3 and VX_4 being the nodes with the highest (positive) out-degrees towards the VIX cluster in graphs like Return-Return, Volatility-Volatility, Volume-Volume, and Volatility-Volume.

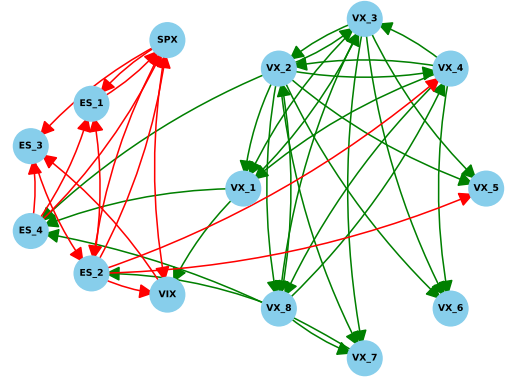
No particular patterns are observed for the Volume-Volume graph, which partly stems from the fact that this graph is derived from a matrix where most correlations are quite high. A few other interesting observations include ES_3 returns having a high negative correlation with some VIX products' volatility; and that for Volume-Volatility, most edges seem to originate from VIX nodes, with high maturities having a large out-degree (not shown in Figure 5.4).

Overall, while SPX and VIX tend to form the two primary clusters in most of these graphs, there is still a significant number of edges between SPX and VIX nodes. Additionally, while higher-liquidity products such as ES_1 and VX_1 have high connectivity, medium and low-liquidity products are also highly connected in certain graphs, like the Return-Volume and Volume-Volatility graphs. While index nodes still have some connections, they are generally among the least connected (as can also be seen in Figure 5.5). Altogether, these observations suggest that a GCN layer would be beneficial, as various interactions exist between nodes that cannot be attributed to a single rule, but rather hint at a complex set of interactions underpinning the dynamics of the market.

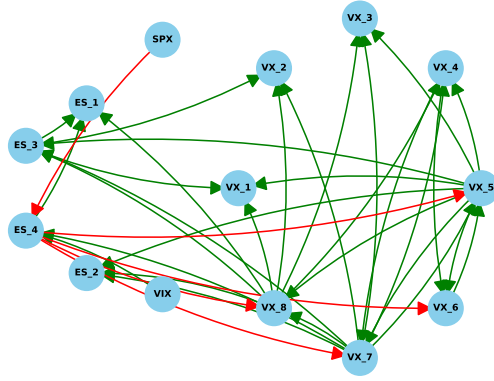
⁴It should be noted that this pair of variables has a graph based on very low correlations (see Table 5.1), which could make the resulting graph less reliable.



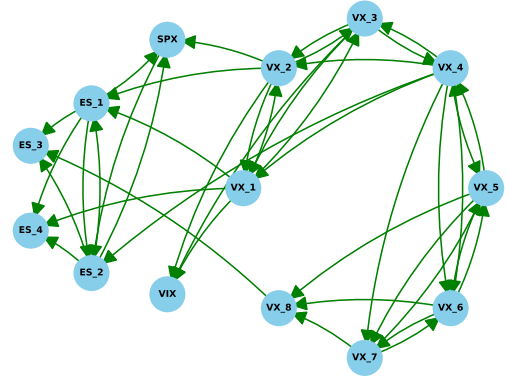
(a) Returns-Returns



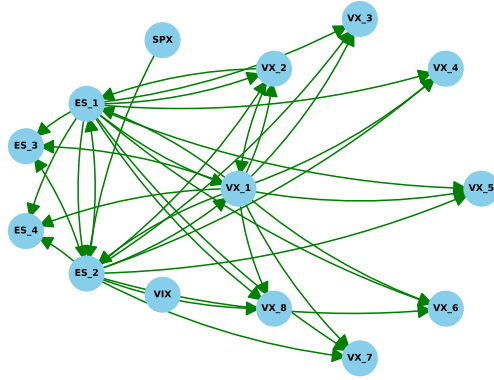
(b) Returns-Volatility



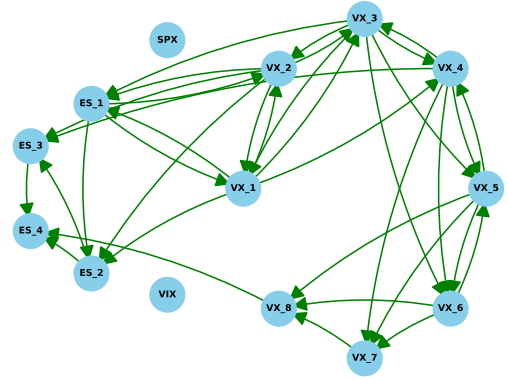
(c) Returns-Volume



(d) Volatility-Volatility



(e) Volatility-Volume



(f) Volume-Volume

Figure 5.4: Visualization of networks. The networks shown are signed, unweighted, and directed. Green arrows correspond to positive edges, while red corresponds to negative edges.

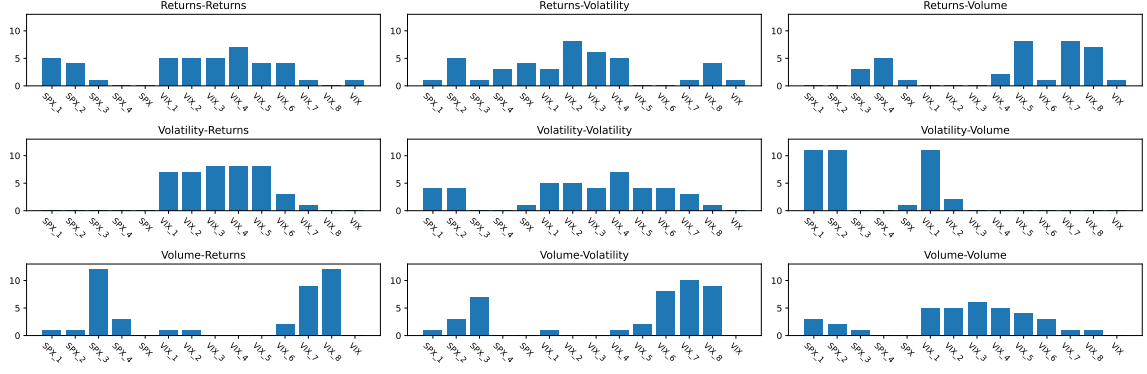


Figure 5.5: Barplot for the out degrees of the contemporaneous unweighted graphs.

5.4 Forecasting Setup

5.4.1 Financial Terms

Denote by $P_t^{(B)}$ (resp. $P_t^{(A)}$) the bid (resp. ask) price. The mid-price $P_t^{(M)}$ is calculated as

$$P_t^{(M)} = \frac{P_t^{(A)} + P_t^{(B)}}{2}. \quad (5.9)$$

The k -minute return observed at time t is then defined as

$$R_t^{k\text{-minute}} = \frac{P_t^{(M)} - P_{t-k}^{(M)}}{P_{t-k}^{(M)}} \approx \log \frac{P_t^{(M)}}{P_{t-k}^{(M)}}. \quad (5.10)$$

Suppose that $\hat{R}_t^{k\text{-minute}}$ is a forecast of $R_t^{k\text{-minute}}$, acting as a trading signal. The forecasts are used as signals for a trading strategy, as a way of evaluating its performance. Specifically, if the sign of $\hat{R}_t^{k\text{-minute}}$ is positive, a theoretical buy transaction is conducted, while a theoretical sell transaction is conducted when the sign is negative.

Consider D_d as the set containing the time indices corresponding to day d . The daily Profit and Loss (P&L) for day d is given by

$$P\&L_d = \sum_{t \in D_d} R_t^{k\text{-minute}} \text{sign}(\hat{R}_t^{k\text{-minute}}). \quad (5.11)$$

The annualized Sharpe Ratio is calculated as

$$\text{SR} = \frac{\text{mean}(\{R_d\}_{d=1}^D) \times \sqrt{252}}{\text{sd}(\{R_d\}_{d=1}^D)}, \quad (5.12)$$

The Profit Per Dollar traded is calculated as

$$\text{PPD} = \text{mean}(\{P\&L_d\}_{d=1}^D). \quad (5.13)$$

We also define the k-minute realized variance calculated as

$$RV_t^{k\text{-minute}} = \sum_{j=0}^{K-1} (R_{t-j}^{1\text{-minute}})^2. \quad (5.14)$$

As realized volatility is highly skewed, we instead forecast $\log(RV_t^{k\text{-minute}})$. The k -minute volume at time t is denoted as $V_t^{k\text{-minute}}$, and is calculated as the number of updates in the orderbook caused by trades, that occurred up to k minutes prior to t . Note that trading volume is not a stationary time-series, as it generally increases over time. Therefore, we forecast the percentage change in volume defined as

$$\delta_t^{k\text{-minute}} = \log \frac{V_t^{k\text{-minute}}}{V_{t-K}^{k\text{-minute}}}. \quad (5.15)$$

5.4.2 Loss Functions and Evaluation Metrics

Denote by $\{y_t\}_{t=1}^T$ the labels of the true values to be forecasted (in our case this is either the returns, realized volatility or trading volume), and by $\{\hat{y}_t\}_{t=1}^T$ the forecasts given by the model. When forecasting trading volume, we use the Mean Absolute Error loss function

$$\text{MAE} = \frac{1}{T} \sum_{t=1}^T |y_t - \hat{y}_t|. \quad (5.16)$$

We also use the Mean Squared Error which is equal to

$$L_{\text{MSE}} = \frac{1}{T} \sum_{t=1}^T (y_t - \hat{y}_t)^2. \quad (5.17)$$

In contrast to trading volume, financial returns are intrinsically much harder to forecast. Furthermore, typical loss functions, such as the MSE, penalize predictions that have a large deviation. However, when forecasting returns, predictions that are correct in sign but wrong in magnitude are much less detrimental than those that are wrong in sign. Therefore, using a variable with economic significance would be a sensible approach. While the Sharpe Ratio would be a natural choice, it is not differentiable with respect to $\hat{y}_t$ due to the usage of the sign function. To circumvent this, let $\epsilon > 0$ be a very small constant. We consider

$$L_{\text{SR}} = -\frac{\text{mean}(\{\tilde{R}_t\}_{t=1}^T)}{\left(\text{sd}(\{\tilde{R}_t\}_{t=1}^T) + \epsilon\right)}, \quad (5.18)$$

where

$$\tilde{R}_t = y_t \tanh\left(\frac{1}{\epsilon} \hat{y}_t\right). \quad (5.19)$$

Equation (5.18) is a slight simplification of the negative annualized Sharpe Ratio, but it is also differentiable, as it does not make use of the sign function. The loss function we use for returns is a combination of these two loss functions

$$L_{\text{mixed}} = L_{\text{MSE}} + \alpha L_{\text{SR}}, \quad (5.20)$$

where α is another hyperparameter to be chosen. Similar loss functions have been used before, for example by [254] and [167].

We also use a custom Quasi-Likelihood (QLIKE) loss function for volatility, defined as

$$L_{\text{QLIKE}} = \frac{1}{T} \sum_{t=1}^T \frac{\exp y_t}{\exp \hat{y}_t} - (y_t - \hat{y}_t) - 1, \quad (5.21)$$

where the -1 term can be omitted for training. QLIKE works independently of the volatility's scale. QLIKE is also used for evaluation of the results, alongside the Heteroskedasticity Adjusted Mean Squared Error (HMSE), defined as

$$\text{HMSE} = \frac{1}{T} \sum_{t=1}^T \left(1 - \frac{\exp \hat{y}_t}{\exp y_t} \right)^2. \quad (5.22)$$

Note that both QLIKE and HMSE have both been assessed to be robust measures for volatility forecasting (e.g by [191]).

5.4.3 Predictive Features

In the following paragraphs, we list the features used in our predictive models. Most of these are variations of predictors that are widely used by practitioners. Denote by $P_t^{(B)}$ (resp., $P_t^{(A)}$) the top-of-book bid (resp., ask) price at time t . Also, denote by $q_t^{(B)}$ ($q_t^{(A)}$) the top-of-book bid (ask) volume, respectively.

We consider the following features

- The Order-Flow Imbalance (OFI) formulated by [66] is used for 5, 10, 30, 60, 90, 120, 180, 270 and 360 minutes. Denote by $E(t, k)$ the set containing the events that occur between time point t and up to k minutes before it. Then, the contribution of each event in $E(t, k)$ calculated as

$$\begin{aligned} e_n = & 1 \left(P_n^{(B)} \geq P_{n-1}^{(B)} \right) q_n^{(B)} - 1 \left(P_n^{(B)} \leq P_{n-1}^{(B)} \right) q_{n-1}^{(B)} \\ & - 1 \left(P_n^{(A)} \leq P_{n-1}^{(A)} \right) q_n^{(A)} + 1 \left(P_n^{(A)} \geq P_{n-1}^{(A)} \right) q_{n-1}^{(A)}. \end{aligned} \quad (5.23)$$

The OFI is then computed as

$$\text{OFI}_t^{(k)} = \sum_{n \in E(t, k)} e_n. \quad (5.24)$$

- k -minute returns and realized volatility of the preceding 5, 10, 30, 60, 90, 180 and 390 minutes, computed using Equations (5.10) and (5.14).
- k -minute positive and negative realized semivolatility of the preceding 5, 10, 30, 60, 90, 180 and 391 minutes, computed as

$$\text{RV-pos}_t^{k\text{-minute}} = \sum_{j=0}^{K-1} \left(R_{t-j}^{1\text{-minute}} \right)^2 \mathbf{1} \left(R_{t-j}^{1\text{-minute}} \geq 0 \right). \quad (5.25)$$

$$\text{RV-neg}_t^{k\text{-minute}} = \sum_{j=0}^{K-1} \left(R_{t-j}^{1\text{-minute}} \right)^2 \mathbf{1} \left(R_{t-j}^{1\text{-minute}} \leq 0 \right). \quad (5.26)$$

- Exponentially weighted return and realized volatility with weight $w \in \{0.75, 0.9, 0.975, 0.99, 0.999\}$, computed as

$$R_t^{(w)\text{-Weighted}} = (1 - w) \sum_{j=0}^{1440} w^j R_{t-j}^{1\text{-minute}}, \quad (5.27)$$

$$\text{RV}_t^{(w)\text{-Weighted}} = (1 - w) \sum_{j=0}^{1440} w^j \left(\text{RV}_{t-j}^{1\text{-minute}} \right)^2. \quad (5.28)$$

- Delta-volume, based on the number of trades occurred in the last 5, 10, 30, 60, 90, 180, 240, 360 and 1440 minutes, as shown in Equation (5.15)
- Binary indicators for the day of the week (Monday, Tuesday, ..., Saturday).
- Binary indicators for the time of day (Hour_00, Hour_01, ..., Hour_22).

5.4.4 Baselines

To compare and assess the potential improvement of the proposed pipeline and the utilization of the forecasts of other instruments, we consider a number of baselines, which are conventional machine learning methods. We include the Ordinary Least Squares (OLS), the LASSO model, and the Principal Component Regression (PCR). These simple methods are necessary to ensure that more complex methods do not overfit the data, especially in the case of forecasting returns. For the LASSO model, we choose the regularization hyperparameter using 10-fold cross-validation, while for PCR, we select the minimum number of principal components that retains 90% of the variance.

Among ensemble models, we consider Random Forests as well as XGBOOST, as both methods are widely used in a variety of applications. For XGBOOST, we use

Layer Name	Layer Units	Activation Function	Other parameters
LSTM	64	tanh	Sequence Length = 12
Dense Layer 1	32	tanh	
Dense Layer 2	16	tanh	
GCN	12 Graphs	tanh	
Output Layer	1	lienar	
Regularization: L1 with parameter 10^{-5}			
Training: ADAM with step size equal to 5e-04 (other parameters set to default) 120 epochs are used for the initial training, and 25 epochs used each time the window is rolled.			
Lookback Window: 20000			
Rolling Window Length: 1500			

Table 5.2: Details for the GCN-LSTM used for the forecasting of Returns, Volatility and Trading Volume.

the default values of the 'xgboost' package in Python, while for Random Forests, we determine the number of trees using cross-validation.

We also consider an Artificial Neural Network (ANN) and an LSTM model as simpler deep learning alternatives. Both of these models follow the same architecture as the GCN model we use. The LSTM model is identical to the GCN-LSTM architecture but without the GCN layer, while the ANN is the same as the LSTM model, without the LSTM layer.

Lastly, we consider a naive forecast. For returns, this is equivalent to holding a long position in the product considered during all trading periods. For volatility, we set the naive forecast value equal to the observed value of the previous period. For differences in trading volume, we set all naive forecasts equal to zero.

5.4.5 Implementation Details

The general architecture of GCN-LSTM is shown in Figure 5.2. The model is implemented using the tensorflow library in Python. A summary of the model's components is provided in Table 5.2.

5.5 Results

5.5.1 Forecasting Returns

The results for forecasting the returns of ES-Mini and VX futures are shown in Tables 5.3 and 5.4, respectively. It should be noted that for returns, we consider data points for training only where there is sufficient liquidity (more than 1) and a "rea-

sonable” spread⁵ for at least one of ES_1 or VX_1. In addition, for the evaluation of each individual product, we only consider data points with sufficient liquidity and a “reasonable” spread for each.

It can be clearly seen that deep learning methods outperform the non-deep learning methods for ES_1 and ES_2 but are outperformed by linear methods for ES_3 and ES_4. As the number of points used for the evaluation of ES_3 and ES_4 is small, these results are less robust. GCN-LSTM is the best-performing deep learning method for ES_1-ES_3, and the results are very similar across the products, demonstrating the effectiveness of propagating forecasts among the different products. Importantly, GCN-LSTM is the highest-performing method among all methods for ES_1 and ES_2. The cumulative P&L plot is shown in Figure 5.6.

For the VX futures, the GCN-LSTM method achieves the highest Sharpe Ratio for six out of eight products, scoring the second highest for the other two. Its PPD is either the highest or second highest for four out of eight scores, while in all cases it is above nine. Furthermore, for all products, GCN-LSTM outperforms ANN and LSTM in both SR and PPD, clearly demonstrating the improvement in forecasts resulting from including the GCN layer in our architecture. It should be noted that in most cases, LASSO and PCR both outperform LSTM and ANN, and they also outperform GCN-LSTM in terms of PPD.

In Table 5.5, we change the setting by considering different horizons. Even though for these cases, trainable points are fewer, results can still be promising. For the most liquid products (ES_1, VX_1, and VX_2), there are a number of cases where we attain SR values higher than 1.5. The highest performances are generally observed for the daily results, with 9 out of 12 products resulting in SR values higher than 1.

5.5.2 Forecasting of Volatility

The results for forecasting the realized volatility of ES-Mini and VX futures are shown in Table 5.6 and Table 5.7, respectively. In this case, deep learning methods outperform the non-deep learning methods for all products considered in terms of QLIKE. In most cases, GCN-LSTM and LSTM have similar performance, with the former generally outperforming the latter. GCN-LSTM is the best-performing method for all SPX products and also for 5 out of 7 products for VIX, being the second-best performer in the other two cases.

⁵We consider a spread “reasonable” if it is below 15 basis points for the SPX futures, and 25 basis points for the VIX futures, respectively.

Returns 'Tradable' Periods	ES_1 11236		ES_2 2487		ES_3 287		ES_4 37	
	SR	PPD	SR	PPD	SR	PPD	SR	PPD
Naive	0.051	0.188	0.054	0.198	0.047	0.219	-0.012	-0.079
LM	0.109	0.401	-0.512	-1.912	-0.1	-0.583	0.08	0.73
LASSO	0.372	1.336	0.432	1.605	2.686	16.342	4.288	42.011
LPCA	0.438	1.582	0.449	1.727	2.675	16.349	4.214	40.744
Random Forests	0.663	2.391	-0.235	-0.855	-0.062	-0.329	0.398	3.265
XGBoost	0.191	0.718	-0.33	-1.227	-0.019	-0.104	0.128	1.018
ANN	1.048	7.823	0.913	7.102	0.918	6.827	0.545	4.196
LSTM	1.281	9.591	1.082	8.109	0.947	8.285	1.301	9.735
GCN-LSTM	1.465	10.941	1.383	9.865	1.341	10.773	1.283	9.471

Table 5.3: Predictive performance of different methods for hourly return forecasting of SPX futures. Here, SR and PPD correspond to Sharpe Ratio and Profit per Dollar respectively. In red (blue) we highlight the highest (second highest) performing method per product.

Returns Tradable Periods	VX_1 6182		VX_2 6243		VX_3 5222		VX_4 4459		VX_5 3798		VX_6 3277		VX_7 792		VX_8 54	
	SR	PPD	SR	PPD	SR	PPD	SR	PPD	SR	PPD	SR	PPD	SR	PPD	SR	PPD
Naive	-0.782	-12.094	-0.455	-5.273	-0.461	-4.048	-0.281	-2.073	-0.324	-2.13	-0.477	-2.776	-0.443	-2.666	-0.418	-2.899
LM	0.339	5.217	0.542	6.344	0.162	1.455	0.045	0.353	0.329	2.136	0.323	1.962	0.358	2.219	-0.237	-1.686
LASSO	0.71	10.901	0.651	7.653	1.527	14.282	0.754	5.896	1.075	14.243	1.099	9.882	1.52	15.72	1.629	11.847
LPCA	1.202	18.43	1.16	14.157	1.362	12.305	1.269	9.916	1.196	13.237	1.195	12.437	1.562	16.044	1.459	12.118
Random Forests	0.69	10.854	0.343	4.024	0.19	1.692	0.095	0.742	0.223	1.519	-0.101	-0.628	-0.0	-0.001	0.422	3.167
XGBOOST	-0.074	-1.16	0.512	5.96	-0.003	-0.027	-0.061	-0.472	0.647	4.49	0.234	1.389	-0.188	-1.232	-0.146	-1.067
ANN	0.729	5.564	0.588	4.278	0.59	4.405	0.933	7.166	1.118	8.396	0.458	3.491	0.171	1.295	-0.127	-0.972
LSTM	0.995	7.415	1.287	9.585	1.24	9.207	1.302	9.658	1.125	8.518	1.326	9.957	1.291	9.655	1.197	9.042
GCN-LSTM	2.488	62.657	2.403	51.649	1.59	31.427	1.496	21.859	1.095	14.430	1.037	11.495	1.259	9.050	0.339	3.816

Table 5.4: Predictive performance of different methods for hourly return forecasting of VX futures. Here, SR and PPD correspond to Sharpe Ratio and Profit per Dollar respectively. In red (resp., blue) we highlight the highest (resp., second highest) performing method for each product.

