ALGORITHMIC DECISION MAKING IN FINANCIAL MARKETS

SID GHOSHAL

Department of Engineering Science

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Balliol College, Michaelmas 2018

In loving memory of my parents, Sumantra and Susmita.

Acknowledgements

I am deeply grateful to the following for their support throughout my research pursuits.

- To my supervisor Steve Roberts, whose peerless clarity of thought was aspirational and inspirational all through this journey. His research guidance and feedback - offered freely and promptly at even the busiest of times - kept my research agenda on-track and on-schedule.
- To the members of Oxford's Machine Learning Research Group, who expanded my understanding of machine learning. Their diverse research interests gave me a fuller appreciation for the seemingly boundless potential of the methods we were developing.
- To the members of the AIMS (Autonomous Intelligent Machines and Systems) CDT, the cohort through which I integrated the DPhil program at Oxford: you were the social bedrock of my doctoral studies. Thank you for the many insightful conversations we've had over the years - socially fulfilling and intellectually challenging in equal measure.
- To my dear brother Ananda, whose encouragement and unconditional support helped me immeasurably in achieving this milestone. His astute proofreading of this thesis reined in my tendency to allow technical minutiae to garble the core message of a chapter.
- To my partner Anna, who kindly and compassionately indulged me in many of the vices of the DPhil life - the late hours, frequent bouts of distracted thinking or mumbling, and interminable coffee consumption. You listened without complaint to my woes and setbacks, and cheered me on excitedly through every milestone all the way to the finish line. For all the little things you do: thank you.

Statement of Originality

I declare that no part of this thesis has been, or is being, submitted for any qualification other than the degree of Doctor of Philosophy at the University of Oxford. Notwithstanding the use of the first person of plural to express personal views, this thesis is the result of my own work unless otherwise stated. The core chapters of this thesis (Chapters 3, 4, 5 and 6) were all co-authored with my supervisor, Steve Roberts.

Several chapters of this work have been published as either journal or conference workshop papers, or are presently under review as described below.

Chapter 3:

Extracting Predictive Information from Heterogeneous Data Streams using Gaussian Processes. Ghoshal, S. and Roberts, S. *Algorithmic Finance*, 5(1-2), pp. 21-30.

Chapter 4:

Reading the Tea Leaves: A Neural Network Perspective on Technical Trading. Ghoshal, S. and Roberts, S. *KDD 2017, Mining and Learning from Time Series.*

Thresholded ConvNet Ensembles: Neural Networks for Technical Forecasting. Ghoshal, S. and Roberts, S. *KDD 2018, Data Science in Fintech.*

Thresholded ConvNet Ensembles: Neural Networks for Technical Forecasting. Ghoshal, S. and Roberts, S. *Neural Computing and Applications*, under review.

Chapter 6:

Short Memories? The Impact of SEC Enforcement on Insider Leakage. Ghoshal, S., Bengtzen, M. and Roberts, S. *Journal of Law, Finance and Accounting*, under review.

Sid Ghoshal, December 2018

Abstract

Machine learning's prowess for automatic pattern recognition at scale is meaningfully reshaping every branch of science. From astronomy to vision, web analytics to medical diagnostics, every data-intensive field is harnessing the potential of modern AI techniques. Though not commonly viewed through the same lens, finance is very much at the forefront of the data revolution. Financial markets present one of the most complex, noisy environments for machine learners: a vast range of factors - not all readily quantifiable - may impact a financial time series, and the relative salience of market variables may evolve through time.

The aim of this thesis is to investigate algorithmic frameworks for the challenging decisions faced by liquidity takers (the 'buy side') and market makers (the 'sell side'), the primary agents in financial markets. By extension, we also consider the behaviour of influential external agents such as regulators, whose actions affect the information landscape for buyers and sellers alike. This thesis deploys recent advances in machine learning to provide rational, data-driven tools to promote market efficiency in the areas of *price discovery*, *liquidity provision* and *financial regulation*.

Contents

Chapter 1

Introduction

The discipline of finance has undergone a radical transformation in the last two decades. Changes in financial regulation have sought to reduce the informational advantage of established, institutional investors and level the playing field for new entrants. The shift from slow, broker-based trading to online electronic platforms in the 2000s significantly reduced transaction costs and improved market liquidity, facilitating the rise of the systematic trader and heightening the need for competitive market making. We propose a fully algorithmic treatment of both pursuits - profit-generating program trading and risk-aware market-making - in a bid to push the current frontier in financial automation.