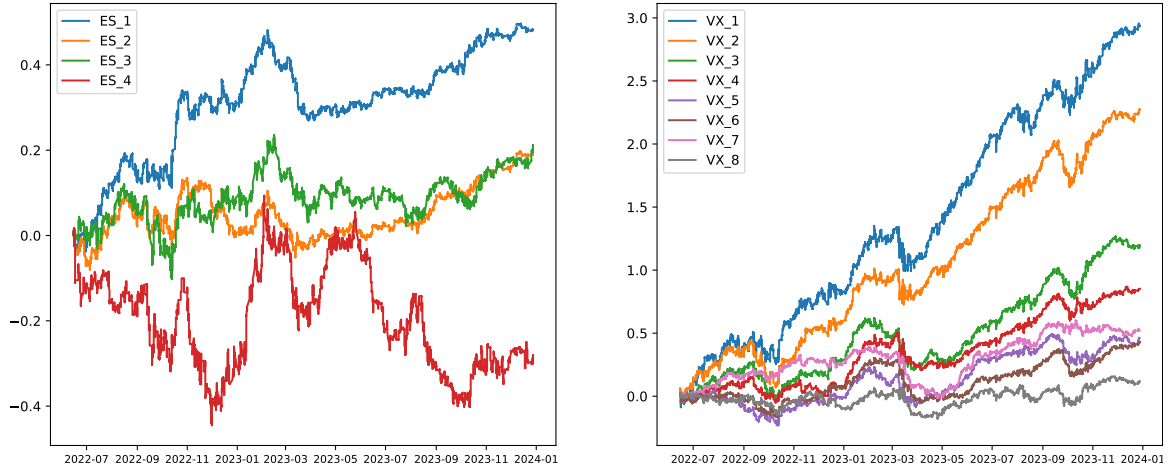


Figure 5.6: Resulting cumulative P&L plot for the hourly return forecasting problem.

Results are quite different when considering the HMSE metric. In this case, deep learning methods appear to perform worse than many of the other considered methods. The best-performing method appears to be Random Forests, being the only one to outperform the naive forecasts.⁶

There appears to be a trade-off between achieving high QLIKE and high HMSE. QLIKE shows the ability of a model to correctly identify the subsequent levels of volatility, while HMSE focuses on the ability of models to have relatively low deviations from the actual predictability. This distinction makes it unclear whether the improvement is significant.

Thus, as GCN-LSTM seems to outperform LSTM for 9 out of 11 products in terms of QLIKE and for 8 out of 11 products in terms of HMSE, the propagation step seems to result in an improvement.

5.5.3 Forecasting Trading Volume

The results for forecasting ES-Mini and VX trading volume are shown in Table 5.8 and Table 5.9, respectively. In terms of MAE, GCN-LSTM is the best-performing method for 3 out of 4 SPX products and is either the best or second-best method for 5 out of 8 VIX products. Importantly, it is the best-performing method for the three most liquid products: ES_1, VX_1, and VX_2. It outperforms LSTM for 3 out of 4 SPX products and the 3 VIX products closer to expiration, but is outperformed by LSTM for 5 out of 8 VIX products.

⁶While not reported here for brevity, results using the Mean Squared Error as an evaluation metric appear to give similar results to HMSE.

	3-Hour		4-Hour		6-Hour		Day	
	SR	PPD	SR	PPD	SR	PPD	SR	PPD
ES_1	0.717	5.196	1.649	13.064	1.074	7.233	1.599	9.269
ES_2	-0.088	-0.654	1.301	9.366	0.665	4.777	1.98	11.509
ES_3	-0.436	-4.186	1.45	12.395	-0.066	-0.551	1.591	12.288
ES_4	-0.426	-5.692	-0.587	-7.789	-0.883	-9.906	-0.404	-3.694
VX_1	-1.158	-28.283	2.936	74.197	1.219	28.802	2.245	40.572
VX_2	3.225	63.552	2.033	39.333	1.949	35.612	1.232	18.686
VX_3	-0.370	-5.898	2.562	41.811	2.313	35.868	0.74	9.623
VX_4	-0.595	-8.055	0.454	6.619	1.459	20.54	1.032	12.019
VX_5	-2.547	-31.665	2.115	28.18	2.146	26.684	1.299	13.704
VX_6	-1.184	-12.915	3.093	35.465	0.708	7.438	2.387	23.377
VX_7	-0.657	-7.909	0.182	2.131	0.589	6.651	0.948	9.677
VX_8	-1.995	-23.842	-0.538	-7.012	1.748	21.197	1.286	14.79

Table 5.5: This table shows the predictive performance for return forecasting at different horizons.

Volatility	ES_1		ES_2		ES_3		ES_4	
	QLIKE	HMSE	QLIKE	HMSE	QLIKE	HMSE	QLIKE	HMSE
Naive	7.68e-01	3.60e-11	1.27e+00	5.80e-11	3.20e+01	2.35e-08	3.74e+03	7.54e-08
LM	7.39e-01	2.38e-09	1.38e+00	2.07e-09	1.43e+01	9.54e-08	2.13e+06	1.38e-03
LASSO	7.35e-01	9.77e-10	1.38e+00	8.10e-11	1.16e+01	3.10e-07	2.78e+02	1.28e-05
LPCA	1.24e+00	9.12e-09	1.38e+00	1.33e-09	9.49e+00	1.88e-08	3.37e+02	8.03e-03
Random Forests	2.56e+00	4.20e-11	1.71e+00	5.20e-11	1.45e+01	1.60e-08	1.98e+02	5.02e-08
XGBOOST	1.79e+00	1.93e-10	1.54e+00	7.86e-10	8.04e+00	9.33e-06	2.42e+04	5.24e-03
ANN	3.95e-07	3.20e-06	2.51e-05	1.86e-04	7.01e-07	5.60e-06	3.37e-07	2.74e-06
LSTM	1.30e-08	1.02e-07	3.02e-09	2.00e-08	1.47e-07	1.13e-06	1.82e-07	1.45e-06
GCN-LSTM	1.00e-09	1.20e-08	3.00e-09	2.10e-08	1.27e-07	9.75e-07	5.00e-08	3.98e-07

Table 5.6: Predictive performance of different methods for hourly volatility forecasting for SPX futures. Here, QLIKE and HMSE correspond to Quasi-likelihood and Heteroskedasticity Adjusted Mean Squared Error. In red (blue) we highlight the highest (second highest) performing method per product.

Volatility	VX_1		VX_2		VX_3		VX_4		VX_5		VX_6		VX_7	
	QLIKE	HMSE	QLIKE	HMSE	QLIKE	HMSE	QLIKE	HMSE	QLIKE	HMSE	MSE	MAE	QLIKE	HMSE
Naive	1.74e+00	9.45e-08	2.75e+00	5.96e-08	3.24e+00	3.68e-08	3.76e+00	2.82e-08	3.79e+00	2.20e-08	3.28e+00	1.50e-08	5.34e+00	2.58e-08
LM	1.93e+00	4.76e-07	2.81e+00	5.91e-05	3.09e+00	1.76e-07	3.31e+00	7.09e-08	3.61e+00	6.48e-08	3.05e+00	3.50e-08	5.17e+00	6.57e-06
LASSO	1.94e+00	5.82e-07	2.84e+00	6.84e-05	3.17e+00	2.35e-07	3.31e+00	5.98e-08	3.68e+00	1.28e-07	3.10e+00	5.04e-08	5.02e+00	1.28e-05
LPCA	1.59e+00	7.94e-06	2.20e+00	3.76e-05	2.26e+00	7.97e-03	2.36e+00	7.49e-05	2.23e+00	9.59e-07	2.07e+00	3.15e-07	2.51e+00	2.16e-08
Random Forests	9.84e-01	6.10e-08	1.53e+00	3.69e-08	1.83e+00	2.30e-08	1.90e+00	1.89e-08	2.03e+00	1.63e-08	1.82e+00	1.01e-08	2.76e+00	1.51e-08
XGBOOST	1.45e+00	1.88e-07	2.79e+00	1.79e-06	1.70e+00	9.37e-07	2.13e+00	2.17e-06	2.13e+00	9.90e-07	2.01e+00	6.82e-07	2.51e+00	8.07e-05
ANN	6.75e-07	5.34e-06	2.08e-07	1.67e-06	2.56e-07	2.04e-06	1.78e-07	1.43e-06	3.38e-07	2.86e-06	2.96e-07	2.38e-06	7.10e-07	5.71e-06
LSTM	1.27e-07	1.01e-06	1.01e-07	8.04e-07	4.00e-08	3.16e-07	3.10e-08	2.52e-07	1.80e-08	1.48e-07	1.20e-08	9.20e-08	2.10e-08	1.71e-07
GCN-LSTM	1.15e-07	9.14e-07	5.50e-08	4.37e-07	2.40e-08	1.92e-07	8.30e-08	6.65e-07	1.30e-08	1.02e-07	9.00e-09	7.20e-08	1.65e-07	1.31e-06

Table 5.7: Predictive performance of different methods for hourly volatility forecasting for VIX futures. Here, QLIKE and HMSE correspond to Quasi-likelihood and Heteroskedasticity Adjusted Mean Squared Error. In red (blue) we highlight the highest (second highest) performing method per product.

For MSE, the results differ significantly, with GCN-LSTM not being the best-performing method for any product. This outcome is potentially due to the use of MAE as the loss function for this architecture, which seems to be outperformed by other methods that tend to produce forecasts with lower variance (closer to the center). In fact, both LSTM and GCN-LSTM are outperformed by the naive models in a number of cases.

Volume	ES_1		ES_2		ES_3		ES_4	
	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
Naive	1.617e-03	6.625e-01	5.171e-03	8.256e-01	4.954e-03	8.556e-01	9.175e-04	2.989e-01
LM	1.370e-02	4.189e-01	6.184e-02	6.146e-01	1.056e-01	6.926e-01	3.967e-02	2.390e-01
LASSO	8.385e+00	8.473e+00	5.379e+00	5.978e+00	9.656e-01	1.518e+00	4.688e-02	3.238e-01
LPCA	2.000e-03	4.190e-01	1.260e-01	6.640e-01	1.140e-01	8.020e-01	6.000e-03	2.390e-01
Random Forests	3.278e-02	3.679e-01	8.158e-04	6.332e-01	6.896e-02	7.376e-01	1.204e-02	2.242e-01
XGBOOST	2.285e-02	3.902e-02	4.226e-03	6.954e-01	3.734e-02	7.611e-01	1.826e-02	2.297e-01
ANN	3.951e-04	7.509e-01	5.713e-05	9.547e-01	6.464e-03	6.758e-01	1.705e-03	2.037e-01
LSTM	1.200e-03	7.520e-01	2.786e-04	9.568e-01	6.766e-03	6.795e-01	6.861e-04	2.090e-01
GCN-LSTM	1.733e-04	7.518e-01	2.992e-04	9.589e-01	4.118e-04	6.859e-01	9.797e-05	2.090e-01

Table 5.8: Predictive performance of different methods for hourly trading volume forecasting for SPX futures. Here, MAE and MAE correspond to the Mean Absolute Error and Mean Squared Error respectively. In red (blue) we highlight the highest (second highest) performing method per product.

Volume	VX_1		VX_2		VX_3		VX_4		VX_5		VX_6		VX_7		VX_8	
	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
Naive	6.532e-04	1.175e+00	1.175e-03	1.301e+00	1.788e-04	1.386e00	5.923e-04	1.490e00	6.767e-03	1.659e00	6.188e-03	1.640e+00	5.030e-03	1.581e00	1.432e-03	1.271e00
LM	3.148e-02	9.600e-01	3.676e-02	9.616e-01	2.776e-02	1.005e00	4.056e-02	1.097e00	4.741e-02	1.188e00	4.545e-02	1.162e00	3.882e-02	1.021e00	6.887e-02	8.977e-01
LASSO	5.786e+00	6.138e+00	5.449e+00	5.894e+00	4.314e00	4.874e00	3.625e00	4.278e00	2.905e00	3.733e00	2.440e00	3.361e00	1.918e+00	2.882e00	1.490e00	2.368e00
LPCA	2.100e-02	9.000e-01	4.400e-02	9.650e-01	1.100e-02	1.083e00	4.300e-02	1.246e00	2.000e-03	1.366e00	3.600e-02	1.338e00	2.100e-02	1.202e00	0.000e00	1.027e00
Random Forests	5.763e-02	7.550e-01	3.673e-02	8.570e-01	9.401e-03	9.759e-01	1.610e-02	1.124e00	3.573e-03	1.183e00	5.550e-03	1.227e00	6.956e-03	1.144e00	9.397e-02	8.967e-01
XGBOOST	3.257e-02	7.920e-01	1.413e-02	9.162e-01	2.750e-02	1.021e00	5.050e-02	1.147e00	8.728e-02	1.309e00	3.588e-03	1.243e00	3.490e-02	1.281e00	1.467e-01	9.500e-01
ANN	6.158e-04	1.253e00	2.799e-03	1.369e00	3.338e-03	1.412e00	1.475e-03	1.479e00	5.859e-04	1.526e00	1.203e-03	1.503e00	2.044e-03	1.415e00	2.099e-03	1.126e-03
LSTM	2.583e-04	1.256e+00	2.106e-03	1.370e+00	1.684e-03	1.417e00	1.232e-04	1.482e00	9.453e-04	1.529e00	1.960e-03	1.506e00	2.265e-03	1.424e00	3.685e-05	1.139e00
GCN-LSTM	1.378e-04	1.256e00	9.283e-04	1.370e00	1.120e-03	1.417e00	4.977e-04	1.482e00	1.225e-03	1.529e00	2.081e-03	1.507e00	3.075e-03	1.417e00	2.692e-04	1.139e00

Table 5.9: Predictive performance of different methods for **hourly trading volume** forecasting for VIX futures. MAE and MAE correspond to the Mean Absolute Error and Mean Squared Error respectively. In red (blue) we highlight the highest (second highest) performing method per product.

5.5.4 Feature Importance

While it can be challenging to extract feature importance scores for deep learning models, the LASSO coefficients obtained during the baseline step act as a readily available and interpretable feature importance metric. Across the 14 instruments,⁷ we obtained the median rank of each feature in terms of the absolute value of the fitted coefficients.

For the return forecasting task, only 6 of the features are deemed significant for more than half the cases, as return forecasting is inherently a high error-to-

⁷For trading volume, there are only 12 products, as the indices are not tradable.

signal ratio problem. These features include the 30, 240, and 360-minute returns ($R_t^{30\text{-minute}}, R_t^{240\text{-minute}}, R_t^{360\text{-minute}}$), the 30-minute negative semi-variance ($\text{RV-neg}_t^{30\text{-minute}}$), as well as the exponentially weighted return and realized volatility with a parameter equal to 0.75 ($R_t^{(0.75)\text{-weighted}}, \text{RV}_t^{(0.75)\text{-weighted}}$).

The exponentially weighted realized volatilities (`weigh_rv`) appear to be among the 20 most important features for all three tasks. For volatility, all `weigh_rv` features are included in the list of most important features, with `weigh_rv_1.0` having the highest median rank and being a non-trivial predictor for 11 products. The 10 and 60-minute realized volatilities are also in the list of the most important features.

For trading volume, some of the most important features unsurprisingly include the preceding trading volumes, with the 90-minute feature scoring the highest median rank. The time-of-day feature also scores high for trading volume and volatility, with many binary hour indicator features after 12:00 appearing in the top 20. Lastly, most days are included in the top 20 for trading volume, with the Monday binary indicator scoring highly for volatility.

5.5.5 Different Configurations

In Table 5.11, we show the results obtained by GCN-LSTM for `ES_1` and `VX_1` for each forecasted quantity. These results are compared with alternative configurations of the same model. In particular, instead of using 12 graphs, we use the Contemporaneous Weighted, Contemporaneous Unweighted, Lagged Weighted, and Lagged Unweighted separately (3 graphs in each case). We further explore different loss functions, including the MAE, MSE, and SR loss functions (where applicable). We also consider a configuration with ‘Non-parallel modules’, meaning that the weights for the LSTM and dense layers are shared across nodes. Note that in the previous subsections, we also considered the LSTM and ANN models, which were simplifications of GCN-LSTM without the GCN layer. We draw the overall conclusions for the different tasks considered.

For returns, we clearly observe that using different loss functions yields significantly worse results. Using all the graphs appears to be beneficial as it outperforms all other cases where we only use a subset, with Contemporaneous Unweighted being the best and second-best performing for `ES_1` and `VX_1`, respectively, among the four types of graphs.

For volatility, the proposed setup achieves the highest performance in terms of QLIKE for `VX_1` and the second-best performance for `ES_1` (outperformed by Contemporaneous Weighted). Using MSE as the loss function clearly diminishes

	Returns		Volatility		Volume	
	Median Rank	Non-zero Coefficients	Median Rank	Non-zero Coefficients	Median Rank	Non-zero Coefficients
OFI_360	1.0	4	47.0	9	43.0	12
OFI_1440	1.0	3	1.0	6	19.0	12
weigh_rets_0.75	90.0	9	30.5	7	35.5	12
weigh_rets_0.9	1.0	2	30.5	7	44.0	11
weigh_rets_0.975	1.0	1	1.0	2	45.0	11
weigh_rets_0.99	1.0		1.0	6	22.5	8
weigh_rets_0.999	1.0	2	21.5	7	46.5	12
weigh_rets_1.0	1.0	1	15.0	7	41.5	10
weigh_rv_0.75	15.5	7	87.0	13	35.0	12
weigh_rv_0.9	1.0	5	78.0	8	63.5	12
weigh_rv_0.975	1.0	4	75.5	8	83.0	12
weigh_rv_0.99	1.0	1	75.5	10	76.0	11
weigh_rv_0.999	1.0	3	82.5	8	64.0	11
weigh_rv_1.0	1.0	5	91.0	11	54.5	11
ret_5	1.0	5	28.0	7	17.5	12
ret_10	1.0	3	15.5	7	20.0	11
ret_30	32.0	7	61.5	10	17.5	11
ret_60	1.0	6	1.0	6	33.5	11
ret_90	1.0	5	41.0	10	21.5	12
ret_180	1.0	3	1.0	5	20.0	10
ret_240	83.5	9	45.0	8	16.0	11
ret_360	89.5	12	36.0	8	23.0	11
ret_1440	1.0	1	1.0	6	15.0	11
rv_5	1.0	1	36.0	7	1.0	6
rv_10	1.0	2	76.0	9	3.5	7
rv_30	1.0	2	1.0	4	10.0	7
rv_60	1.0	0	83.5	9	1.0	6
rv_90	1.0	3	1.0	3	9.0	8
rv_pos360	1.0	3	1.0	3	33.5	10
rv_pos1440	1.0	3	75.5	10	19.0	8
rv_neg5	1.0	1	22.5	7	17.5	11
rv_neg10	1.0	3	18.0	7	14.0	7
rv_neg30	87.0	8	1.0	5	23.5	11 F
num_trades_5	1.0	3	1.0	3	49.5	12
num_trades_10	1.0	1	1.0	2	74.0	12
num_trades_30	1.0	1	1.0	5	95.0	12
num_trades_60	1.0	2	28.0	7	84.0	12
num_trades_90	1.0	1	1.0	3	91.0	11
num_trades_180	1.0	1	1.0	3	73.5	12
num_trades_240	1.0	0	1.0	2	43.0	12
num_trades_360	1.0	1	1.0	2	84.5	12
num_trades_1440	1.0	2	1.0	5	85.0	12
Monday	1.0	5	64.5	10	64.5	12
Tuesday	1.0	1	1.0	6	81.5	12
Wednesday	1.0	1	19.5	7	78.0	12
Thursday	1.0	1	31.5	8	78.5	12
Friday	1.0	3	42.5	8	76.0	12
Saturday	1.0	0	1.0	0	1.0	0
Hour_11	1.0	3	1.0	6	56.0	12
Hour_12	1.0	4	60.5	9	69.5	12
Hour_13	1.0	4	82.0	12	84.5	12
Hour_14	1.0	2	82.5	11	75.5	12
Hour_15	1.0	6	63.0	12	64.0	12
Hour_16	1.0	3	41.0	8	62.5	12
Hour_17	1.0	1	44.0	8	62.0	12
Hour_18	1.0	3	53.5	10	68.5	12
Hour_19	1.0	6	72.0	10	76.0	12
Hour_20	1.0	6	42.5	8	63.0	12
Hour_21	1.0	5	69.5	12	88.5	12
Hour_22	1.0	3	16.5	7	71.5	12

Table 5.10: Feature important metric extracted from LASSO models for the three major forecasting problems studied. In red, we show the 20 features with the highest median rank. For presentation purposes, we exclude some features which are not highly important for any of the three problems (the excluded Features include: The OFI features below 240, realized volatility of 180 minutes or more, positive realized volatility of between 5 and 240 minutes, negative realized volatility of above 90 minutes, and hours 1-10.)

	Returns				Volatility				Volume			
	ES_1		VIX_1		ES_1		VIX_1		ES_1		VIX_1	
	SR	PPD	SR	PPD	QLIKE	HMSE	QLIKE	HMSE	MSE	MAE	MSE	MAE
Contemporaneous Weighted	0.766	6.910	0.786	10.013	1.699e-10	3.397e-10	1.342e-07	2.683e-07	7.524e-01	1.745e-03	1.256e00	7.950e-04
Contemporaneous Unweighted	1.184	6.484	0.765	5.955	7.694e-08	1.539e-07	1.659e-07	3.317e-07	7.524e-01	2.366e-03	1.256e00	2.784e-03
Lagged Weighted	0.510	2.694	0.688	8.895	1.576e-04	2.467e-04	2.791e-04	4.113e-04	7.520e-01	8.715e-04	1.256e00	5.011e-04
Lagged Unweighted	0.640	3.125	0.615	7.647	4.697e-09	9.397e-09	5.422e-07	1.084e-06	7.519e-01	8.872e-04	1.256e00	5.011e-04
Loss Function: MSE	0.368	1.223	0.332	4.328	5.902e-04	1.811e-03	6.666e-07	1.353e-06	7.719e-01	4.707e-04	1.271e00	1.730e-03
Loss Function: MAE	-0.138	-0.463	-0.316	-4.027	2.296e-09	4.591e-09	1.278e-07	2.554e-07	NA	NA	NA	NA
Loss Function: SR	0.007	0.023	-0.254	-3.272	NA	NA	NA	NA	NA	NA	NA	NA
Non-parallel modules	0.213	0.710	0.486	6.281	1.523e-07	3.031e-07	1.740e-07	3.459e-07	7.706e-01	1.420e-04	1.270e00	2.001e-05
Used model	1.465	13.941	1.328	10.243	1.00e-09	1.20e-08	1.15e-07	9.140e-07	7.518e-01	1.733e-04	1.256e00	1.346e-04

Table 5.11: Predictive performance of different methods for hourly return, volatility and volume forecasting for ES_1 and VX_1 futures. In red (blue) we highlight the highest (second highest) performing method per product.

performance, while using MAE appears to perform almost comparably to the QLIKE loss.