1.1 Market Microstructure

How exactly do trades occur? Much of the research in this thesis was seeded by a single, simple observation on market microstructure.

Trades in financial markets occur when one agent, the liquidity taker, chooses to buy or sell at the ask or bid price of a liquidity provider, the market maker.

Markets are driven by two distinct types of agent: liquidity takers (also known as the 'buy side', because they pay for access to markets), and market makers (also

known as the 'sell side', because they sell the service of continuous access to markets via the prices they display). To be successful, the liquidity taker must move ahead of the market: they must buy before it rises, and sell before it drops. By contrast, the market maker isn't formally required to take any views: they merely show two prices where they're willing to buy and sell, and offset buy orders with sell orders, capturing the gap as risklessly and frequently as possible. Embedded in the italicised statement above are three of the most fundamental questions in finance:

1. When is it optimal for the liquidity taker to act, i.e. what data should they be considering when optimising the timing of their trades? Another way of framing it is: can we, even softly, forecast market movements? The literature refers to this as the problem of *price discovery*.

2. Given this liquidity taker, what are the optimal bid and ask prices that the market maker should show, to balance their dual imperative of maximising profit while minimising risk? Can we improve on current methods for inventory control, by learning the buy side's behaviour and in some way modelling adverse selection? This topic falls under the umbrella of *liquidity provision*.

3. Is the interaction between these two agent sets affected by changes in the regulatory environment? In what manner do the actions (or inaction) of *financial regulators* impact the price formation process?

1.2 Price Discovery

At the heart of the liquidity taker's dilemma lies the crucial question of price discovery: determining the fair price for a security. How should a trader select or extract features from data, in order to optimise the timing of their trades? And assuming they can identify predictive content in a variety of datasets, how should they combine this information to generate strong forecasts?

1.2.1 Extracting Insight from Heterogeneous Data

Finance provides one of the most data-rich environments for machine learning researchers. The range of datatypes that can potentially influence an asset's price spans a wide gamut of numerical, textual and categorical input domains. Historical prices of the asset and its comparables, macroeconomic indicators as well as the shape of the order book can potentially influence the evolution of a time series. Event-driven changes in the market often first take form in textual disclosures: regulatory filings, press reports, even subjective editorials by respected commentators are all factors in efficient price discovery. Brokerage houses and equity research firms furnish us with another form of structured data via their buy, hold and sell recommendations, producing a potentially informative clustering of Wall Street's perspective on individual securities.

1.2.1.1 Contribution

Fusion of this vast array of heterogeneous data into a viable probabilistic model for time series forecasting remains an open challenge for both the finance and machine learning communities, and each of these four signal domains - technicals, sentiment, options markets, broker recommendations - warrants closer inspection before their fusion can be addressed. Though there are marginal gains in using multiple time series from the same domain, our contribution to this field lies in examining data fusion across domains, and assessing the incremental value of modelling inter-domain dependencies. This work, the first of our research papers, was peer-reviewed and published in 2016 ("Extracting Predictive Information from Heterogeneous Data Streams using Gaussian Processes", *Algorithmic Finance, Volume 5, Issues 1-2, 2016*).

1.2.2 Neural Networks for Technical Forecasting

Of the four data domains mentioned in the preceding section, technicals have garnered the most attention. The foremost reason for this is perhaps accessibility: actively traded financial assets possess an often publicly available price history. As price processes evolve in real-time, tick by tick, the quantity of data at hand permits the use of statistical tools denied to fundamental analysts, whose inputs update at a much lower frequency (e.g. quarterly for corporate earnings announcements in the US).

1.2.2.1 The Curse of 'Domain Knowledge' in Finance

Much of modern practice in financial forecasting relies on technical analysis, an umbrella term for several heuristics applying visual pattern recognition to price charts. Despite its ubiquity in financial media, the reliability of its signals remains a contentious and highly subjective form of 'domain knowledge'. Our aim in this domain is two-fold: firstly, to critically evaluate the predictive prowess of commonly-cited visual patterns in financial time series, and secondly, to assess the potential for deep learning in the technical domain.