For volume, the proposed setup seems to be outperformed by the configuration where weights are shared across nodes, which is most likely a natural outcome of higher correlation among nodes. Furthermore, the lagged networks appear to mostly outperform the contemporaneous networks.

5.6 Conclusion

We have analyzed the interaction between the different E-mini S&P 500 futures and Cboe Volatility Index futures. The analysis of the graphs in our study provided several key insights into the relationships between the different futures products, offering a visual representation of the correlation structure at play in this market. The clear clustering of SPX and VIX nodes, the prevalence of negative correlations between returns and volatility, and the varied influence of liquidity levels on the connectivity of different graphs all provide valuable insights and highlight the complexity of financial markets. These findings solidify the utility of establishing a cross-asset flow of information for the different forecasting tasks.

By employing a novel architecture of the GCN-LSTM class, we aimed to enhance the predictability for a number of forecasting tasks. We have demonstrated that the proposed GCN-LSTM architecture provides significant improvements in forecasting accuracy over the base LSTM model. The forecasting results for returns in Section 5.5.1 indicate that deep learning methods, particularly GCN-LSTM, generally outperform traditional methods for more liquid products like ES_1 and ES_2. Similarly, Section 5.5.2 and Section 5.5.3 show that the already high performance of the LSTM model for the volatility and volume forecasting tasks, respectively, is further improved by our GCN-LSTM model.

Importantly, in this paper we integrate economically significant loss functions into our model. By using the Sharpe Ratio for financial returns and a quasi-likelihood loss function for volatility, we ensure that the predictions of our model are not only reliable but also economically meaningful. This aligns our forecasting objectives with practical financial decision-making, rendering our approach highly relevant for practitioners in the futures markets.

While our study has provided valuable insights, it also opens up several avenues for future research. One potential direction is the exploration of other types of derivative products, such as options, where larger networks can be constructed as the strike price also varies. The combination of both futures and options data would result in an interesting examination of the resulting graphs, providing insights into the interactions between options and futures contracts. Another interesting angle would be to explore this methodology in a different market, such as the commodities futures market.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

We have conducted four projects that are all centered around forecasting the main financial trading quantities (returns, realized volatility, and trading volume). All three of these quantities are of crucial importance to practitioners, who could utilize a number of tools and valuable insights found throughout this dissertation.

The central theme of the thesis has been the formulation of data-driven approaches for a variety of financial forecasting problems. In large part, this is showcased by the usage of options-based data for forecasting problems in the equity market.

Firstly, in Chapter 2, we used the normalized difference between option volumes corresponding to bullish and bearish speculators, as predictor for the next day returns of the corresponding underlyings. We showed the benefit of decomposing option volumes by market participant, with the Market Maker OVI, scoring an exceptionally good performance in predicting next day returns. The conducted analysis was quite comprehensive, showing its robustness in multiple levels, with the results being confirmed using two different data sets. In addition, the analysis shed light on the specific channels where predictability stems from. Volume corresponding to put options, options with high implied volatility, and options with negative values for Rho, had a higher informational content than call options, options with low implied volatility and options with positive values for Rho respectively. For each market participant, most of the predictability came either from sell or buy orders.

Secondly, in Chapter 4 we have tackled the usage of options data for volatility forecasting. The literature generally concludes that either implied volatility surface data does not improve forecasts or that only certain of its features are useful. We have provided two approaches that were proven successful in providing forecasting power for realized volatility, which demonstrates the need of proper manipulation of

such data. The first approach was based on dimensionality reduction used on the Implied Volatility Surface data generated by the Black Scholes model. The second approach was a framework for extracting the fitted volatility, which we have applied to the Heston and Bates models. As the calibration of option models is used for option pricing, it entails that they are inherently capturing the look ahead risk and volatility of the underlying assets. Our approach effectively capitalizes on that, while simultaneously avoiding the computational and optimizational complexity that is common in calibration processes.

A secondary theme permeating the thesis, is the usage of metrics with economic meaning for the calibration of predictive models, which are often used by academics and practitioners only during the evaluation step. In particular, we have used the Profit and Loss, and Sharpe Ratio not only as evaluation metrics, but also incorporated them in the optimization steps of our models. In Chapter 2, we formulate P&L regression, which is a linear model that maximizes the P&L instead of minimizing the mean squared error, which showed improved results. In Chapter 3, we used P&L to select the optimal model in an online manner, when forecasting ETF returns. Lastly, in Chapter 5 we used the Sharpe Ratio as the objective function for forecasting futures' returns, which was showed to outperform models using Mean Absolute Error and Mean Squared Error as objective functions. In the same project, we also used a quasi-likelihood objective function when forecasting volatility, which also showed significant improvements in performance. These methods show the utility of using problem tailored objective functions.

A third theme is the usage of computationally efficient methodology. This has been the focus of the study given in Chapter 3, where we formulate a generic pipeline tailored for mid-sized financial problems. Importantly, the methodology is non-parametric as there are no assumptions about the form of the relationship between predictors and target due to the model being based on the kNN and GRNN models. This pipeline is especially tailored for low signal to error ratio, mid-sized problems, where models with high expressive power such as deep learning models are often incapable of outperforming simple models. In particular, OFTER is shown to outperform models like LSTM for a large number of problems. OFTER exemplifies a secondary theme of the thesis, which is the usage of computationally efficient methodologies.

As a final step in Chapter 4 we had shown how the individual equity volatility forecasts we obtained could be utilized for forecasting the index VIX. A methodology involving the simultaneous processing of all features across models, would have been impossible and very computationally intensive. This led into the final chapter of the

main text, Chapter 5. In that chapter, we went one step further propagating signals across different instruments with LSTM-GCN, a neural network structure that is able to process temporal structures via the use of a Long Short Term Model, as well as spatial structures via the use of a Graph Convolutional Network. We proposed an iteration of that model, tailored for the joint forecasting of E-mini S&P 500 and CBOE Volatility Index futures. In that setting, the nodes of the graph represent the different instruments of the two underlying instruments. We used LSTM-GCN to forecast all three main financial trading quantities, with promising results for all three, that outperformed LSTM and other baselines.

6.2 Future Work

In each individual chapter, further work opportunities were outlined, including extending the analysis of volume imbalance by market participant shown in Chapter 2 to other markets, extending the analysis of Chapter 3 to include base learners from other machine learning models, further augmenting the models in Chapter 4 using sentiment data and utilizing more complex machine learning methods.

6.2.1 Application of GCN to an Equity Network

A natural continuation of this thesis would be an expansion of the methodology presented in Chapter 5. In particular, a graph could be constructed for the S&P500 stocks. Multiple graphs could be constructed for such a model, including the clustering by sector, using knowledge graphs, as well as using a similar correlation analysis to the one conducted in Chapter 5, as different inputs into a multi-channel GCN layer.

Some modifications would need to be applied to our model. First, as the network would be substantially bigger than the one we worked with in Chapter 5, the parallel module component would need to be dropped. Secondly, an additional GCN layer could become necessary, especially if networks such as supply chain or knowledge graphs were to be used. While a similar setup was conducted by [215], their approach was limited to using daily data and the need of more features such as oscillators.

Such a network could also be utilized to model volatility shocks. A similar research paper is [247] but that did not include the temporal utility of Chapter 5. In addition, instead of a prediction task, this would be a classification task, where a label of 1 would indicate the occurrence of a price jump in a pre-determined time span. Trading

volume forecasting could also be extended to equities, with the methodology of [240], extended via the use of GCN-LSTM.

6.2.2 Extensions of Methodology for Extracting volatility under the Heston Model

While extensions of methodology have been listed for all sections, one part open to most extensions and modifications would be the methodology regarding extracting the current volatility by calibrating a Heston model. As calibration of the Heston model is generally difficult, we focus on a one-dimensional optimization of the current volatility by fixing the two least significant parameters and creating a grid for the other two parameters. We could extend our methodology in a few ways, which we discuss below.

Calibration using deep learning The calibration of option pricing models is yet another area where deep learning has seen use. We could also integrate deep learning into our methodology for calibrating the Heston model.

For example, a promising pipeline is offered by [142], who train an unstacked network for each of the Heston parameters using extracted features from simulated price time series data with [21] and principal component analysis. The training is conducted on data readily simulated using discretization methods. An alternative method could be the usage of the inverse map approach (see, e.g., the discussion in [119]), whereby the calibration process is modeled as a Deep Neural Network, where an increase in computational speed is achieved by replacing the intensive pricing of options with an Artificial Neural Network approximation.

While neither methodology makes the calibration of the Heston model inherently easier, a deep learning alternative could provide several advantages. First, unlike the fast Fourier transform we used, a neural network would only need to be trained once, providing instant estimates for new data, a significant advantage for practitioners. Another advantage is the toolbox accompanied by neural network methods. For example, regularization can be easily utilized, making the model less prone to overfitting.

De-Americanization Pricing American options is inherently difficult because exercising an option does not necessarily happen at the expiration of the product. De-Americanization is the process of converting American option data into European option data using approximations and heuristics. In our methodology, we have only

considered stocks that do not pay dividends, as the conversion process is trivial. For instruments that pay dividends, this is often achieved using binomial trees (see, e.g., [46]). This is already an approximation as an unrealistic parametric model is assumed. Therefore, this is another case where we could incorporate a deep learning method to recover an approximation in converting between European and American option prices by simulating data, potentially from a large variety of parametric models.

Bibliography

- [1] Kenneth r. french - data library. http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html. Accessed: 2023-02-02.
- [2] Daron Acemoglu, Ufuk Akcigit, and William R. Kerr. Innovation Network. CESifo Working Paper Series 6173, CESifo, 9 2016.
- [3] Anup Agrawal, Jeffrey F. Jaffe, and Gershon N. Mandelker. The post-merger performance of acquiring firms: A re-examination of an anomaly. *The Journal of Finance*, 47(4):1605–1621, 1992.
- [4] Huson Ali Ahmed, Ali Hajj Hassan, and Annuar M. D. Nasir. The relationship between trading volume, volatility and stock market returns: A test of mixed distribution hypothesis for a pre- and post crisis on kuala lumpur stock exchange. *Investment management & financial innovations*, 2:146–158, 2005.
- [5] Nesreen K. Ahmed, Amir F. Atiya, Neamat El Gayar, and Hisham El-Shishiny. An empirical comparison of machine learning models for time series forecasting. *Econometric Reviews*, 29(5-6):594–621, 2010.
- [6] Yacine Aït-Sahalia, Chenxu Li, and Chen Xu Li. Closed-form implied volatility surfaces for stochastic volatility models with jumps. *Journal of Econometrics*, 222(1):364–392, May 2021.
- [7] Ahmad Jobran Al-Mahasneh, Sreenatha Anavatti, Matthew Garratt, and Mahardhika Pratama. Applications of general regression neural networks in dynamic systems. In Vahid Asadpour, editor, *Digital Systems*, chapter 8, pages 133–154. IntechOpen, Rijeka, 2018.
- [8] J. Cross Amanda, A. Rohrer Gary, M. Brown-Brandl Tami, P. Cassady Joseph, and N. Keel Brittney. Feed-forward and generalised regression neural networks in modelling feeding behaviour of pigs in the grow-finish phase. *Biosystems*

- Engineering*, 173:124–133, 2018. Advances in the Engineering of Sensor-based Monitoring and Management Systems for Precision Livestock Farming.
- [9] Kaushik I. Amin and Charles M. C. Lee. Option trading, price discovery, and earnings news dissemination*. *Contemporary Accounting Research*, 14(2):153–192, 1997.
 - [10] Torben G. Andersen and Tim Bollerslev. Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. *International Economic Review*, 39(4):885–905, 1998.
 - [11] Joseph H. Anthony. The interrelation of stock and options market trading-volume data. *The Journal of Finance*, 43(4):949–964, 1988.
 - [12] Faheem Aslam, Yasir Tariq Mohmand, Paulo Ferreira, Bilal Ahmed Memon, Maaz Khan, and Mrestyal Khan. Network analysis of global stock markets at the beginning of the coronavirus disease (covid-19) outbreak. *Borsa Istanbul Review*, 20:S49–S61, 2020. Financial markets under growing uncertainty: Information transmission and investment strategies.
 - [13] Francesco Audrino and Simon D. Knaus. Lassoing the har model: A model selection perspective on realized volatility dynamics. *Econometric Reviews*, 35(8-10):1485–1521, 2016.
 - [14] Francesco Audrino, Fabio Sigrist, and Daniele Ballinari. The impact of sentiment and attention measures on stock market volatility. *International Journal of Forecasting*, 36(2):334–357, 2020.
 - [15] Marco Avellaneda, Brian Healy, Andrew Papanicolaou, and George Papanicolaou. Pca for implied volatility surfaces, 2020.
 - [16] Ihsan Badshah. Modeling the dynamics of implied volatility surfaces. *SSRN Electronic Journal*, 02 2009.
 - [17] David Bailey and Marcos Lopez de Prado. The deflated sharpe ratio: Correcting for selection bias, backtest overfitting, and non-normality. *The Journal of Portfolio Management*, 40:94–107, 09 2014.
 - [18] Christine Bangsgaard and Thomas Kokholm. The lead lag relation between vix futures and spx futures. *Journal of Financial Markets*, 67:100851, 2024.

- [19] Bank for International Settlements. Bis statistics explorer: Table d5.1 and d8, May 2021. Data retrieved from <https://stats.bis.org/statx>.
- [20] Yong Bao. Estimation Risk-Adjusted Sharpe Ratio and Fund Performance Ranking under a General Return Distribution. *Journal of Financial Econometrics*, 7(2):152–173, 12 2008.
- [21] Marília Barandas, Duarte Folgado, Leticia Fernandes, Sara Santos, Mariana Abreu, Patrícia Bota, Hui Liu, Tanja Schultz, and Hugo Gamboa. Tsfel: Time series feature extraction library. *SoftwareX*, 11:100456, 2020.
- [22] Ole E. Barndorff-Nielsen, Silvia Kinnebroek, and Neil Shephard. *Measuring downside risk: realised semivariance*, pages 117–136. Oxford University Press, (edited by t. bollerslev, j. russell and m. watson) edition, 2010.
- [23] Ole E. Barndorff-Nielsen and Neil Shephard. Econometric analysis of realized volatility and its use in estimating stochastic volatility models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 64(2):253–280, 2002.
- [24] Ole E. Barndorff-Nielsen and Neil Shephard. Power and Bipower Variation with Stochastic Volatility and Jumps. *Journal of Financial Econometrics*, 2(1):1–37, 01 2004.
- [25] S. Basu. Investment performance of common stocks in relation to their price-earnings ratios: A test of the efficient market hypothesis. *The Journal of Finance*, 32(3):663–682, 1977.
- [26] David S. Bates. Jumps and stochastic volatility: Exchange rate processes implicit in deutsche mark options. *The Review of Financial Studies*, 9(1):69–107, 1996.
- [27] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.*, 15(6):1373–1396, June 2003.
- [28] Austin R. Benson, Paul Liu, and Hao Yin. A simple bipartite graph projection model for clustering in networks, 2020.

- [29] Christoph Bergmeir and Jos M. Ben tez. On the use of cross-validation for time series predictor evaluation. *Information Sciences*, 191:192–213, 2012. Data Mining for Software Trustworthiness.
- [30] Mihir Bhattacharya. Price changes of related securities: The case of call options and stocks. *The Journal of Financial and Quantitative Analysis*, 22(1):1–15, 1987.
- [31] Bruno Biais and Pierre Hillion. Insider and Liquidity Trading in Stock and Options Markets. *The Review of Financial Studies*, 7(4):743–780, 05 2015.
- [32] Barbara Będowska-Sójka and Agata Kliber. Information content of liquidity and volatility measures. *Physica A: Statistical Mechanics and its Applications*, 563:125436, 2021.
- [33] Fischer Black. Fact and fantasy in the use of options. *Financial Analysts Journal*, 31(4):36–72, 1975.
- [34] Fischer Black and Myron Scholes. The pricing of options and corporate liabilities. *Journal of Political Economy*, 81(3):637–654, 1973.
- [35] Friedhelm Bliemel. Theil’s forecast accuracy coefficient: A clarification. *Journal of Marketing Research*, 10(4):444–446, 1973.
- [36] Tim Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3):307–327, 1986.
- [37] Tim Bollerslev, Andrew J. Patton, and Rogier Quaadvlieg. Exploiting the errors: A simple approach for improved volatility forecasting. *Journal of Econometrics*, 192(1):1–18, 2016.
- [38] Maria Rosa Borges. Efficient market hypothesis in european stock markets. *The European Journal of Finance*, 16(7):711–726, 2010.
- [39] Jean-Phillipe Bouchaud and Marc Potters. Financial applications of random matrix theory: a short review. In *The Oxford Handbook of Random Matrix Theory*. Oxford University Press, 09 2015.
- [40] Elie Bouri, Chi Keung Marco Lau, Brian Lucey, and David Roubaud. Trading volume and the predictability of return and volatility in the cryptocurrency market. *Finance Research Letters*, 29:340–346, 2019.

- [41] Leo Breiman and Jerome H. Friedman. Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, 80(391):580–598, 1985.
- [42] R.G. Brown. *Exponential Smoothing for Predicting Demand*. Little, 1956.
- [43] S Bryzgalova, M Pelger, and J Zhu. Forest through the trees: Building cross-sections of stock returns. Working Paper 10.2139/ssrn.3493458, Social Sciences Research Network, September 2020.
- [44] Andrea Bucci. Realized Volatility Forecasting with Neural Networks. *Journal of Financial Econometrics*, 18(3):502–531, 06 2020.
- [45] James R Bunch, Christopher P Nielsen, and Danny C Sorensen. Rank-one modification of the symmetric eigenproblem. *Physica Scripta*, 1978.
- [46] Olena Burkovska, Maximilian Gaß, Kathrin Glau, Mirco Mahlstedt, Wim Schoutens, and Barbara Wohlmuth. Calibration to american options: Numerical investigation of the de-americanization, 2016.
- [47] Linda Canina and Stephen Figlewski. The informational content of implied volatility. *The Review of Financial Studies*, 6(3):659–681, 1993.
- [48] Charles Cao, Zhiwu Chen, and John M. Griffin. Informational content of option volume prior to takeovers. *The Journal of Business*, 78(3):1073–1109, 2005.
- [49] Defu Cao, Yujing Wang, Juanyong Duan, Ce Zhang, Xia Zhu, Conguri Huang, Yunhai Tong, Bixiong Xu, Jing Bai, Jie Tong, and Qi Zhang. Spectral temporal graph neural network for multivariate time-series forecasting, 2021.
- [50] Hieu K. Cao, Han K. Cao, and Binh T. Nguyen. Delafo: An efficient portfolio optimization using deep neural networks. In *Advances in Knowledge Discovery and Data Mining*, pages 623–635, Cham, 2020. Springer International Publishing.
- [51] Mark M. Carhart. On persistence in mutual fund performance. *The Journal of Finance*, 52(1):57–82, 1997.
- [52] Kalok Chan, Y. Peter Chung, and Wai-Ming Fong. The Informational Role of Stock and Option Volume. *The Review of Financial Studies*, 15(4):1049–1075, 06 2015.

- [53] Rong Chen and Ruey S. Tsay. Functional-coefficient autoregressive models. *Journal of the American Statistical Association*, 88(421):298–308, 1993.
- [54] Wei Chen, Manrui Jiang, Wei-Guo Zhang, and Zhensong Chen. A novel graph convolutional feature based convolutional neural network for stock trend prediction. *Information Sciences*, 556:67–94, 2021.
- [55] Wei-Jie Chen, Jing-Jing Yao, and Yuan-Hai Shao. Volatility forecasting using deep neural network with time-series feature embedding. *Economic Research-Ekonomska Istraživanja*, 36(1):1377–1401, 2023.
- [56] Yu-Lun Chen and Wei-Che Tsai. Determinants of price discovery in the vix futures market. *Journal of Empirical Finance*, 43:59–73, 2017.
- [57] Eliodoro Chiavazzo, C. Gear, Carmeline Dsilva, Neta Rabin, and Ioannis Kevrekidis. Reduced models in chemical kinetics via nonlinear data-mining. *Processes*, 2, 07 2013.
- [58] Chicago Board Options Exchange. Vix index: Cboe volatility index methodology. https://cdn.cboe.com/api/global/us_indices/governance/VIX_Methodology.pdf, 2024. Accessed: 2024-06-08.
- [59] Rémy Chicheportiche and Jean-Philippe Bouchaud. The fine-structure of volatility feedback i: Multi-scale self-reflexivity. *Physica A: Statistical Mechanics and its Applications*, 410:174–195, 2014.
- [60] Ray Yeutien Chou, Hengchih Chou, and Nathan Liu. *Range Volatility Models and Their Applications in Finance*, pages 1273–1281. Springer US, Boston, MA, 2010.
- [61] Kim Christensen, Mathias Siggaard, and Bezirgen Veliyev. A Machine Learning Approach to Volatility Forecasting*. *Journal of Financial Econometrics*, 06 2022. nbac020.
- [62] Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS 2014 Workshop on Deep Learning, December 2014*, 2014.
- [63] Peter K. Clark. A subordinated stochastic process model with finite variance for speculative prices. *Econometrica*, 41(1):135–155, 1973.

- [64] Aaron Clauset, Cosma Rohilla Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703, 2009.
- [65] R. R. Coifman and Lafon S. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5 – 30, 2006. Special Issue: Diffusion Maps and Wavelets.
- [66] R. Cont, A. Kukanov, and S. Stoikov. The price impact of order book events. *Journal of Financial Econometrics*, 12(1):47–88, June 2013.
- [67] Rama Cont, Mihai Cucuringu, and Chao Zhang. Cross-impact of order flow imbalance in equity markets. *Quantitative Finance*, 23(10):1373–1393, 2023.
- [68] Rama Cont and José da Fonseca. Dynamics of implied volatility surfaces. *Quantitative Finance*, 2(1):45–60, 2002.
- [69] Thomas E. Copeland. A model of asset trading under the assumption of sequential information arrival. *The Journal of Finance*, 31(4):1149–1168, 1976.
- [70] Savelie Cornegruta, Robert Bakewell, Samuel Withey, and Giovanni Montana. Modelling radiological language with bidirectional long short-term memory networks. In *Proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis*, pages 17–27, Austin, TX, November 2016. Association for Computational Linguistics.
- [71] Fulvio Corsi. A simple approximate long-memory model of realized volatility. *The Journal of Financial Econometrics*, 7(2):174–196, 2009.
- [72] Yiran Cui, Sebastian del Baño Rollin, and Guido Germano. Full and fast calibration of the heston stochastic volatility model, 2015.
- [73] Chester Curme, Michele Tumminello, Rosario N. Mantegna, H. Eugene Stanley, and Dror Y. Kenett. Emergence of statistically validated financial intraday lead-lag relationships, 2014.
- [74] Zhifeng Dai, Huiting Zhou, Fenghua Wen, and Shaoyi He. Efficient predictability of stock return volatility: The role of stock market implied volatility. *The North American Journal of Economics and Finance*, 52:101174, 2020.
- [75] Domingos S. de O. Santos J nior, Jo o F.L. de Oliveira, and Paulo S.G. de Matos Neto. An intelligent hybridization of arima with machine learning models for time series forecasting. *Knowledge-Based Systems*, 175:72–86, 2019.

- [76] Francis Diebold and Roberto Mariano. Comparing predictive accuracy. *Journal of Business Economic Statistics*, 13(3):253–63, 1995.
- [77] Kydros Dimitrios and Oumbailis Vasileios. A network analysis of the greek stock market. *Procedia Economics and Finance*, 33:340–349, 2015. The Economies of Balkan and Eastern Europe Countries in the Changed World (EBEEC 2015).
- [78] Youcef Djenouri, Asma Belhadi, Gautam Srivastava, and Jerry Chun-Wei Lin. Hybrid graph convolution neural network and branch-and-bound optimization for traffic flow forecasting. *Future Generation Computer Systems*, 139:100–108, 2023.
- [79] Doris Dobi. Modeling Systemic Risk in The Options Market. *PhD Thesis, New York University*, 2014.
- [80] D. L. Donoho and C. Grimes. Hessian Eigenmaps: Locally Linear Embedding Techniques for High-Dimensional Data. *Proceedings of the National Academy of Sciences*, 100(10):5591–5596, 2003.
- [81] Bernard Dumas, Jeff Fleming, and Robert Whaley. Implied volatility functions: Empirical tests. *The Journal of Finance*, 53, 07 2001.
- [82] Jan Dvorsky, Jaroslav Belas, Beata Gavurova, and Tomas Brabenec. Business risk management in the context of small and medium-sized enterprises. *Economic Research-Ekonomska Istraživanja*, 34(1):1690–1708, 2021.
- [83] David Easley, Maureen O’Hara, and P.S. Srinivas. Option volume and stock prices: Evidence on where informed traders trade. *The Journal of Finance*, 53(2):431–465, 1998.
- [84] Janina Engel, Michela Nardo, and Michela Rancan. *Network Analysis for Economics and Finance: An Application to Firm Ownership*, pages 331–355. Springer International Publishing, Cham, 2021.
- [85] Robert F. Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica*, 50(4):987–1007, 1982.
- [86] David Enke and Suraphan Thawornwong. The use of data mining and neural networks for forecasting stock market returns. *Expert Syst. Appl.*, 29:927–940, 2005.

- [87] Jr. Everette S. Gardner and Ed. Mckenzie. Forecasting trends in time series. *Management Science*, 31(10):1237–1246, 1985.
- [88] Eugene F. Fama and Kenneth R. French. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33:3–56, 1993.
- [89] Eugene F. Fama and Kenneth R. French. A five-factor asset pricing model. *Journal of Financial Economics*, 116(1):1 – 22, 2015.
- [90] Yizhou Fang, Martin Lysy, and Don Mcleish. Common-factor stochastic volatility modelling with observable proxy. *Canadian Journal of Statistics*, 48(1):36–61, 2020.
- [91] Stephen Figlewski. Estimating the Implied Risk-Neutral Density for the US Market Portfolio. In *Volatility and Time Series Econometrics: Essays in Honor of Robert Engle*. Oxford University Press, 03 2010.
- [92] Paulo Renato A. Firmino, Paulo S.G. de Mattos Neto, and Tiago A.E. Ferreira. Correcting and combining time series forecasters. *Neural Networks*, 50:1–11, 2014.
- [93] Black Fischer. Studies in stock price volatility changes. *Proceedings of the American Statistical Association, Business and Economic Statistics*, 64:171–181, 1976.
- [94] Evelyn Fix and J. L. Hodges. Discriminatory analysis. nonparametric discrimination: Consistency properties. *International Statistical Review / Revue Internationale de Statistique*, 57(3):238–247, 1989.
- [95] Valentin Flunkert, David Salinas, and Jan Gasthaus. Deepar: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36, 04 2017.
- [96] Masaaki Fukasawa. Asymptotic analysis for stochastic volatility: martingale expansion. *Finance and Stochastics*, 15:635–654, 2011.
- [97] Mark B. Garman and Michael J. Klass. On the estimation of security price volatilities from historical data. *The Journal of Business*, 53(1):67–78, 1980.
- [98] Li Ge, Tse-Chun Lin, and Neil D. Pearson. Why does the option to stock volume ratio predict stock returns? *Journal of Financial Economics*, 120(3):601–622, 2016.

- [99] Eric Ghysels, Andrew Harvey, and Eric Renault. Stochastic volatility. Cahiers de recherche, Centre interuniversitaire de recherche en Économie quantitative, CIREQ, 1996.
- [100] M. D. Godfrey. Elements of Numerical Analysis, by Peter Henrici (John Wiley and Sons, New York, 1964), pp. 328. *Naval Research Logistics Quarterly*, 12(2):195–195, June 1965.
- [101] Luis Goncalves-Pinto, B. Grundy, A. Hameed, T. V. D. Heijden, and Yichao Zhu. Why do option prices predict stock returns? the role of price pressure in the stock market. *Manag. Sci.*, 66:3903–3926, 2020.
- [102] Marco Grassia, Manlio De Domenico, and Giuseppe Mangioni. mgnn: Generalizing the graph neural networks to the multilayer case, 2021.
- [103] Sanford J. Grossman. The Existence of Futures Markets, Noisy Rational Expectations and Informational Externalities. *The Review of Economic Studies*, 44(3):431–449, 10 1977.
- [104] Julien Guyon and Jordan Lekeufack. Volatility is (mostly) path-dependent. *Quantitative Finance*, 23(9):1221–1258, 2023.
- [105] William L. Hamilton. Graph representation learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 14(3):1–159, 2020.
- [106] Yufeng Han and Fang Liu. The information content of the implied volatility surface: Can option prices predict jumps? *SSRN Electronic Journal*, 01 2019.
- [107] Peter R. Hansen, Asger Lunde, and James M. Nason. The model confidence set. *Econometrica*, 79(2):453–497, 2011.
- [108] M.R. Hassan and B. Nath. Stock market forecasting using hidden markov model: a new approach. In *5th International Conference on Intelligent Systems Design and Applications (ISDA’05)*, pages 192–196, 2005.
- [109] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.
- [110] Bernard Herskovic, Bryan Kelly, Hanno Lustig, and Stijn Van Nieuwerburgh. The common factor in idiosyncratic volatility: Quantitative asset pricing implications. *Journal of Financial Economics*, 119(2):249–283, 2016.

- [111] Steven L. Heston. A Closed-Form Solution for Options with Stochastic Volatility with Applications to Bond and Currency Options. *The Review of Financial Studies*, 6(2):327–343, 04 2015.
- [112] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, nov 1997.
- [113] Mamoru Hosaka. *Bézier Curves and Control Points*, pages 117–139. Springer Berlin Heidelberg, Berlin, Heidelberg, 1992.
- [114] Mohammad Asiful Hossain, Rezaul Karim, Ruppa K. Thulasiram, Neil D. B. Bruce, and Yang Wang. Hybrid deep learning model for stock price prediction. *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1837–1844, 2018.
- [115] Jianfeng Hu. Does option trading convey stock price information? *Journal of Financial Economics*, 111(3):625–645, 2014.
- [116] Dashan Huang, Fuwei Jiang, Kunpeng Li, Guoshi Tong, and Guofu Zhou. Scaled pca: A new approach to dimension reduction. *Management Science*, 68(3):1678–1695, 2022.
- [117] Bujar Huskaj and Marcus Nossman. A term structure model for vix futures. *Journal of Futures Markets*, 33(5):421–442, 2013.
- [118] Takashi Isogai. Dynamic correlation network analysis of financial asset returns with network clustering. *Appl. Netw. Sci.*, 2:8, 2017.
- [119] A Itkin. Deep learning calibration of option pricing models: some pitfalls and solutions, 2019.
- [120] J. Zhang J. Wang and H. Wang M. Xu. Classification of imbalanced data by using the smote algorithm and locally linear embedding. *Signal Processing, 2006 8th International Conference on. Vol. 3. IEEE*, 2007.
- [121] Teo Jašić and Douglas Wood. The profitability of daily stock market indices trades based on neural network predictions: Case study for the s and p 500, the dax, the topix and the ftse in the period 1965-1999. *Applied Financial Economics*, 14:285–297, 02 2004.

- [122] Narasimhan Jegadeesh. Evidence of predictable behavior of security returns. *The Journal of Finance*, 45(3):881–898, 1990.
- [123] Ming Jin, Huan Yee Koh, Qingsong Wen, Daniele Zambon, Cesare Alippi, Geoffrey I. Webb, Irwin King, and Shirui Pan. A survey on graph neural networks for time series: Forecasting, classification, imputation, and anomaly detection, 2023.
- [124] Travis L. Johnson and Eric C. So. The option to stock volume ratio and future returns. *Journal of Financial Economics*, 106(2):262–286, 2012.
- [125] Philippe Jorion. Predicting volatility in the foreign exchange market. *The Journal of Finance*, 50(2):507–528, 1995.
- [126] Kissan Joseph, M. Babajide Wintoki, and Zelin Zhang. Forecasting abnormal stock returns and trading volume using investor sentiment: Evidence from on-line search. *International Journal of Forecasting*, 27(4):1116–1127, 2011.
- [127] Zura Kakushadze and Juan Andrés Serur. *151 Trading Strategies*. Springer, 2018.
- [128] Dimos Kambouroudis and David Mcmillan. Does vix or volume improve garch volatility forecasts? *Applied Economics*, 48:1–19, 10 2015.
- [129] J. Kennedy and R. Eberhart. Particle swarm optimization. In *Proceedings of ICNN’95 - International Conference on Neural Networks*, volume 4, pages 1942–1948 vol.4, 1995.
- [130] Jong-Min Kim, Dong H. Kim, and Hojin Jung. Applications of machine learning for corporate bond yield spread forecasting. *The North American Journal of Economics and Finance*, 58:101540, 2021.
- [131] Kyoung-jae Kim and Ingoo Han. Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index. *Expert Systems with Applications*, 19:125–132, 08 2000.
- [132] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 12 2014.
- [133] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

- [134] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks, 2017.
- [135] Burak Korkusuz, Dimos Kambouroudis, and David G. McMillan. Do extreme range estimators improve realized volatility forecasts? evidence from g7 stock markets. *Finance Research Letters*, 55:103992, 2023.
- [136] Mark A. Kramer. Nonlinear principal component analysis using autoassociative neural networks. *AIChE Journal*, 37(2):233–243, 1991.
- [137] Alexander Kurov and Dennis J Lasser. Price dynamics in the regular and e-mini futures markets. *Journal of Financial and Quantitative Analysis*, 39(2):365–384, 2004.
- [138] Christopher G. Lamoureux and William D. Lastrapes. Forecasting Stock-Return Variance: Toward an Understanding of Stochastic Implied Volatilities. *The Review of Financial Studies*, 6(2):293–326, 04 2015.
- [139] Oliver Ledoit and Michael Wolf. Robust performance hypothesis testing with the sharpe ratio. *Journal of Empirical Finance*, 15(5):850–859, 2008.
- [140] Hei-Wai Lee, Vivek Sharma, and Kelly Nianyun Cai. Are stocks dumped or neglected by analysts’ inferior investments to covered stocks? *Journal of Business Research*, 64(5):501–507, 2011.
- [141] Ming-Chi Lee. Using support vector machine with a hybrid feature selection method to the stock trend prediction. *Expert Systems with Applications*, 36(8):10896–10904, 2009.
- [142] Igor Michel Santos Leite, João Daniel Madureira Yamim, and Leonardo Goliatt da Fonseca. The deepoanets for finance: An approach to calibrate the heston model. In Goreti Marreiros, Francisco S. Melo, Nuno Lau, Henrique Lopes Cardoso, and Luís Paulo Reis, editors, *Progress in Artificial Intelligence*, pages 351–362, Cham, 2021. Springer International Publishing.
- [143] Shangzhe Li, Junran Wu, Xin Jiang, and Ke Xu. Chart gen: Learning chart information with a graph convolutional network for stock movement prediction. *Knowledge-Based Systems*, 248:108842, 2022.

- [144] Wei Li, Ruihan Bao, Keiko Harimoto, Deli Chen, Jingjing Xu, and Qi Su. Modeling the stock relation with graph network for overnight stock movement prediction. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 4541–4547. International Joint Conferences on Artificial Intelligence Organization, 7 2020. Special Track on AI in FinTech.
- [145] Yong Li, Zhaoxuan Li, Qiang Mei, Peng Wang, Wenlong Hu, Zhishan Wang, Wenxin Xie, Yang Yang, and Yuhaoran Chen. Research on multi-port ship traffic prediction method based on spatiotemporal graph neural networks. *Journal of Marine Science and Engineering*, 11(7), 2023.
- [146] Daniel Libman, Simi Haber, and Mary Schaps. Volume prediction with neural networks. *Frontiers in Artificial Intelligence*, 2, 2019.
- [147] Bryan Lim, Sercan O. Arık, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4):1748–1764, 2021.
- [148] Tse-Chun Lin and Xiaolong Lu. Why do options prices predict stock returns? evidence from analyst tipping. *Journal of Banking & Finance*, 52, 03 2015.
- [149] Tse-Chun Lin, Xiaolong Lu, and Joost Driessen. Why do options prices predict stock returns? *Netspar Discussion Paper*, No. 07/2013-079, 2013.
- [150] Min Liu, Wei-Chong Choo, Chi-Chuan Lee, and Chien-Chiang Lee. Trading volume and realized volatility forecasting: Evidence from the china stock market. *Journal of Forecasting*, 42(1):76–100, 2023.
- [151] Xiaotao Liu and Kin keung Lai. Intraday volume percentages forecasting using a dynamic svm-based approach. *Journal of Systems Science and Complexity*, 30:421–433, 2017.
- [152] Yang Liu. Novel volatility forecasting using deep learning–long short term memory recurrent neural networks. *Expert Systems with Applications*, 132:99–109, 2019.
- [153] Guillermo Llorente, Roni Michaely, Gideon Saar, and Jiang Wang. Dynamic volume-return relation of individual stocks. *The Review of Financial Studies*, 15(4):1005–1047, 2002.

- [154] G.G. Lorentz. *Bernstein Polynomials*. AMS Chelsea Publishing Series. Chelsea Publishing Company, 1986.
- [155] Yutong Lu, Gesine Reinert, and Mihai Cucuringu. Co-trading networks for modeling dynamic interdependency structures and estimating high-dimensional covariances in us equity markets, 2023.
- [156] Zhongjin Lu and Yingzi Zhu. Volatility components: The term structure dynamics of vix futures. *Journal of Futures Markets*, 30(3):230–256, 2010.
- [157] Chuong Luong and Nikolai Dokuchaev. Forecasting of realised volatility with the random forests algorithm. *Journal of Risk and Financial Management*, 11(4), 2018.
- [158] Ritva Luukkonen, Pentti Saikkonen, and Timo Teräsvirta. Testing linearity against smooth transition autoregressive models. *Biometrika*, 75(3):491–499, 09 1988.
- [159] Shaojun Ma and Pengcheng Li. Predicting daily trading volume via various hidden states, 2021.
- [160] Aktham Maghyereh, Hussein Abdoh, and Basel Awartani. Have returns and volatilities for financial assets responded to implied volatility during the covid-19 pandemic? *Journal of Commodity Markets*, 26:100194, 2022.
- [161] Steven Manaster and Richard J. Rendleman. Option prices as predictors of equilibrium stock prices. *The Journal of Finance*, 37(4):1043–1057, 1982.
- [162] Benoit Mandelbrot. The Variation of Certain Speculative Prices. *The Journal of Business*, 36:394–394, 1963.
- [163] Gary Marcus. Deep learning: A critical appraisal, 2018.
- [164] Daiki Matsunaga, Toyotaro Suzumura, and Toshihiro Takahashi. Exploring graph neural networks for stock market predictions with rolling window analysis. *arXiv preprint arXiv:1909.10660*, 2019.
- [165] Sheila Mbadi. Predicting stock market movement using an enhanced naïve bayes model for sentiment analysis classification. 02 2018.
- [166] A. Melman. Numerical solution of a secular equation. *Numerische Mathematik*, 1995.

- [167] Felix Prenzel Milena Vuletić and Mihai Cucuringu. Fin-gan: forecasting and classifying financial time series via generative adversarial networks. *Quantitative Finance*, 24(2):175–199, 2024.
- [168] Camelia Minoiu and Javier Reyes. A network analysis of global banking: 1978–2010. *Journal of Financial Stability*, 9(2):168–184, 2013.
- [169] Stefan Mittnik, Nikolay Robinzonov, and Martin Spindler. Stock market volatility: Identifying major drivers and the nature of their impact. *Journal of Banking & Finance*, 58:1–14, 2015.
- [170] Roy Mitz, Nir Sharon, and Yoel Shkolnisky. Symmetric rank-one updates from partial spectrum with an application to out-of-sample extension. *SIAM J. Matrix Anal. Appl.*, 40:973–997, 2017.
- [171] Roy Mitz and Yoel Shkolnisky. ROIPCA: an online PCA algorithm based on rank-one updates. *CoRR*, abs/1911.11049, 2019.
- [172] P. A. P. Moran and Peter Whittle. Hypothesis testing in time series analysis. 1951.
- [173] Jorge J. Moré. The levenberg-marquardt algorithm: Implementation and theory. In G. A. Watson, editor, *Numerical Analysis*, pages 105–116, Berlin, Heidelberg, 1978. Springer Berlin Heidelberg.
- [174] Fernando Moreno-Pino and Stefan Zohren. Deepvol: Volatility forecasting from high-frequency data with dilated causal convolutions, 2022.
- [175] Mbodja Mougou and Ann Marie Whyte. Stock returns and volatility: An empirical investigation of the german and french equity markets. *Global Finance Journal*, 7(2):253–263, 1996.
- [176] Thiago J. M. Moura, George D. C. Cavalcanti, and Luiz S. Oliveira. Evaluating competence measures for dynamic regressor selection, 2019.
- [177] Milan Mrazek, Jan Pospíšil, and Tomáš Sobotka. On calibration of stochastic and fractional stochastic volatility models. *European Journal of Operational Research*, 254, 04 2016.
- [178] Milan Mrázek and Jan Pospíšil. Calibration and simulation of heston model. *Open Mathematics*, 15(1):679–704, 2017.

- [179] E. A. Nadaraya. On estimating regression. *Theory of Probability & Its Applications*, 9(1):141–142, 1964.
- [180] Antonio Naimoli, Richard Gerlach, and Giuseppe Storti. Improving the accuracy of tail risk forecasting models by combining several realized volatility estimators. *Economic Modelling*, 107:105701, 2022.
- [181] Nicoletta Nicolaou and Timothy G. Constandinou. A nonlinear causality estimator based on non-parametric multiplicative regression. *Frontiers in Neuroinformatics*, 10(JUNE), June 2016.
- [182] Michael O’Neill and Gulasekaran Rajaguru. Causality of price movements in vix exchange-traded products and vix futures contracts. *Journal of Accounting Literature*, pages 1–17, April 2023.
- [183] Option Metrics. Ivydb, 12/2021.
- [184] Joerg Osterrieder, Daniel Kucharczyk, Silas Rudolf, and Daniel Wittwer. Neural networks and arbitrage in the VIX. *Digital Finance*, 2(1):97–115, September 2020.
- [185] Ping-Feng Pai and Chih-Sheng Lin. A hybrid arima and support vector machines model in stock price forecasting. *Omega*, 33(6):497–505, 2005.
- [186] Mahesh Pal. Modelling pile capacity using generalised regression neural network. In *Proceedings of Indian geotechnical conference December*, pages 15–17, 2011.
- [187] Jun Pan and Allen M. Poteshman. The information in option volume for future stock prices. *Review of Financial Studies*, 19(3):871–908, 2006.
- [188] Sibarama Panigrahi and H.S. Behera. A hybrid ets ann model for time series forecasting. *Engineering Applications of Artificial Intelligence*, 66:49–59, 2017.
- [189] Panagiotis G. Papaioannou, Ronen Talmon, Ioannis G. Kevrekidis, and Constantinos Siettos. Time-series forecasting using manifold learning, radial basis function interpolation, and geometric harmonics. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 32(8):083113, 2022.
- [190] Michael H. Parkinson. The extreme value method for estimating the variance of the rate of return. *The Journal of Business*, 53:61–65, 1980.