1.2.2.2 Contribution

By reframing technical analysis as a poorly specified, arbitrarily preset feature-extractive layer in a deep neural network, we learn better convolutional filters directly from the data, and provide visual representations of the features being identified. This work was the topic of an oral presentation at KDD 2017's workshop on *Mining and Learning from Time Series* ("Reading the Tea Leaves: A Neural Network Perspective on Technical Trading"). In a subsequent extension of the work showcased at KDD 2018's *Data Science in Fintech*, we found that an ensemble of shallow, thresholded CNNs optimised over different resolutions achieves state-of-the-art performance on this domain, outperforming technical methods while retaining some of their interpretability.

This work is presently being reviewed under the title "Thresholded ConvNet Ensembles: Neural Networks for Technical Forecasting" for publication in *Neural Computing and Applications*.

1.3 Liquidity Provision

Market makers face the complex optimisation task of maximising their profit while minimising their inventory risk. This dual mandate exerts inherently opposing forces: the simplest way to avoid inventory is to maintain wide bid-offer spreads, yet doing so will prevent deal flow and therefore neuter profitability. Conversely, showing competitive two-way prices will ensure heavier volumes on both sides of the order book and therefore lock-in gains from clients crossing the bid-offer spread, but opens the hazard of building up sizeable inventory risk if trade flow is temporally asymmetric. A further concern for dealers is the threat of adverse selection by market professionals exploiting an information advantage.

1.3.1 Prevalent Frameworks

Existing work in finance and stochastic control provides a robust framework based on the Hamilton-Jacobi-Bellman (HJB) equation for dynamically balancing inventory (Ho and Stoll, 1981 and Avellaneda and Stoikov, 2008). However, existing methods do little to address adverse selection in over-the-counter (OTC) markets where prices can be individually tuned for each counterparty, such as the foreign exchange market.

1.3.2 Contribution

As part of a Bayesian approach to market making, we propose a data-driven adjustment to the inventory-optimal bid and offer prices. We build a high-dimensional Gaussian Process for each counterparty's behaviour, drawing on market maker data about client activity in response to time of day, anticipatory and reactive volatility,

returns at various timescales (1-minute return, 1-hour return, 1-day return) and trading volume. By learning a Gaussian Process representation for each counterparty, we identify patterns corresponding to opportunistic trading and construct a framework for countering adverse selection risk. Specifically, clients whose trades systematically pre-empt large moves (measured via anticipatory volatility and return metrics) may be exploiting an information advantage over the market maker. Persistently high returns provide a speculative basis for measuring adverse selection, with which we augment the Avellaneda and Stoikov model to control for both inventory risk and adverse selection risk simultaneously.

Another practically-minded contribution of this work involves the classification of clients on the basis of their historical opportunism. By constructing a matrix of Bhattacharyya distances, we measure the similarity between the probability distributions representing each counterparty pair. The resulting adjacency matrix can then be used to cluster different client types with community detection techniques built on Bayesian non-Negative Matrix Factorisation (Psorakis et al, 2011, 2012).

1.4 Financial Regulation

Significant research in both the legal and econometrics literature has gone into assessing the fairness of financial markets and the impact of regulatory changes. A particular area of interest to regulators is the flow of non-public information (NPI) from insiders to institutional investors: does the abuse of selective disclosure laws produce detectable footprints for law enforcement to investigate?

1.4.1 Footprints of Informed Trading

Empirical findings on this topic have already proven controversial. Recent research in the econometric literature supports the hypothesis that institutional investors have advance information about a wide range of unexpected firm-specific events and trade

on such information ahead of the broader market (Hendershott et al, 2015). Illustratively, mutual fund managers who share educational ties with corporate managers overweight and outperform in such investments, with nearly all of the outperformance concentrated around corporate news announcements (Cohen et al, 2008).

1.4.2 Contribution

In our contribution to this subject, we infer firm-specific information flow from the behaviour of equity time series. Pairing statistical testing methodologies with Gaussian Process-based function learning, we find indications of microeconomic NPI leakage in US equity markets, manifesting as a prescient price drift in the 24-hour timeframe ahead of earnings announcements. We hypothesise that this abnormality reflects the abuse of a particular loophole in the design of Regulation Fair Disclosure, which permits a 24-hour delay in the market-wide disclosure of *unintentional* private information leaks.