- [191] Andrew J. Patton. Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics*, 160(1):246–256, 2011. Realized Volatility.
- [192] Andrew J. Patton and Kevin Sheppard. Optimal combinations of realised volatility estimators. *International Journal of Forecasting*, 25(2):218–238, 2009. Forecasting Returns and Risk in Financial Markets using Linear and Nonlinear Models.
- [193] Luca Di Persio and Oleksandr Honchar. Recurrent neural networks approach to the financial forecast of google assets. 2017.
- [194] Les Piegl and Wayne Tiller. *B-spline Curves and Surfaces*, pages 81–116. Springer Berlin Heidelberg, Berlin, Heidelberg, 1997.
- [195] Francisco J. Pulgar, Francisco Charte, Antonio J. Rivera, and Mar a J. del Jesus. Aeknn: An autoencoder knn-based classifier with built-in dimensionality reduction. *International Journal of Computational Intelligence Systems*, 12:436–452, 2019.
- [196] Jorge V. Pérez-Rodríguez. Another look at the implied and realised volatility relation: a copula-based approach. *Risk Management*, 22(1):38–64, March 2020.
- [197] Rafael Reisenhofer, Xandro Bayer, and Nikolaus Hautsch. Harnet: A convolutional neural network for realized volatility forecasting, 2022.
- [198] L. Rogers and Stephen Satchell. Estimating variance from high, low and closing prices. *The Annals of Applied Probability*, 1, 11 1991.
- [199] Richard Roll, Eduardo Schwartz, and Avanidhar Subrahmanyam. O/s: The relative trading activity in options and stock. *Journal of Financial Economics*, 96(1):1–17, 2010.
- [200] Yu Rong, Wenbing Huang, Tingyang Xu, and Junzhou Huang. Dropedge: Towards deep graph convolutional networks on node classification, 2020.
- [201] Murtaza Roondiwala, Harshal Patel, and Shraddha Varma. Predicting stock prices using lstm. *International Journal of Science and Research (IJSR)*, 6, 04 2017.
- [202] Robert W. Rudd. Theil, henri, applied economic forecasting, chicago, rand mc-nally & company, 1966. *American Journal of Agricultural Economics*, 49:241–243, 1967.

- [203] Venkatesh Satish, Abhay Saxena, and Max Palmer. Predicting intraday trading volume and volume percentages. *The Journal of Trading*, 9:15–25, 06 2014.
- [204] Arjun Singh Saud and Subarna Shakya. Analysis of look back period for stock price prediction with rnn variants: A case study on banking sector of nepse. *Procedia Computer Science*, 167:788–798, 2020. International Conference on Computational Intelligence and Data Science.
- [205] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80, 2009.
- [206] Christian Schlag and Hans Stoll. Price impacts of options volume. *Journal of Financial Markets*, 8(1):69–87, 2005.
- [207] Bernhard Schölkopf, Robert Williamson, Alex Smola, John Shawe-Taylor, and John Platt. Support vector method for novelty detection. In *Proceedings of the 12th International Conference on Neural Information Processing Systems*, NIPS’99, page 582–588, Cambridge, MA, USA, 1999. MIT Press.
- [208] Wim Schoutens, Erwin Simons, and Jurgen Tistaert. A perfect calibration! now what? *Wilmott*, 2004:66–78, 2004.
- [209] Rob Schumaker and Hsiu-chin Chen. Textual analysis of stock market prediction using breaking financial news: The azfin text system. *ACM Trans. Inf. Syst.*, 27, 02 2009.
- [210] H.Nejat Seyhun. Insiders’ profits, costs of trading, and market efficiency. *Journal of Financial Economics*, 16(2):189–212, 1986.
- [211] Omer Berat Sezer, Mehmet Ugur Gudelek, and Ahmet Murat Ozbayoglu. Financial time series forecasting with deep learning : A systematic literature review: 2005–2019. *Applied Soft Computing*, 90:106181, 2020.
- [212] William F. Sharpe. Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance*, 19(3):425–442, 1964.
- [213] William F. Sharpe. The sharpe ratio. *The Journal of Portfolio Management*, 21(1):49–58, 1994.

- [214] Jingyi Shen and M. Shafiq. Short-term stock market price trend prediction using a comprehensive deep learning system. *Journal of Big Data*, 7, 08 2020.
- [215] Yong Shi, Yunong Wang, Yi Qu, and Zhensong Chen. Integrated gcn-lstm stock prices movement prediction based on knowledge-incorporated graphs construction. *International Journal of Machine Learning and Cybernetics*, 15(1):161–176, 2024.
- [216] William L Silber. *The economic role of financial futures*. Salomon Brothers Center for the Study of Financial Institutes, 1985.
- [217] Eraylson. G. Silva, George D. C. Cavalcanti, João Fausto L. de Oliveira, and Paulo S. G. de Mattos Neto. On the evaluation of dynamic selection parameters for time series forecasting. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7, 2020.
- [218] David P Simon and Jim Campasano. *The VIX futures basis: Evidence and trading strategies*. SSRN, 2014.
- [219] D.F. Specht. A general regression neural network. *IEEE Transactions on Neural Networks*, 2(6):568–576, 1991.
- [220] Jens A. Stephan and Robert E. Whaley. Intraday price change and trading volume relations in the stock and stock option markets. *The Journal of Finance*, 45(1):191–220, 1990.
- [221] Nick Taylor. Forecasting returns in the vix futures market. *International Journal of Forecasting*, 35(4):1193–1210, 2019.
- [222] J. B. Tenenbaum, V. Silva, and J. C. Langford. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 290(5500):2319–2323, 2000.
- [223] Manoj Thakur and Deepak Kumar. A hybrid financial trading support system using multi-category classifiers and random forest. *Applied Soft Computing*, 67:337–349, 2018.
- [224] Erik Theissen. Price discovery in spot and futures markets: A reconsideration. *European Journal of Finance*, 18, 01 2011.
- [225] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

- [226] Michele Tumminello, Salvatore Miccichè, Fabrizio Lillo, Jyrki Piilo, and Rosario N. Mantegna. Statistically validated networks in bipartite complex systems. *PLOS ONE*, 6(3):1–11, 03 2011.
- [227] Marcel Visser. Ranking and combining volatility proxies for garch and stochastic volatility models. MPRA Paper 4917, Korteweg-de Vries Institute for Mathematics, University of Amsterdam, Amsterdam, Netherlands, 11 2008. 31 pages.
- [228] Stefania Vitali and Stefano Battiston. The community structure of the global corporate network. *PLOS ONE*, 9:1–13, 08 2014.
- [229] Jianian Wang, Sheng Zhang, Yanghua Xiao, and Rui Song. A review on graph neural network methods in financial applications, 2021.
- [230] Geoffrey S. Watson. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, 26(4):359–372, 1964.
- [231] Yu Wei, Chao Liang, Yan Li, Xunhui Zhang, and Guiwu Wei. Can cboe gold and silver implied volatility help to forecast gold futures volatility in china? evidence based on har and ridge regression models. *Finance Research Letters*, 35:101287, 2020.
- [232] Bin Weng, Lin Lu, Xing Wang, Fadel M. Megahed, and Waldyn Martinez. Predicting short-term stock prices using ensemble methods and online data sources. *Expert Systems with Applications*, 112:258–273, 2018.
- [233] Robert Whaley. Understanding the vix. *The Journal of Portfolio Management*, 35:98–105, 02 2009.
- [234] Larry Williams. The vix fix. pages 24–32, 12 2007.
- [235] Peter R. Winters. Forecasting sales by exponentially weighted moving averages. *Management Science*, 6(3):324–342, 1960.
- [236] World Bank. World bank: Gross domestic product 2019, 2 2021. Data retrieved from <https://databank.worldbank.org/data>.
- [237] Yulei Wu, Hong-Ning Dai, and Haina Tang. Graph neural networks for anomaly detection in industrial internet of things. *IEEE Internet of Things Journal*, PP:1–1, 07 2021.

- [238] Sophie Xiaoyan, Jun Pan, Allen Potesman, Joe Chen, Jun Liu, Josh Pollet, Dimitri Vayanos, Jiang Wang, and Josh White. The information in option volume for future stock volatility. *Journal of Finance*, 2008.
- [239] Yuhang Xing, Xiaoyan Zhang, and Rui Zhao. What does the individual option volatility smirk tell us about future equity returns? *Journal of Financial and Quantitative Analysis*, 45(3):641–662, 2010.
- [240] Xiaojie Xu and Yun Zhang. A high-frequency trading volume prediction model using neural networks. *Decision Analytics Journal*, 7:100235, 2023.
- [241] Xiang Yan, Jiancheng Bai, Xiafei Li, and Zhonglu Chen. Can dimensional reduction technology make better use of the information of uncertainty indices when predicting volatility of chinese crude oil futures? *Resources Policy*, 75:102521, 2022.
- [242] Dennis Yang and Qiang Zhang. Drift-independent volatility estimation based on high, low, open, and close prices. *The Journal of Business*, 73(3):477–492, 2000.
- [243] Qiwei Yin, Ruixun Zhang, Yiwei Liu, and XiuLi Shao. Forecasting of stock price trend based on cart and similar stock. In *2017 4th International Conference on Systems and Informatics (ICSAI)*, pages 1503–1508, 2017.
- [244] Xingkun Yin, Da Yan, Abdullateef Almudaifer, Sibbo Yan, and Yang Zhou. Forecasting stock prices using stock correlation graph: A graph convolutional network approach. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2021.
- [245] Tao You, Paweł Fiedor, and Artur Hołda. Network analysis of the shanghai stock exchange based on partial mutual information. *Journal of Risk and Financial Management*, 8(2):266–284, 2015.
- [246] Chao Zhang, Xingyue Pu, Mihai Cucuringu, and Xiaowen Dong. Graph neural networks for forecasting multivariate realized volatility with spillover effects, 2023.
- [247] Chao Zhang, Xingyue Pu, Mihai Cucuringu, and Xiaowen Dong. Graph neural networks for forecasting multivariate realized volatility with spillover effects, 2023.

- [248] Chao Zhang, Yihuang Zhang, Mihai Cucuringu, and Zhongmin Qian. Volatility forecasting with machine learning and intraday commonality, 2022.
- [249] Chao Zhang, Zihao Zhang, Mihai Cucuringu, and Stefan Zohren. A universal end-to-end approach to portfolio optimization via deep learning, 2021.
- [250] G.Peter Zhang. Time series forecasting using a hybrid arima and neural network model. *Neurocomputing*, 50:159–175, 2003.
- [251] Jin E. Zhang and Yingzi Zhu. Vix futures. *Journal of Futures Markets*, 26(6):521–531, 2006.
- [252] Muhan Zhang and Yixin Chen. Link prediction based on graph neural networks, 2018.
- [253] Wenjun Zhang, Zhensong Chen, Jianyu Miao, and Xueyong Liu. Research on graph neural network in stock market. *Procedia Computer Science*, 214:786–792, 2022. 9th International Conference on Information Technology and Quantitative Management.
- [254] Zihao Zhang, Stefan Zohren, and Stephen Roberts. Deep learning for portfolio optimization. *The Journal of Financial Data Science*, 2(4):8–20, 2020.
- [255] Gilles Zumbach. Volatility conditional on price trends. *Quantitative Finance*, 10(4):431–442, 2010.
- [256] Kathleen Zyga, Richard Price, and Brenton Williams. A generalized regression neural network for logo recognition. In *KES'2000. Fourth International Conference on Knowledge-Based Intelligent Engineering Systems and Allied Technologies. Proceedings (Cat. No.00TH8516)*, volume 2, pages 475–478 vol.2, 2000.
- [257] Zeynep İltüzzer. Option pricing with neural networks vs. black-scholes under different volatility forecasting approaches for bist 30 index options. *Borsa Istanbul Review*, 22(4):725–742, 2022.

Appendix A

Appendix for Option Volume Imbalance as a Predictor for Equity Returns

A.1 Financial Definitions

The following definitions are from NASDAQ's website <https://www.nasdaq.com/glossary>. The types of market participants used in the analysis are:

- **Firm (Proprietary Trades):** *Principal trading in which a firm seeks direct gain rather than commission dollars.*
- **Broker:** *An individual who is paid a commission for executing customer orders. Either a floor broker who executes orders on the floor of the exchange, or an upstairs broker who handles retail customers and their orders. Also, person who acts as an intermediary between a buyer and seller, usually charging a commission. A "broker" who specializes in stocks, bonds, commodities, or options acts as an agent and must be registered with the exchange where the securities are traded.*
- **Market Makers:** *One who maintains firm bid and offer prices in a given security by standing ready to buy or sell round lots at publicly quoted prices.*
- **Customers:** *A person or entity that is not a broker or dealer in securities (and is not a Professional Customer as defined below).*
- **Professional Customers:** *A person or entity that is not a broker or dealer in securities and places more than 390 orders in listed options on average per day during a calendar month for its own beneficial account(s)*

For each market participant class other than Market Makers, the volumes are further decomposed into four types based on transaction intent: Opening Buy, Opening Sell, Closing Buy and Closing Sell, which are defined as follows

- **Opening Buy:** *An option where the intention of the purchaser is to create or increase a long position.*
- **Opening Sell:** *An option where the intention of the writer is to create or increase a short position.*
- **Closing Buy:** *An option where the intention of the purchaser is to reduce or eliminate a short position in a stock.*
- **Closing Sell:** *An option where the intention of the writer is to reduce or eliminate a short position in a stock.*

A.2 Approximation of trade flow between MPCs

In Section 3 of the main text, we construct an approximation of the trade flow between different market participant classes.

We approximate the trade flow, by inspecting the 10-minute intraday aggregate volume reports, and then interpreting and matching the flow. Within a 10-minute window, if there is exactly one MPC buying and/or selling a particular product, we can easily record the flow between the two classes. However, if several MPCs trade a specific product during the window, we can estimate the flow between the different participants, by interpreting the buy and sell volumes of each MPC as left and right nodes of a bipartite graph. The existence and volume of transactions between the two types of participants is represented by the existence and weight of an edge. Inspired by the Chung-Lu generative model for bipartite graphs [28], we can then approximate the flow between two MPCs, for each asset, at each window, by assuming the flow is proportional to the product of the buyer and seller volume.

Let $V_{i,j,d,m}^{\text{Buy},t}$ (resp. $V_{i,j,d,m}^{\text{Sell},t}$), correspond to the volume of option contract j for underlying asset i bought (resp. sold) during day d , during intraday time window $t \in \{1, \dots, 39\}$, by MPC m . Then, we denote the nominal flow between MPCs m_1 and m_2 on day d as

$$\text{NF}_{d,m_1,m_2} = \sum_{i=1}^{N_d} \sum_{j=1}^{C_{i,d}} \sum_{t=1}^{39} P_{i,j,d} V_{i,j,d,m_1}^{\text{Buy},t} \times \frac{V_{i,j,d,m_2}^{\text{Sell},t}}{\sum_{m'} V_{i,j,d,m'}^{\text{Sell},t}}. \quad (\text{A.1})$$

Next, we normalize the flow for each day so that

$$\sum_{m_1} \sum_{m_2} \text{NF}_{d,m_1,m_2} = 1, \forall d \in \{1, \dots, D\},$$

and average across all D days, to obtain the average nominal flow between MPCs m_1 and m_2

$$\widetilde{\text{NF}}_{m_1,m_2} = \text{median} \left(\frac{\text{NF}_{d,m_1,m_2}}{\sum_{m'_1,m'_2} \text{NF}_{d,m'_1,m'_2}}; d = 1, \dots, D \right). \quad (\text{A.2})$$

A.3 OVI of different Market Participant Classes: Additional Results

In this part, we provide additional numeric results regarding Section 5 of the main paper.

A.3.1 Contemporaneous Returns

The main paper was focused on the usage of Option Volume Imbalance as a predictor for future returns. We are also interested in testing whether the OVI has explanatory power on contemporaneous returns. To this end, we test for the linear relationship

$$\tilde{f}_{i,d-1}^{\text{CL_tmCL}} = \alpha + \beta_{i,d} \text{OVI}_{i,d,m} + \epsilon_{i,d}. \quad (\text{A.3})$$

For comparison purposes, we also consider the relationship of OVI with future returns

$$\tilde{f}_{i,d}^{\text{CL_tmCL}} = \alpha + \beta_{i,d} \text{OVI}_{i,d,m} + \epsilon_{i,d}. \quad (\text{A.4})$$

In a sliding window manner, with a window length of 200 days and a shift size of 50 days, we regress both CL_tmCL (Equation (A.4)) and prevCL_CL (Equation (A.3)) returns against OVI, by filtering non-zero OVI values. In Figure A.1, we show the median p and R^2 value for contemporaneous and future returns by using the above method.

The OVI is a significant explanatory variable for all MPCs, for both contemporaneous and future returns. However, we observe that the p -values for the contemporaneous returns are much lower than those corresponding to the future returns. The R^2 's obtained are generally quite low across all cases, but slightly higher for contemporaneous returns.

	CL_tmCL	CL_tmOP	tmOP_tmCL		prevCL_CL	prevCL_OP	OP_CL
Firm	4.85e-05 (1.34e-01)	1.45e-05 (3.83e-01)	6.57e-05 (9.07e-02)	Firm	2.72e-02 (0e+00)	1.15e-02 (0e+00)	1.72e-02 (0e+00)
Broker	5.32e-05 (1.64e-02)	7e-05 (9.09e-03)	1.64e-05 (2.04e-01)	Broker	1.57e-02 (0e+00)	5.43e-03 (0e+00)	1.05e-02 (0e+00)
Market Maker	1.39e-04 (7.61e-04)	1.65e-04 (8.45e-05)	3.26e-05 (1.17e-01)	Market Maker	5.55e-03 (0e+00)	2.47e-03 (0e+00)	2.7e-03 (0e+00)
Customer	8.94e-05 (4.25e-03)	1.1e-04 (4.52e-04)	1.15e-05 (2.52e-01)	Customer	1.71e-02 (0e+00)	5.91e-03 (0e+00)	1.09e-02 (0e+00)
Professional Customer	3.52e-05 (3.46e-01)	5.28e-05 (2.72e-01)	5.95e-05 (2.25e-01)	Professional Customer	6.97e-04 (7.25e-05)	5.18e-04 (2.08e-04)	2.84e-04 (1.36e-02)

Figure A.1: Adjusted R^2 -values obtained for the coefficient of the OVI in (Equation (A.4)) and (Equation (A.3)) respectively for different MPCs, for the Q3 QR group. The p -values for the regression are shown in brackets.

A.3.2 Alternative Bet-size weighting schemes

In the main paper, we assess performance of the P&Ls of strategies based on OVI-based strategies that allocate equal capital across traded instruments (we refer to this as the Uniform Weighting Scheme). In this section, we use a number of different bet weighting schemes, where the bet size (ie., capital) allocated to each equity (see equation (2.8)) depends either on the liquidity of the asset in the option market for that financial day (as measured by traded volume) or on the magnitude of its signal (OVI). However, we could take into account our confidence in the used signals, by basing the bet scheme either on the magnitude of the signal or the liquidity of the option contract (as for more liquid assets, the OVI is calculated using more transaction data). In each case, we first define a “raw” bet size $b'_{i,d}$ for each equity on the basis of an appropriate measure of the market, and set

$$b_{i,d} = \frac{b'_{i,d}}{\sum_j b'_{j,d}}.$$

In this way we ensure comparability of schemes by betting a total of 1 dollar, distributed across all the underlying assets for which trading is indicated, each day. This way, the cumulative P&L (2.8) has the interpretation of being the cumulative percentage return.

- **Uniform Weighting Scheme.** This is the weighting scheme we used in the main paper. Here $b'_{i,d} = 1, \forall i, d$. In this case we do not differentiate between the different underlying assets, allocating equal capital to each asset i with $s_{i,d} \neq 0$.
- **Imbalance Weighting Scheme.** Here $b'_{i,d} = |s_{i,d}^h|$. Using this betting scheme is equivalent to using the signal $s_{i,d}$ instead of just its sign as in Equation (2.8). As higher-magnitude signals (for example, higher OVI magnitude) indicate a higher confidence level, this weighting scheme allows us to incorporate our prediction confidence in the bet size. Splitting equities into quantile rank portfolios is an alternative way to differentiate based on the magnitude of the signals, but using this scheme allows us to use the whole portfolio, without splitting into groups.
- **(Nominal) Volume Weighting Scheme.** Here the bet size is set to $b'_{i,d} = \log(1 + \sum_{j=1}^{C_{i,d}} V_{i,j,d,m})$, or $b'_{i,t} = \log(1 + \sum_{j=1}^{C_{i,d}} P_{i,j,d} V_{i,j,d,m})$ for the nominal volume case (depending on the MPC m). The (nominal) volume is used as a proxy for option liquidity in order to de-emphasize illiquid cases. The log transformation is applied to soften the significant differences in volumes between different contracts.

- **Relative Volume Weighting Scheme.** Here the bet size is set to $b'_{i,d} = \log(1 + \sum_{j=1}^{C_{i,d}} \frac{V_{i,j,d,m}}{OI_{i,j,d}})$ (depending on the MPC m), where $OI_{i,j,d}$ is the open interest for option contract j (of asset i). One issue with the volume weighting scheme, is that equities with lower volumes in the option market, will tend to get a lower weight, even when their liquidity is relatively high. The relative volume weighting scheme, takes this shortcoming into account, by normalizing volumes by the open interest of each contract.
- **IV-Volume Weighting Scheme.** Given how the implied volatility feature has the most significant and consistent effect on the predictability of the OVI, we consider a betting weighting scheme based on it. Similarly to the volume weighted betting scheme, we consider a strategy based on the product's volume, multiplied by the implied volatility of the asset. The best size is set as

$$b'_{i,d} = \sum_{i=1}^{N_d} \sum_{j=1}^{C_{i,d}} \sum_m \log(1 + \min\{\sigma_{i,j,d}, 2\}) V_{i,j,d,m},$$

where $\sigma_{i,j,d}$ is the implied volatility corresponding to option contract j of asset i on day d .

We compare the Uniform Weighting Scheme with the alternative schemes defined. We use the Market Maker OVI, and record the P&L corresponding to excess market overnight returns. In Figure A.2, we carry out a SR difference test to compare the performance of different weighting schemes for the overnight returns with the uniform scheme. In Figure A.2, we plot the cumulative P&L for excess market overnight returns obtained by following a strategy for each of the different weighting schemes, for the Market Maker OVI.

Imbalance Weighting Scheme: In most cases across MPCs and quantile rank groups, when performing pairwise SR difference tests between the Uniform and Imbalance Weighting Schemes, the hypothesis is not rejected, and if we correct for multiple testing, none of the hypothesis' is rejected. However, as seen in Figure A.3, the SRs corresponding to the Imbalance Scheme are slightly higher than those for the Uniform Weighting Scheme, for all quantile rank groups for Brokers and Customers, as well as Q2, Q4 and Q5 for Market Makers. Therefore, this suggests that the Imbalance Weighting Scheme may be a better way to incorporate the information embedded in the OVI, than a simple uniform scheme.

Furthermore, as seen in the cumulative plot Figure A.2(e), we can clearly see that the quantile rank groups are more indistinguishable (both in terms of SR and PPD)

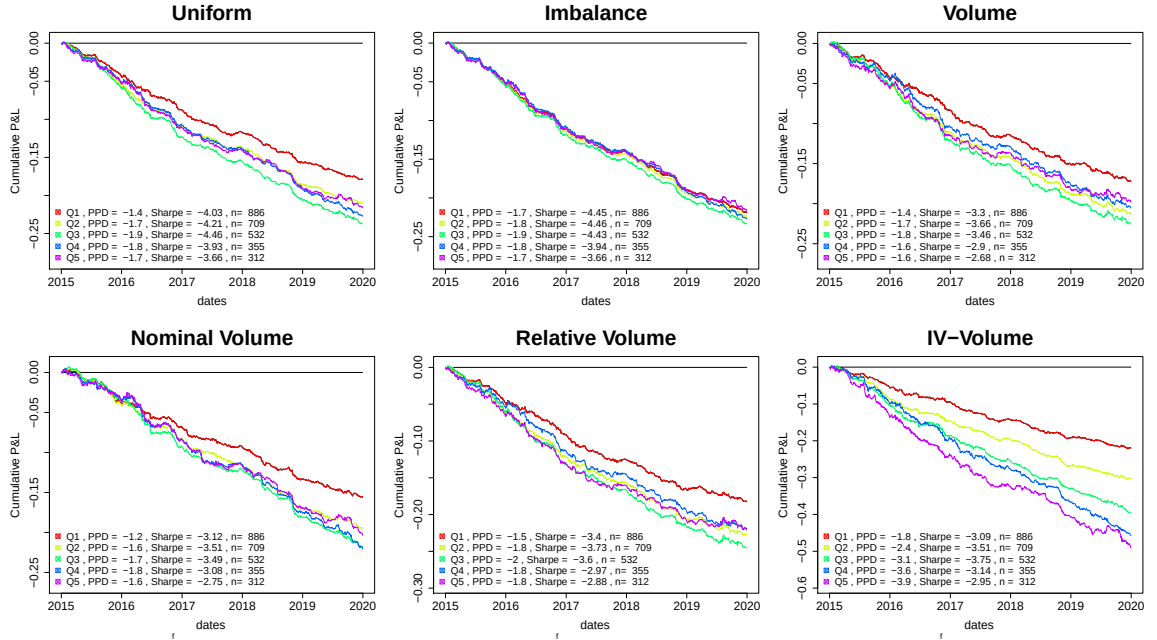


Figure A.2: Cumulative Returns for Different Betting Schemes: This shows the cumulative EMR CL_tmOP returns P&L by following the different betting schemes with one-day holding period. The OVI used is for Market Makers, and the five quantile rank portfolios Q1-Q5. Each portfolio Q-k corresponds to considering the assets corresponding to the $(20k)$ % highest OVIs in absolute value, for six different target returns. Also given, are Sharpe (the Annualized Sharpe Ratio), PPD (Profit per dollar in basis points), and n (the average number of assets used on the portfolio).

in comparison with the uniform scheme, especially for Q1-Q3. In addition, the 5th QR group is the worst performing, despite being one of the highest performing groups for the other weighting schemes.

Volume-Weighting Scheme: In Figure A.3, it can be seen that in most cases, both the Volume and Nominal Volume Weighting Schemes are outperformed by the Uniform Weighting Scheme both in terms of SR as well as PPD. The Relative Volume Scheme, is also outperformed by the Uniform Scheme in terms of SR, but slightly outclasses it in terms of PPD. In addition it outclasses the other two schemes in terms of both SR and PPD, for most of the QR Groups. The lower SRs are a result of higher variability introduced when incorporating volume into the betting scheme. However, none of the volume betting schemes seems to be suitable in replacing the uniform betting scheme. Note, that defining a weighting scheme based on the raw option volume (without the log transform), the square-root-transformed volume, or the liquidity of the underlying, also fails to outperform the uniform scheme, even

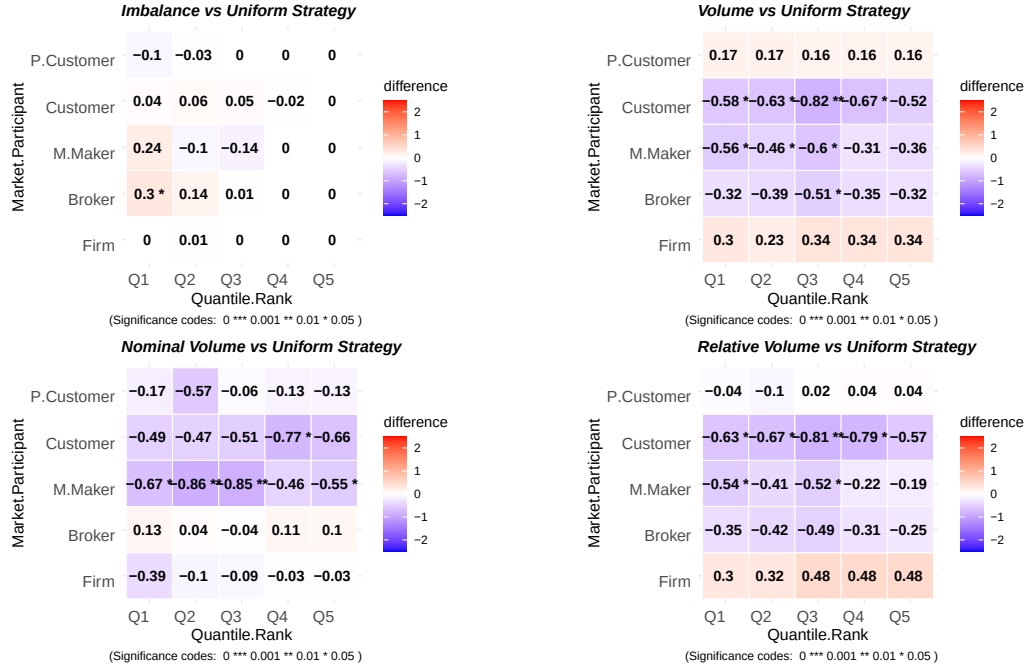


Figure A.3: Sharpe Ratio Difference Tests for different betting scheme: Cells colored red (resp. blue) denote that the scheme considered outperforms (resp. is outperformed) by the Uniform Weighting Scheme. The significance level is also given after the number of stars (1-3), if applicable.

though the raw-Volume Weighting Scheme does outperform it in terms of PPD (but has a very high variability).

As can be seen from Figure A.2, this scheme results in a higher PPD than all the other schemes, even though its SR is lower than most schemes with the exception of the volume scheme. Therefore, similarly to the Relative Volume Weighting Scheme, this is a more profitable yet less consistent scheme than the Uniform Scheme.

A.3.3 Cumulative P&L for non-market maker participants

In Section 5 of main the paper, we showed the cumulative P&L plots for the various return modes for Market Makers. Below, in Figures Figure A.4-Figure A.7, we show the equivalent plots for Firms, Brokers, Customers and Professional Customers.

Similarly to Market Makers, the highest performance observed for Brokers and Customers corresponds to overnight EMR CL_tmOP returns. OVI also seems to have predictive power on tmOP_tmCL returns, especially for the years 2015-2018. However, in both cases, performance drops in 2019. The CL_tmCL return data is essentially a combination of the other two types of returns, thereby resulting in a cumulative P&L plot with a clear trend, and a drop in performance for 2019. For

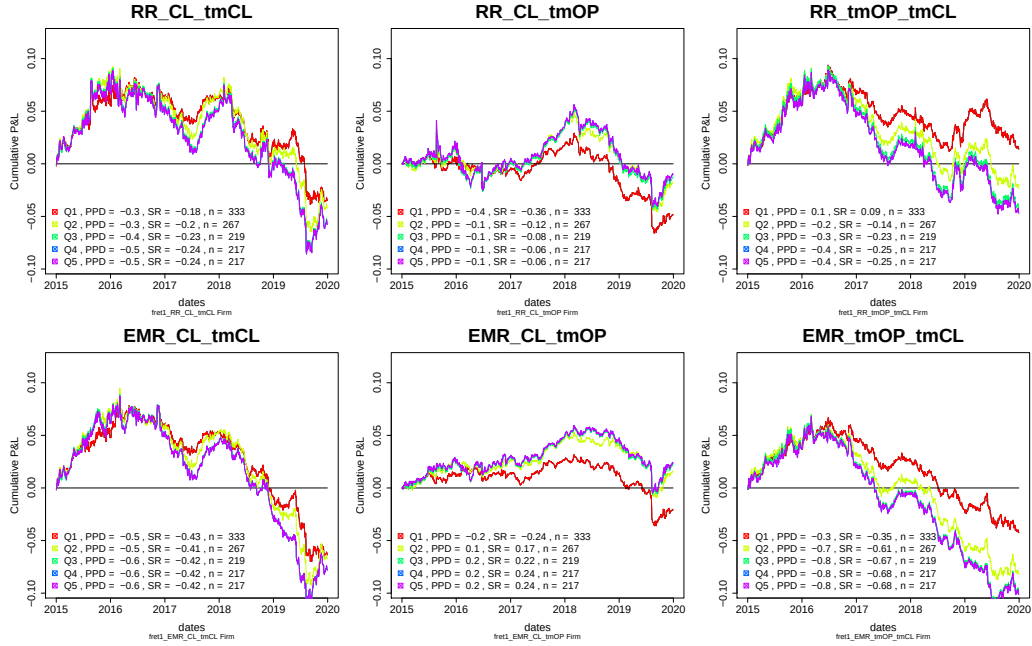


Figure A.4: Cumulative P&L plot - Firm: This shows the cumulative P&L from following a Uniform Weighting Scheme - one-day holding period strategy, for the five quantile rank groups Q1-Q5, using the OVI of Market Makers for six different types of returns. Also given, are SR (Annualized Sharpe Ratio), PPD (Profit per dollar in basis points), and n (average number of instruments in the portfolios).

Professional Customers, most of the OVI predictability corresponds to overnight returns, which follow a clear upward trend. For EMR tmOP_tmCL, the cumulative P&L seems to have a somewhat downward trend for 2015 and 2016 and a somewhat upward trend for 2018 and 2019. For Firms, the EMR tmOP_tmCL returns correspond to a upward trend for 2015, which later reverses. On the other hand, Firm OVI seems to be uninformative for overnight returns, as the cumulative P&L remains close to zero.

A.3.4 Proportion of “profitable” days

An alternative way to evaluate the performance of the different OVI strategies, would be to consider the percentage of profitable days. Denote the proportion of positive daily P&L's by π corresponding to a particular OVI. In Figure A.8, we show the value of $\max\{\pi, 1 - \pi\}$ for different OVI signals, for a Uniform Weighting Scheme, evaluated for EMR CL_tmOP returns. This is the proportion of P&L's that are either positive (e.g by using the Customer OVI) or negative (e.g by using the Market Maker OVI). The OVIs of Firms and Professional Customers have values close to 0.5,

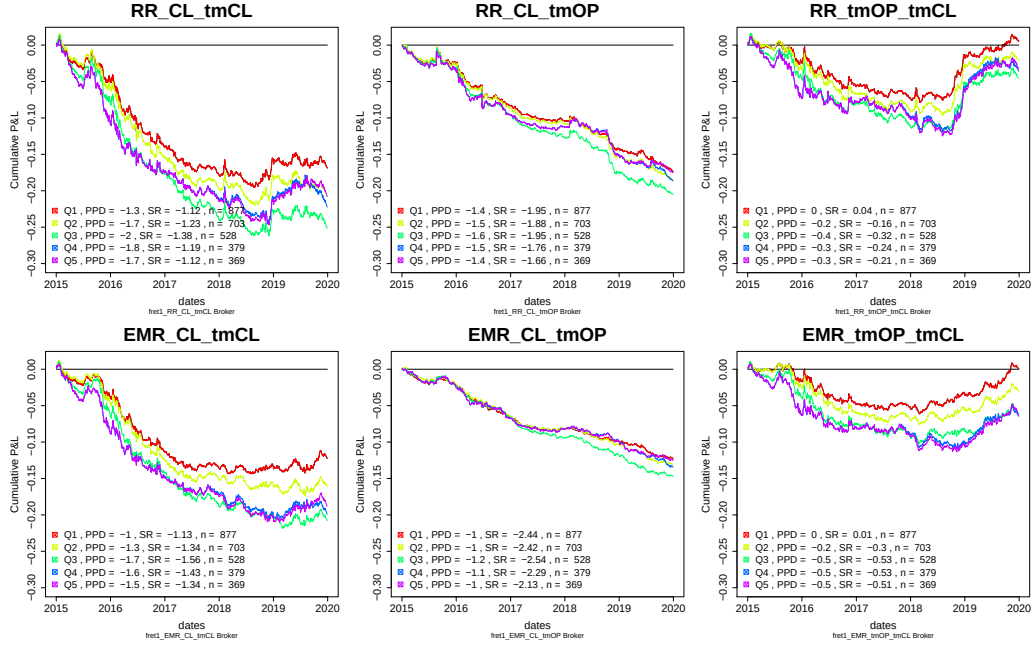


Figure A.5: Cumulative P&L plot - Broker: This shows the cumulative P&L from following a Uniform Weighting Scheme - one-day holding period strategy, for the five quantile rank groups Q1-Q5, using the OVI of Market Makers for six different types of returns. Also given, are SR (Annualized Sharpe Ratio), PPD (Profit per dollar in basis points), and n (average number of instruments in the portfolios).

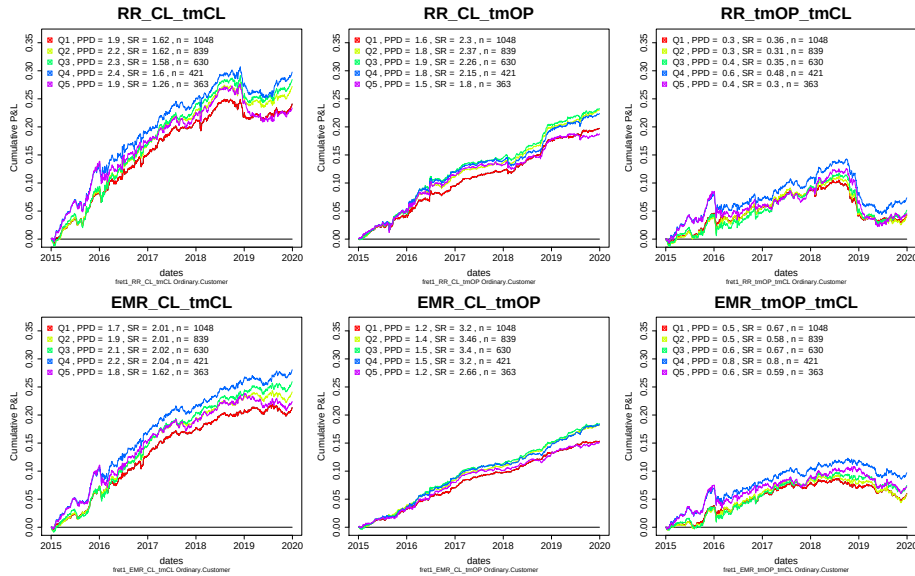


Figure A.6: Cumulative P&L plot - Customer: This shows the cumulative P&L from following a Uniform Weighting Scheme - one-day holding period strategy, for the five quantile rank groups Q1-Q5, using the OVI of Market Makers for six different types of returns. Also given, are SR (Annualized Sharpe Ratio), PPD (Profit per dollar in basis points), and n (average number of instruments in the portfolios).

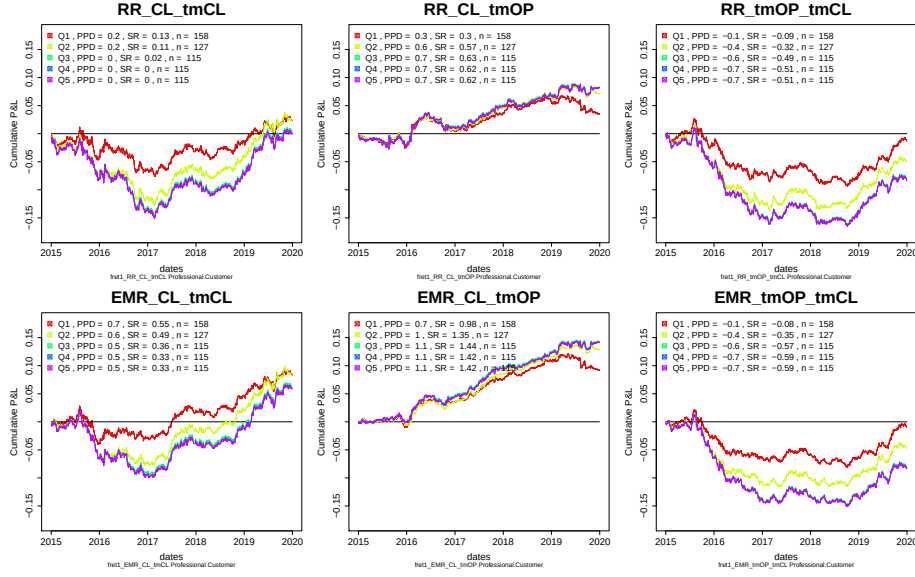


Figure A.7: Cumulative P&L plot - Professional Customer: This shows the cumulative P&L from following a Uniform Weighting Scheme - one-day holding period strategy, for the five quantile rank groups Q1-Q5, using the OVI of Market Makers for six different types of returns. Also given, are SR (Annualized Sharpe Ratio), PPD (Profit per dollar in basis points), and n (average number of instruments in the portfolios).

showing how uninformative they are as predictors for future returns. The other three MPCs fare much better, with the Market Makers having a rate of informative P&L's above 60%, which is much higher than what would be expected at random (around 50%). Depending on the quantile rank, the OVI of Customers and Brokers results in a proportion lying between 55 and 60%. Note that using the total number of trades or the nominal volume results in similar results.

A.4 Broader Scope Analysis: Additional results

In this section, we list additional results for Sections 6 and 7 of the main paper.

A.4.1 Decomposition based on Option Features

As mentioned in Section 6 of the main paper, we consider how certain option features affect the predictability of the OVI. Table A.1 shows a summary of the analysis, including results corresponding to Broker and Customer OVI, which were not shown in the main paper.

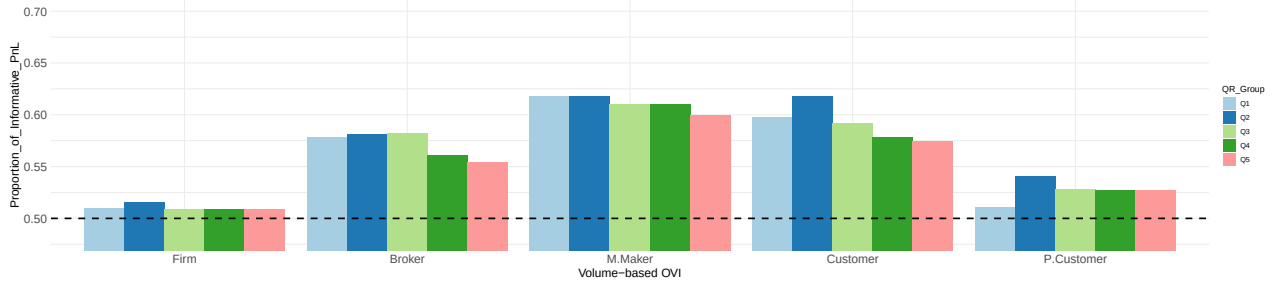


Figure A.8: P&L accuracy: The ratio of correct predictions for the EMR CL_tmOP returns across days is shown for each of the OVI measures, for the five quantile rank groups. Accuracy is defined as $\max\{\pi, 1 - \pi\}$, where π is defined as the ratio of positive daily P&L.

A.4.2 Difference between Opening and Closing Transactions

For MPCs other than Market Makers, we have a distinction between opening and closing transactions. In both the cases of Brokers and Customers, the number of closing transactions is much lower than for opening transactions. Therefore, in both cases there is no significant difference between opening transactions and the complete data set, whereas the performance of closing transactions is quite low.

A.4.3 Influence of underlying equity's features on OVI's predictive power

Similar to the experiments for the influence of option features on the predictive power of OVI, we divide underlying equities into four buckets based on some characteristic of the underlying. For Market Cap, there is no significant difference across buckets. For Turnover (both the daily, as well as the median of the last 21 trading days), higher turnover seems to be linked with higher predictability. However, the differences in SR are not significant in either cases.

We also consider differences in performance between different sectors. In Figure A.9, we show the boxplots for SRs obtained when running a Q3-Market Maker strategy, grouped by sector¹. We observe that most of the sectors have a positive SR on average, even though the lower quantile of P&L for most sectors falls below 0. However, the discrepancies across sectors are not statistically significant, as we do not get a rejection for an ANOVA test when testing for group differences.

The results seem to suggest that predictability does not depend as much on the features of the underlying, but more so on the features of the option features, as shown

¹Equity sectors are taken from <https://www.marketbeat.com/>.

Option Feature	Definition	Best Performing Bucket		
		Brokers	Market Makers	Customers
Delta (Δ)	$\Delta = \frac{\partial P^{\text{Option}}}{\partial S}$	1 st and 3 rd	1 st	1 st and 3 rd
Gamma (Γ)	$\Gamma = \frac{\partial \Delta}{\partial S} = \frac{\partial^2 P^{\text{Option}}}{\partial S^2}$	None	4 th	3 rd -4 th
Theta (Θ)	$\Theta = \frac{-\partial P^{\text{Option}}}{\partial \tau}$	None	2 nd -4 th	3 rd -4 th
Vega ($\mathcal{V}$)	$\mathcal{V} = \frac{\partial P^{\text{Option}}}{\partial \sigma}$	None	None	1 st and 3 rd
Rho (ρ)	$\rho = \frac{\partial P^{\text{Option}}}{\partial r}$	2 nd	1 st -2 nd	1 st -2 nd
Moneyiness (Mon)	$\begin{cases} \text{Mon} = +\frac{\log\left(\frac{Se^{\tau\sigma}}{K}\right)}{\sigma\sqrt{\tau}}, & \text{for call options} \\ \text{Mon} = -\frac{\log\left(\frac{Se^{\tau\sigma}}{K}\right)}{\sigma\sqrt{\tau}}, & \text{for put options} \end{cases}$	None	2 nd	2 nd -4 th
Maturity	Number of days until expiration.	None	None	None
Implied Volatility (σ)	σ that solves (2.27) / (2.28)	4 th	4 th	4 th

Table A.1: Summary information for option feature decomposition: For each option feature, its definition is given as well as the "Best Performing Bucket" for each MPC, which is the bucket that displays the highest Sharpe Ratio.

in Section 6 of the main paper. Finally, note that when we test the SR obtained by following the Q3 Market Maker strategy (with a Uniform Weighting Scheme) for each individual instrument, 148 of instruments reject the SR significance test ($\text{SR} = 0$) at a 5% level, while only one instrument rejects the hypothesis for a Bonferonni corrected test.

A.4.4 Cross-Impact and Networks

We now turn our attention to the network produced by following a Q3 and row-level corrected Bonferonni level (i.e. $\frac{0.05}{\text{\#Number of Instruments}}$) network, as discussed in Section 7 of the main paper. The resulting network has one large connected component. In the weak sense (dropping the directions of the edges), it includes all but two instruments, while in the strong sense, it includes 942 out of 1792 instruments, where the rest of the instruments form no other component.

In Figure A.10, we show a histogram of the in- and out-degrees, in the normal and log-scale. The in-degrees give a p -value of 0 when testing the null hypothesis that they follow a power-law distribution. Out-degrees do not reject and give a coefficient of 3.75, but from Figure A.10, it seems quite clear that the resulting network does not follow a power-law, as about half of the produced networks reject such a hypothesis.

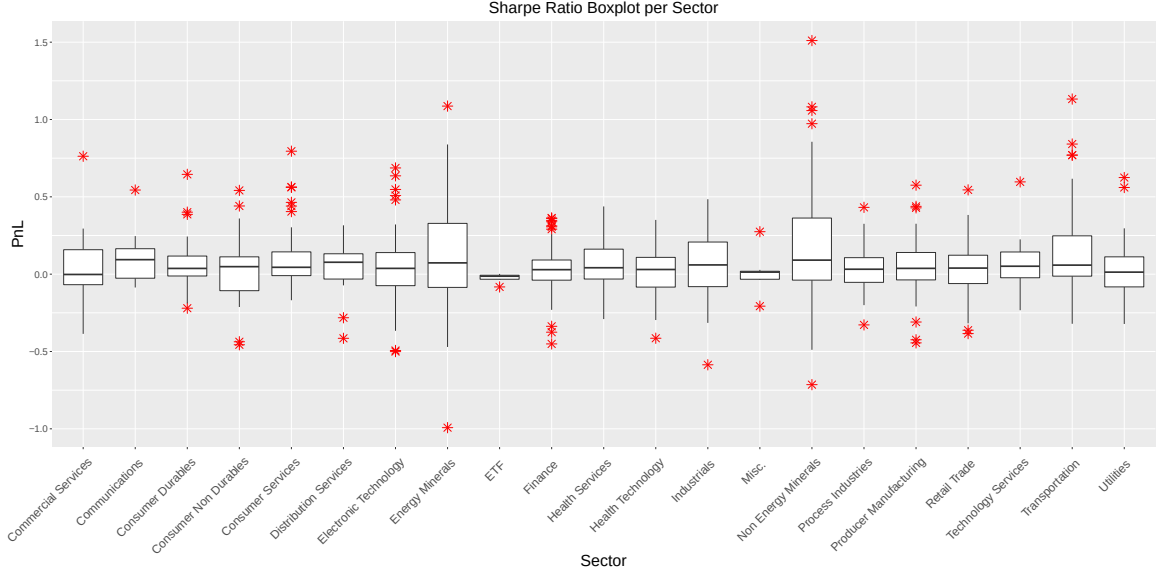


Figure A.9: Boxplots of Sharpe Ratios for each instrument, obtained using the Market Maker Q3 strategy. Results are shown in the form of boxplots, broken down by sector.

We consider the average degrees based on each underlying sector in Figure A.11, based on the Q3-market maker network with row-level Bonferonni correction. Some interesting observations are:

- Transportation is the sector with the highest average in degree, and the lowest average out degree.
- Distribution Services, Process Industries and Utilities sectors also have higher in-degree than the average node. On the other side, Health Services, Consumer-Durables and Energy Minerals have lower than average in-degrees.
- The out-degrees seem to vary more than than in-degrees. The communication sector as well as ETF's have much higher out-degrees than average, while Consumer Non Durables and Producer Manufacturing sectors also have higher than average out-degrees. Reversely, the Distribution service has a lower than average out-degree.

We also consider a bipartite graph, under the Chung-Lu model similarly to the main paper, where weights are proportional to the out-degree of the outgoing node and the in-degree of the incoming node. No pair of sectors has a value significantly different from the actual number of links observed. In other words, there is no significant evidence for a sector-sector cluster structure.

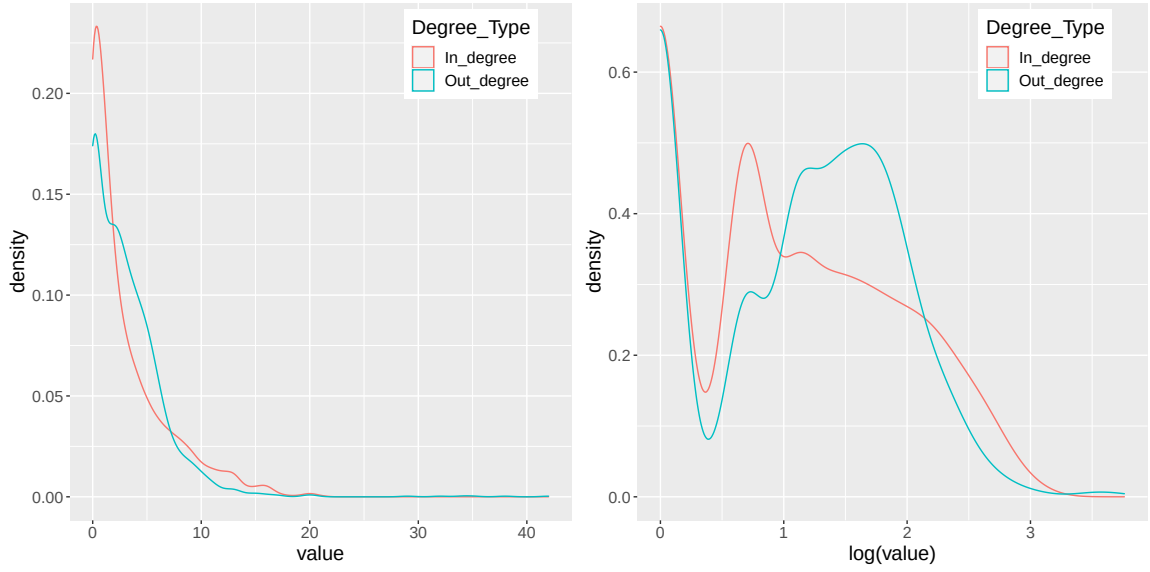


Figure A.10: In and Out-Degrees of Network. The left sub figure shows the degree density plot, while the right sub figure shows the density of the logarithm of the density, corresponding to a Q3 row-level Bonferonni Corrected network.

Lastly, we consider the centrality of the nodes. Besides the in- and out-degrees, we also consider four more centrality measures: Eigenvector, Katz, Page-Rank and Betweenness Centrality. In Figure A.12, we show the instruments with the highest centralities (we choose the five instruments with the highest value for each centrality). Note that instrument with high in-degrees also tend to have high Katz centralities and vice versa, while instruments with high Katz Centrality also tend to have a high Page rank.

A.5 Intra-day strategies.

The interplay between OVI and future returns in the main paper was conducted on a daily level. We also briefly consider an extension to the intraday case and report on some of the preliminary observed patterns. We consider the five MPCs, using three QR Groups defined similarly as before, and consider various types of intraday excess market returns for different time horizons ($\{1, 5, 10, 30, 60, 90, 180\}$ -minute future returns, as well as overnight returns). We refer to the OVI used as “open- t ” OVI, as it captures all transactions conducted from the market’s opening up to current time t . Note that a limitation of this analysis is that the universe of underlying assets is smaller and constrained only to 2018-2019 due to the availability

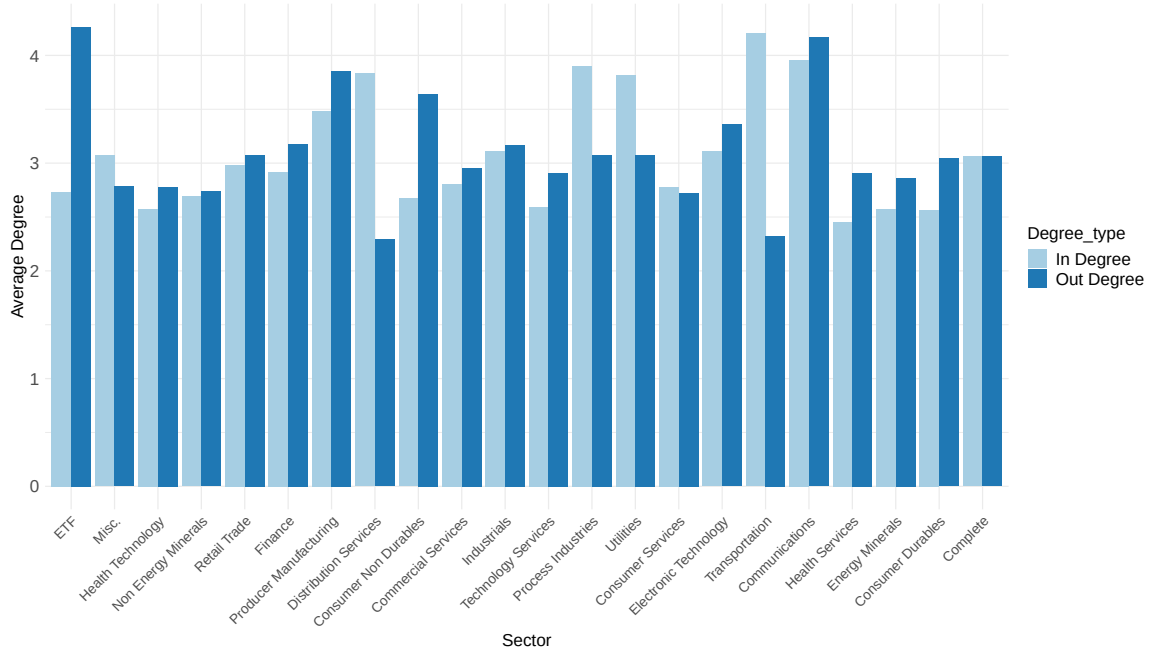


Figure A.11: In and Out Degrees per Sector: This figure shows the average degree for nodes corresponding to each sector of instruments, for the Q3 row-level Bonferonni Corrected network discussed in Section 8 of the main paper. The right-most bar shows the average degree of the whole network.

of intraday data². This renders any direct comparison between the intraday and daily returns not possible at this stage. Therefore, we limit ourselves to some interesting qualitative observations. Indeed, running a fully Bonferonni Corrected test taking into account all tests (correcting for the three QR Buckets, ten return modes, five MPCs and 39 values for t). In this case, none of the SRs comes out as significant; using the less conservative, Benjamini-Hochberg Procedure does not end up in any rejection either.

In most cases, Market Makers attain the most significant and consistent performance. The open-t OVI at mid-day results in a moderately high SR (maximized at 13:10), when we consider returns for 30-180 minutes (see for example Figure A.15). Furthermore, setting t near midday, results in positive CL_tmOP returns. Thus, a holding period from midday to opening of next day seems to be the optimal strategy when using the Market Maker OVI. If the holding period is further extended to include overnight returns (i.e CL_tmCL returns), then the SR remains positive but is decreased in magnitude.

Across MPCs, for either open-t or t-close OVIs and a return window of 60, 90 or

²Returns are calculated using LOB data, by considering the midpoint of the bid and ask price

180 minutes, a seasonal-like effect can be observed for SRs. For example, we observe this pattern in Figure A.14, for the open- t Firm OVI used for 90 minute EMR. Setting t to any time earlier than 10:30am, yields a consistently positive profit. For t after 10:30am, we observe a decrease in predictability and this drop eventually results in a negative SR for t later than 11:00am. Fluctuations continue with returns changing signs again at 12:30pm, 13:20pm and 15:10pm, displaying a somewhat periodic pattern. This seasonality of the time series appears in the other MPCs as well, albeit to a lesser extent. Across MPCs, these oscillations also appear to be dumped, as they usually decrease in magnitude (either in terms of SR or PPD).

In Figure A.13, we observe that the option volumes follow a U-shape similar to what is observed for the spot market. Higher volumes should make OVI a more reliable signal, which could explain the highest performance observed when using the volumes from the early part of the day.

Due to the rather restrictive data sets employed in our analysis, the above intraday results are difficult to quantify. On the other hand, a lot of observed SRs exceed 1, and are considered statistically significant (when ignoring Bonferonni correction). The oscillations of the observed SR with respect to t , also explain why CL_tmCL are worse performing than CL_tmOP. The Market Maker OVI is the exception, as its corresponding SR does not revert sign as t increases within the day, and remains mostly negative. This appears to be in line with the observation that the Market Maker's OVI has the highest predictability for longer horizon returns.

Instrument		In-Degree	Out-Degree	Eigenvector Centrality	Katz Centrality	Page Rank	Betweenness Centrality
CGA	China Green Agriculture, Inc	20	5	0.065	2.426	0.002	32927
DLNG	Dynagas LNG Partners LP	21	2	0.048	2.635	0.003	8963
FFC	Flaherty & Crumrine	20	3	0.058	2.821	0.004	25454
TG	Tredegar Corporation	20	1	0.065	2.889	0.003	1663
TKC	Turkcell Iletisim Hizmetleri A.S.	20	6	0.084	2.946	0.004	37277
GLD	SPDR Gold Trust	0	34	0.031	0	0	0
CL	Colgate-Palmolive Company	9	38	0.122	1.123	0.001	69515
HNI	HNI Corp	0	35	0.08	0	0	0
INN	Summit Hotel Properties Inc	4	42	0.121	0.529	0.001	64119
TNC	Tennant Company	4	32	0.092	0.489	0	33523
HIG	Hartford Financial Services Group Inc	16	11	0.101	2.237	0.002	55785
URI	United Rentals, Inc.	1	29	0.087	0.275	0.001	53813
HNP	Huaneng Power International Inc ADR Class N	17	2	0.058	2.752	0.005	21177
STNG	Scorpio Tankers Inc.	16	5	0.077	2.657	0.003	39681
AVK	Advent Convertible and Income Fund	13	3	0.06	2.119	0.005	29248
BCX	BlackRock Resources Comments Strategy Trust	16	4	0.056	2.446	0.005	27603
ITW	Illinois Tool Works Inc.	12	4	0.046	1.783	0.005	47104
KMPR	Kemper Corp	9	1	0.029	1.52	0.006	689
PFO	Flaherty & Crumrine Pref. Income Opportunity Fund	4	0	0.009	0.675	0.006	0
TGI	Triumph Group Inc	10	3	0.028	1.409	0.005	18344
BXP	Boston Properties Inc.	5	6	0.056	0.967	0.001	60198
GPK	Graphic Packaging Holding Company	13	3	0.047	1.958	0.004	88019
RGA	Reinsurance Group of America Inc	13	9	0.065	1.969	0.004	56358

Figure A.12: Selection of Instruments with high centralities for different centrality measures. Red cells denote high centrality for the metric corresponding to the particular column.

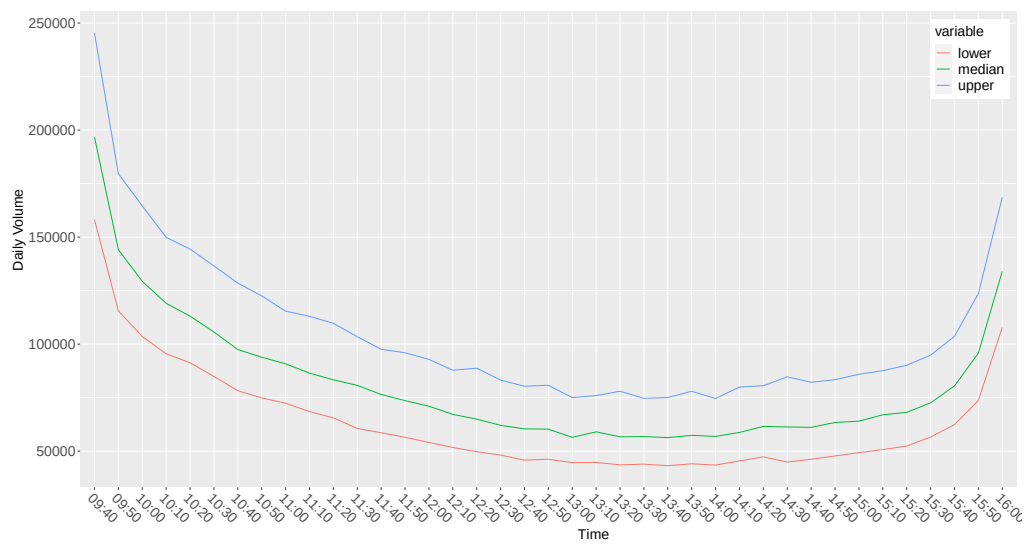


Figure A.13: Intraday Volume: The median total option volume at different time points is shown with the black line, while the upper quantile (resp. lower quantile) volume is given by the red line (resp. blue line).

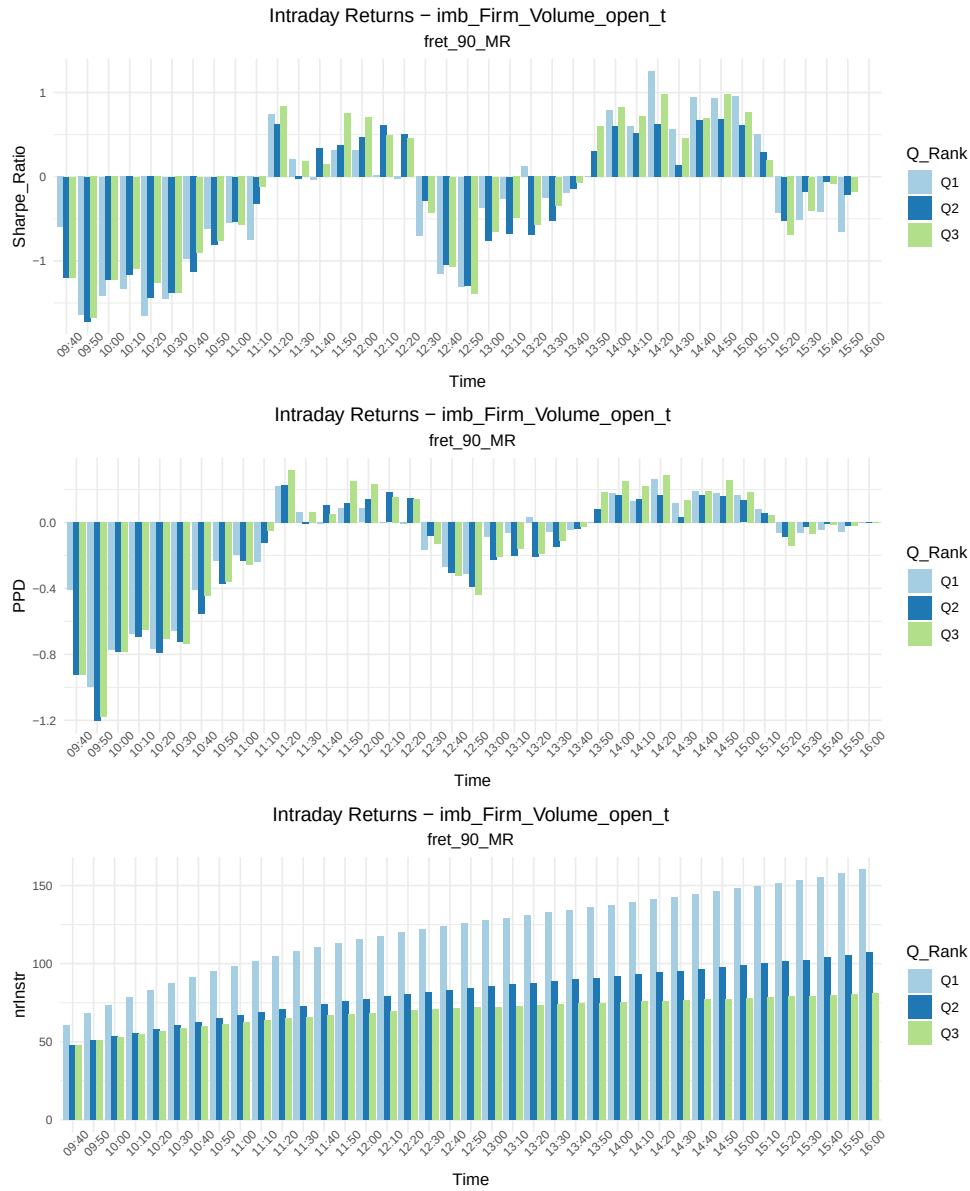


Figure A.14: Intraday Returns for Firm open_t OVI: This Figure shows the 90-minute market excess returns by following an open_t (use all volumes from opening up to t) strategy against t. The quantities shown are given for 3 cumulative quantile rank groups. The top figure shows the (Annualized) SR corresponding to each strategy, the second figure shows the Profit Per Dollar (PPD) and the third figure shows the average number of instruments used for each portfolio.

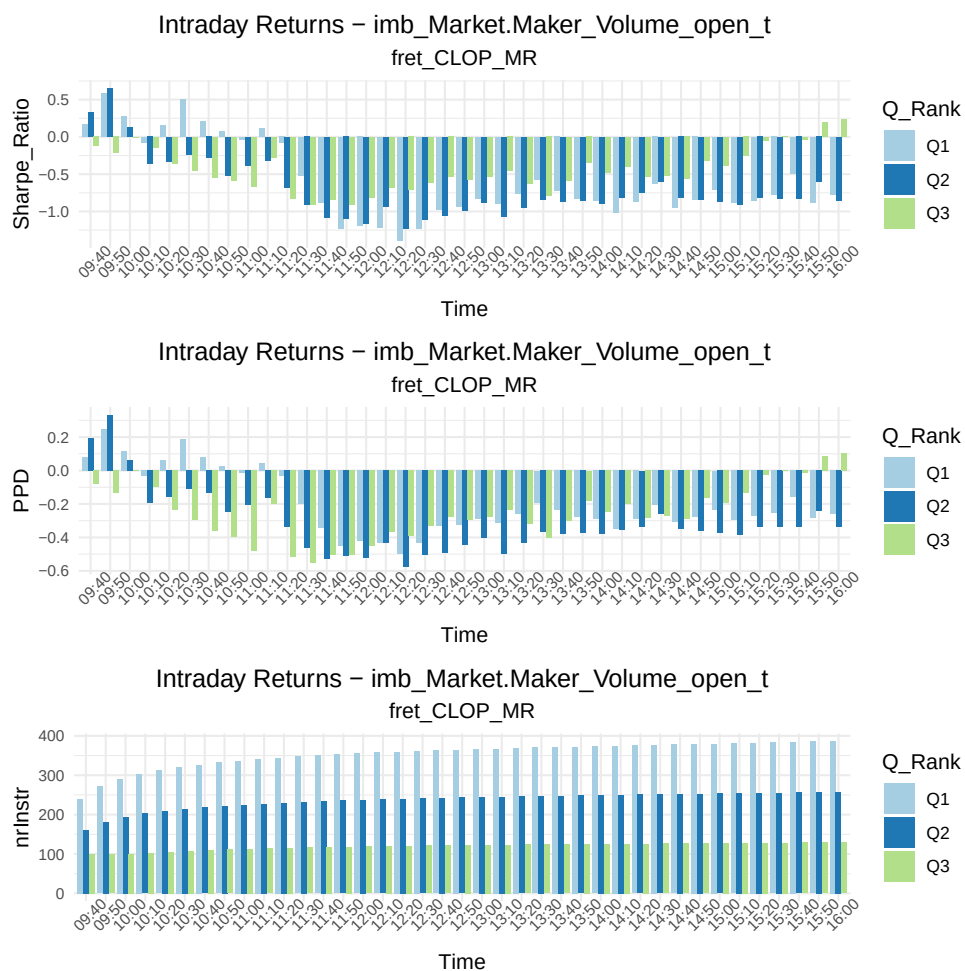


Figure A.15: Intraday Returns for Market Maker open_t OVI: This Figure shows the overnight (CL_tmOP) market excess returns by following an open_t (use all volumes from opening up to t) strategy against t. The quantities shown are given for 3 cumulative quantile rank portfolios. The top figure shows the (Annualized) Sharpe Ratio corresponding to each strategy, the second figure shows the Profit Per Dollar (PPD) and the third figure shows the average number of instruments used for each portfolio.

Appendix B

Appendix for OFTER: An Online Pipeline for Time Series Forecasting

B.1 A convexifying approach for solving the Secular equation

In this section, we give a brief description of the method we use to solve the secular equation described in Section 3.3.1, which amounts to finding the roots of

$$f(\kappa) = 1 + \rho \sum_{j=1}^d \frac{z_j^2}{\lambda_j - \kappa}. \quad (\text{B.1})$$

The main difficulty in directly solving Equation (B.1) lies in the multitude of discontinuities in the domain of f . An efficient approach to find the i^{th} root of the equation, would be to consider the reparameterizations $\kappa = \lambda_i + \rho\xi$ and $\delta_j = \frac{\lambda_j - \lambda_i}{\rho}$, which leads to the formulation of Equation (B.1) as

$$f_i(\xi) = 1 + \sum_{j=1}^d \frac{z_j^2}{\delta_j - \xi}. \quad (\text{B.2})$$

By assuming that each eigenvalue is distinct (a reasonable assumption when dealing with real data), we notice that

$$\delta_1 < \delta_2 < \dots < \delta_i = 0 < \dots < \delta_n. \quad (\text{B.3})$$

Therefore, we look for solutions in the set $(0, \delta_{i+1})$, for $i < n$, and in $\mathbb{R}_+$ for $i = n$. We can further transform $f_i(x)$ into a convex function, by considering $F_i(\gamma) = f_i(\gamma^{-1/2})$, which can be shown to transform F_i into a convex function in the interval $(\frac{1}{\delta_{i+1}}, \infty)$ for $i < n$ and $(0, \infty)$ for $i = n$.

Finally, by using the fact that, for a decreasing convex function, the Newton method¹ monotonically converges to a solution [100], it can be established that, if γ^* is the solution to $F_i(\gamma) = 0$, then the Newton method converges monotonically to γ^* with any starting point in $\left[\frac{1}{\delta_{i+1}^2}, \gamma^*\right]$. For a more detailed explanation and an alternative based on the secant method, we refer the reader to [166].

B.2 Brief summary of baseline methods

The following methods are used as baselines against OFTER. We provide a short description of each one.

¹We remind the reader that the Newton method searches for a solution to $f(x) = 0$, starting from point x_0 , and then considers the sequence

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)},$$

until a solution is obtained.

Ordinary Least Squares (OLS)

Ordinary Least Squares (OLS) is by far the most well known regression method, where the target variable y is modeled as a linear combination of the features

$$y_i = \beta_0 + \sum_{j=1}^d \beta_j X_{i,j} + \epsilon_i. \quad (\text{B.4})$$

The coefficient vector $\beta = (\beta_0, \dots, \beta_d)$, is chosen as the one maximizing $\sum \epsilon_i^2$, and can be found using the closed form formula

$$\beta = (X^T X)^{-1} X^T y. \quad (\text{B.5})$$

Since we include lagged features in the feature matrix, this method is very similar to a Vector Auto-regressive Regression (VAR) model.

Principal Component Regression

To combat the high correlation present in the features, Principal Component Regression (PCR) first identifies the principal components of X , obtaining $\tilde{X}$, and then a number of components $p < d$ is retained. OLS is then used, but on the first p columns of $\tilde{X}$ instead of X . Similar to OFTER, we use $\delta = 0.9$ for choosing the number of components.

Least Absolute Shrinkage and Selection Operator

An alternative to reducing the dimensionality of X before OLS, is to include a regularization term in the objective function Equation (B.4). Least Absolute Shrinkage and Selection Operator (LASSO) is a widely popular method, where the modified objective function takes the following form

$$(y - X\beta)^T (y - X\beta) + \lambda \sum_j |\beta_j|. \quad (\text{B.6})$$

Note that in this case, there is no closed-form solution, and thus a numerical optimization method is utilized. We choose λ using a cross-validation procedure.

Random Forests

A decision tree is a simple method inspired by decision analysis, where the visualization of the model resembles a tree consisting of nodes and branches. The initial

feature space constitutes the root node. The observations in the root node are partitioned into 'successor nodes', based on a splitting rule chosen so as to optimize some criterion (usually either entropy or Gini Impurity). The process is repeated recursively on each of the successor nodes until some stopping criterion is met. For each new data point, its features are then used to determine its corresponding leaf node. Forecasts can then be obtained by averaging the target value of all observations that fall within each leaf node. This is a non-parametric method, but with a very high variance.

A Random Forest is built by fitting multiple decision trees. Each decision tree is fitted using only a random sample of observations and a random sample of features. The forecasts of the decision trees are then combined into a single forecast (usually by averaging). Random Forests are based on the idea of bagging aimed at reducing variance and avoiding overfitting.

For the benchmark task, we use 200 trees, which are built using the squared error criterion. The proportion of used features is set to 0.333, as recommended by [109].

XGBoost

Boosting is a class of ensemble models, where the weak learners usually take the form of decision trees. At each iteration, a new weak learner is fitted whose training phase focuses on the previously misclassified data points. The eXtreme Gradient Boosting (XGBoost) is one of the most widely used variants of boosting. We use the standard sklearn implementation of XGBoost, with the default parameters.

Support Vector Machines (SVM)

SVM is another popular machine learning method that is motivated by learning theory. In particular, a vector β and a value β_0 are found that minimize

$$\frac{1}{2} \|\beta\|^2, \quad \text{s.t.} \quad |(y_i - \beta^T x_i - \beta_0)| \leq \epsilon. \quad (\text{B.7})$$

A prediction for a new point x' is then given by $\beta_0 + \beta^T x'$. We use the default settings of sklearn, by choosing $\epsilon = 0.1$.

Recurrent Neural Networks

Recurrent neural networks (RNN) form a class of neural networks specifically aimed at data sets with temporal structures. This is generally achieved by introducing

additional skip connections that pass the output of a past layers to the current one. One issue caused is that gradients have the tendency to vanish or explode.

Long-Short Term Memory is a variant of Recurrent Neural Networks, specifically designed to avoid the issue of exploding or vanishing gradients. This is achieved by introducing a cell state, which contains a summary of the previous cells and is controlled by a number of different gates. Denote by x_t the feature vector of a data point, c_t the input of the cell state, and h_t the output of the module. The formula of the basic LSTM takes the form

$$\begin{cases} f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f), \\ i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i), \\ o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o), \\ \hat{c}_t = \tanh(W_c x_t + U_c h_{t-1} + b_c), \\ c_t = f_t \odot c_{t-1} + i_t \odot \hat{c}_t, h_t = c_t \odot \tanh(c_t), \end{cases} \quad (\text{B.8})$$

where $W_f, U_f, b_f, W_i, U_i, b_i, W_o, U_o, b_o, W_c, U_c$ and b_c are trainable parameters.

We consider two more variants of LSTM. Bi-LSTM which utilizes two parallel and reverse networks of LSTMs so that information can flow in both directions. We also use Gated Recursion Unit (GRU) which has a similar yet simpler architecture than LSTM.

For each variant, the layout consists of two layers, each with a unit size of 20, and a fully connected layer. The lookback period is set equal to 3, to match the data used for the other methods. For the training, 50 epochs appear to be enough. For optimization ADAM is used with learning parameter equal to 0.0001, and the other parameters set as default in the corresponding Keras package.

B.3 Robustness analysis

In this short section, we demonstrate how changing five of the main parameters used in OFTER-DR-FT affects its performance on the synthetic data set. We fix five of the main parameters given in OFTER as follows: $\delta = 0.9$, $L = 800$, $c_{\min} = 0.05$, $c_{\text{Original}} = 0.05$, and $p_{\text{ADF}} = 0.05$. By keeping the other four parameters fixed, we vary each of these parameters in turn and note the variation in its performance. In Table B.1, we show the correlation between the target and the forecasts when changing each of the noted parameters to the given value.

In general, the performance of the algorithm does not significantly depend on the choice of parameters, as forecasting each feature has consistent performance.

However there are exceptions; for example, $c_{\min}$ appears to have a significant effect on the performance of the algorithm for M2.3 and M2.4.

Value	δ		p_{ADF}		L		$C_{\min}$		C_{original}	
	0.8	0.95	0.01	0.1	500	1500	0.01	0.1	0.01	0.1
M1.1	0.470	0.459	0.467	0.467	0.441	0.477	0.464	0.464	0.465	0.461
M1.2	0.519	0.511	0.509	0.509	0.503	0.521	0.509	0.514	0.504	0.518
M1.3	0.077	0.071	0.085	0.085	0.085	0.113	0.119	0.116	0.085	0.117
M1.4	0.554	0.558	0.552	0.552	0.555	0.556	0.553	0.556	0.545	0.559
M1.5	0.094	0.107	0.101	0.101	0.147	0.120	0.102	0.110	0.101	0.113
M2.1	0.599	0.623	0.599	0.599	0.596	0.603	0.675	0.600	0.599	0.60
M2.2	0.535	0.535	0.535	0.535	0.534	0.539	0.897	0.535	0.536	0.536
M2.3	0.864	0.799	0.815	0.815	0.776	0.850	0.759	0.401	0.804	0.817
M2.4	0.962	0.964	0.963	0.963	0.963	0.963	0.962	0.963	0.963	0.963
M2.5	0.679	0.681	0.681	0.681	0.678	0.682	0.681	0.678	0.680	0.679
M3.1	0.758	0.758	0.758	0.758	0.751	0.762	0.758	0.758	0.758	0.757
M3.2	0.852	0.851	0.851	0.851	0.85	0.852	0.851	0.850	0.849	0.850
M3.3	0.669	0.664	0.665	0.665	0.662	0.668	0.665	0.666	0.664	0.667
M3.4	0.902	0.903	0.903	0.903	0.902	0.905	0.903	0.902	0.903	0.902
M3.5	0.826	0.827	0.827	0.827	0.825	0.828	0.827	0.826	0.825	0.825

Table B.1: Robustness analysis of OFTER. The table displays the correlation between the target and forecasts for each of the 15 features of the synthetic data set, by varying a subset of five parameters of the algorithm.

B.4 Non-Financial Application

As a non-financial application we also consider the Beijing air quality data set from the UCI Machine Learning Repository. Data is collected by the Beijing Municipal Environmental Monitoring Center, and considers various air quality measurements from 12 monitoring points. We aim at forecasting each of the six types of pollution (PM2.5, PM10, O3, CO, SO2, NO2) concentrations (in ug/m^3), for each of the stations. Other features in the data set, used as predictors, include Temperature, Dewpoint Temperature, Precipitation, Pressure as well as Wind speed and Direction.

We use $L = 800$ and set $c_{\min} = c_{\text{Original}} = 0.01$, $L_D = 500$. In Figure B.1, we show a grid of the proportions of cases where one method outperforms another based on the RMSE metric. As it can be seen, all four OFTER configurations are outperforming all of the other eight benchmarks. In this particular case, OFTER-DR-FT outperforms the other three versions of OFTER.

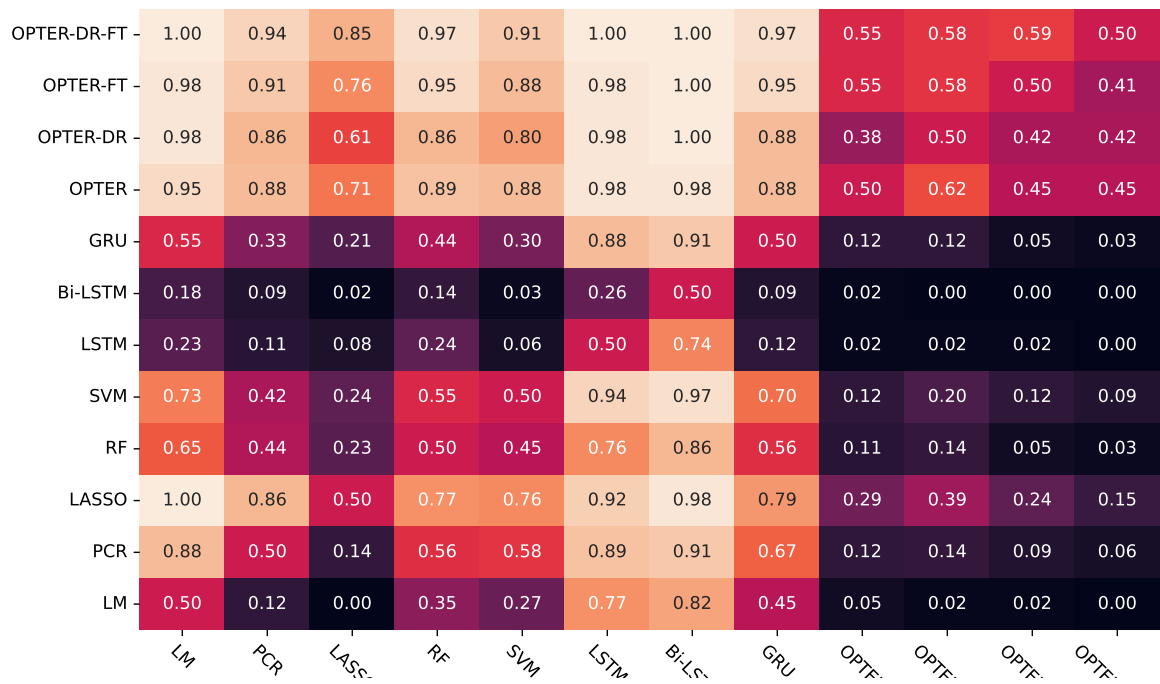


Figure B.1: This heatmap is a visualization of the comparison between the different forecasting models for the air pollution data set. The proportion of the number of cases where a method (label of the vertical axis) outperforms another method (label of the horizontal axis) in terms of RMSE given in the corresponding cell.

Appendix C

Appendix for Options-driven volatility forecasting

C.1 Heston Model Calibration

In this section we describe how to fully fit the Heston model described in Section 4.4, for the initial parameter estimates. Our methodology uses a combination of the particle swarm optimization (PSO) method and a gradient descend method (in this case ADAM). This way we take advantage of the faster gradient method, and employ the slower (but less prone to getting stuck in local minima) PSO only when the obtained parameter set is a bad fit.

Specifically, the parameter set for the first day of the data set is found using PSO. Then for each subsequent day a gradient descent method is used with the starting parameter vector being equal to the one fitted for the previous trading day. If the obtained RMSE (see equation Equation (4.49)) is lower than a pre-defined threshold $\text{RMSE}_{\max}$, we accept it, otherwise we run PSO. Finally, between the parameter vectors obtained using PSO and the gradient method, we choose the one that corresponds to a lower RMSE.

In the rest of this section we give a description for each of the two methods described. Note that the function to be minimized is the RMSE as defined in Equation (4.49).

$$f(\theta_d) = \sqrt{\frac{1}{N_d} \sum_{i=1}^{N_d} w_i r_i(\theta_d)^2}. \quad (\text{C.1})$$

For convenience, we drop the day index d .

C.1.1 Particle Swarm Optimization

Particle Swarm Optimization [129] (PSO) is a computational optimization method, that does not require the underlying problem to be convex, and can find a solution even when there is a multitude of local minima. A closed box (Cartesian product of closed intervals) U within which a solution needs to be specified. Thus, with this restriction, we aim to solve $\min_{\theta \in U} f(\theta)$. The algorithm works by considering moving particles $p_1, \dots, p_N \in U$. The central idea is that each of these particles explores the parameter space U , with varying velocities $v_1, \dots, v_N > 0$. The multiplicity of particles ensures sufficient exploration, thus avoiding local optima. Each of the particles' velocities depends on the optimal solution encountered by that particle so far, while the velocity of the swarm depends on the overall optimal solution. In Algorithm 3, we give the complete algorithm we use.

Algorithm 3: Particle Swarm Optimization

Input: U (parameter space), p (number of particles), w (inertia parameter),
 ϕ_p, ϕ_g (velocity update parameters), $t_{\max}, \epsilon, t_{\text{conv}}$ (convergence
parameters)

Output: θ^*

Initialize $t \leftarrow 1, C_1 = C_2 = C_3 = 0$

for $i \leftarrow 1$ to p **do**

 Initialize the particle's position $x_{1,i}$ by uniformly drawing a point from U

 Initialize the particle's velocity $v_{1,i}$ to 0

 Initialize the particle's best known position to $\rho_{1,i} \leftarrow x_{1,i}$

end for

Initialize the swarm's best known position to $g_1 \leftarrow \operatorname{argmax}_{x \in \{\rho_{1,1}, \dots, \rho_{1,p}\}} f(x)$

while $C_1 + C_2 + C_3 = 0$ (Termination Criteria are not met) **do**

for $i \leftarrow 1$ to p **do**

 Draw $r_p, r_g \sim U(0, 1)$

 Update the particle's velocity as

$v_{t+1,i} \leftarrow wv_{t,i} + \phi_p r_p (\rho_{t,i} - x_{t,i}) + \phi_g r_g (g_t - x_{t,i})$

 Update the particle's position as $x_{t+1,i} \leftarrow x_t + v_{t+1,i}$

 Update the particle's best known position as $\rho_{t+1,i} \leftarrow \operatorname{argmax}_{x \in \{x_{t,i}, x_{t+1,i}\}} f(x)$

end for

 Update the swarm's best known position as $g_{t+1} \leftarrow \operatorname{argmax}_{x \in \{\rho_{t+1,1}, \dots, \rho_{t+1,p}\}} f(x)$

$t \leftarrow t + 1$; $C_1 \leftarrow 1_{t \geq t_{\max}}$; $C_2 \leftarrow 1_{f(g_{t+1}) \geq \epsilon}$; $C_3 \leftarrow 1_{g_{t+1} = g_{t+1-t_{\text{conv}}}}$

end while

Set $\theta^* = g_t$

ADAM Out of the large family of gradient methods, we opt to use ADAM [133], a stable in optimization, due its simplicity and efficiency. This is an iterative algorithm which starts with a parameter vector $\theta^{(0)}$, and iteratively carries out the following updating step

$$\theta_j^{(n)} = \theta_j^{(n-1)} - \eta \frac{\frac{m_j^{(n)}}{1-\beta_1^n}}{\sqrt{\frac{v_j^{(n)}}{1-\beta_2^n} + \epsilon}}, \quad (\text{C.2})$$

$$m_j^{(n)} = \beta_1 m_j^{(n-1)} + (1 - \beta_1) \nabla_j f(\theta^{(n-1)}), \quad (\text{C.3})$$

$$v_j^{(n)} = \beta_2 v_j^{(n-1)} + (1 - \beta_2) (\nabla_j f(\theta^{(n-1)}))^2, \quad (\text{C.4})$$

where we set $m_j^{(0)} = v_j^{(0)} = 0, \forall j$. The update step is repeated until one of the following terminal criteria is met.

- $f(x) \leq \epsilon_1$
- $|\theta_n - \theta_{n-1}| \leq \epsilon_2$
- $n \geq n_{\max}$

We use the configuration of the original ADAM paper, with $\eta = 0.001, \epsilon = 10^{-8}, \beta_1 = 0.9, \beta_2 = 0.999$. For the terminal conditions, we use $n_{\max} = 500, \epsilon_1 = 0.01, \epsilon_2 = 10^{-4}$.

C.2 Financial Trading Metrics

In Section 4.6.1 we consider a hypothetical trading strategy. In this section, we shortly define the financial metrics used in Figure 4.8. Let P_d^{open} and P_d^{close} denote the opening and closing price of an asset for day d . The daily open to close return is defined as

$$f_d = \frac{P_{d+1}^{\text{close}} - P_{d+1}^{\text{open}}}{P_{d+1}^{\text{open}}}. \quad (\text{C.5})$$

Denote by $\hat{f}_d$, the forecast for f_d . The Profit and Loss (P&L) metric captures the total profit that is achieved by following the set of forecasts $\{\hat{f}_d\}_{d=1}^D$, for a trading strategy where as the underlying instrument is longed (shorted) whenever the sign of $\hat{f}_d$ is positive (negative).

$$\text{P\&L}_d = \sum_{d=1}^D f_d \text{sign}(\hat{f}_d). \quad (\text{C.6})$$

A measure of the performance of a strategy is the Sharpe Ratio [212], defined as

$$\tilde{\text{SR}} = \frac{\text{mean}(\mathbf{P\&L})\sqrt{252}}{\text{sd}(\mathbf{P\&L})}. \quad (\text{C.7})$$

We also define the P&L return per Day (PPD) as

$$\text{PPD} = \frac{\sum_{d=1}^{D-1} \text{P\&L}_d}{D-1}. \quad (\text{C.8})$$

It should be noted that we do not consider a trading strategy in a realistic setting, as the trading costs and liquidity constraints are ignored.