The anomaly systematically dampens in the aftermath of selective disclosure investigations by the Securities and Exchange Commission (SEC), much the same way insiders avoid opportunistic trading in their own stock whenever the SEC publicly prosecutes insider traders (Cohen et al, 2012). Yet within 24 months of a major investigation, the inferred leakage has resumed. The market's short memory of punitive enforcement actions suggests an insufficiency in the SEC's efforts to deter selective disclosure. This work is presently being reviewed under the title "Short Memories? The Impact of SEC Enforcement on Insider Leakage" for publication in the *Journal of Law, Finance and Accounting*.

Chapter 2

Background

This chapter aims to provide a cogent understanding of the modelling frameworks commonly adopted in the study of financial markets. To be effective, their complexity will need to be adapted to the difficulty of the pursuit (e.g., producing market forecasts). After describing some of the most widespread financial models and the limitations imposed by their parameter spaces, we will relax the assumption of a parametric solution and explore ultra-parametric and non-parametric machine learning approaches for regression and classification problems. These techniques form the bedrock of the research findings in subsequent chapters.

2.1 Model Complexity

Prediction tasks are not all equally complex. Given a car travelling at a constant speed k (in kilometers per hour), the distance d (in kilometers) covered after t hours is readily expressed:

$$d = k \times t \tag{2.1}$$

Simple models fully capture dependencies between variables for a wide range of real-world physical phenomena. Indeed,

- Hooke's Law: Force $= -$ Stiffness $\times$ Displacement (2.2)

- Ohm's Law: Current $= \frac{\text{Voltage}}{\text{Resistance}}$ (2.3)

- Mass-Energy Equivalence: Energy = Mass $\times$ Celerity^2 (2.4)

are all iconic linear models drawn from classical mechanics, electromagnetism and special relativity. In each instance, variables of interest are linearly related to one another through a factor held constant in the model ($-$ Stiffness, Resistance^{-1} and Celerity^2 respectively), leading to models exhibiting linearity in their parameters.

Standard linear regression models are frequently deployed in finance, where they do not reflect rigorous dependencies but rather the hypothesised dynamics of economic variables. As such, these models do not aim to deliver a 'correct' solution like the foregoing physical models, but merely an insightful one.[1]

- Capital Asset Pricing Model (CAPM): (2.5)

  $\mathbb{E}$[Stock Return] - Risk-free Return = $\beta\times$ ($\mathbb{E}$[Market Return] - Risk-free Return)

- Auto-Regressive Moving Average Model (ARMA(p,q)): (2.6)

$$x_t = c + \sum_{i=1}^{p} \phi_i x_{t-i} + \sum_{i=1}^{q} \theta_i \epsilon_{t-i} + \epsilon_t, \quad x_i, \phi_i, \theta_i \in \mathbb{R}, \quad \epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$$

Linear regression models of this type may facilitate our understanding of the world, but they are often just approximations of the dynamics of the underlying phenomenon. The linear approximation may prove woefully inadequate if the true relationship obeys a power law, or exhibits periodicity. Their framework imposes two types of restriction which we will need to relax, in order to build more complex models with greater explanatory power.

1. The linear assumption: linear models are inherently biased by choices made in the data gathering process. The ARMA model (Equation (2.6)) depends

[1] In the immortal words of the statistician George Box, 'all models are wrong but some are useful.'

on observed variables x_i. Had we instead collected the quadratic, cubic or exponential function of x_i instead, the resulting ARMA coefficients $\{\phi_i,\ \theta_i\}$ would likely change - and with them, our explanation of the phenomenon under study. A solution to this involves the inclusion of additional features $k(x_i)$ in the model. Although the relationship between output and inputs may not be linear, linearity may still hold in a different feature space spanned by an appropriate choice of basis functions k.

2. The parametric assumption: linear models assume that a finite set of parameters can adequately capture the relationship between output and inputs. Better (albeit more complex) solutions may be found if we make no assumptions about the functional form of the mapping and allow instead for an infinite dimensional parameter, capable of growing with the volume of data and the required complexity of the solution.

2.2 Parametric Models

We begin by defining widely used parametric models before addressing the notion of overfitting and methods to combat it.

2.2.1 Linear Regression Models

Linear regression models remain some of the most widely used in practical applications, in part due to their simplicity, interpretability and the relative ease of determining the statistical properties of their estimators. Provided we are given n inputs each of dimension k in a design matrix $\mathbf{X}$, along with the n corresponding scalar outputs in a vector $\mathbf{y}$, the linear regression model maps $\mathbf{X}$ to $\mathbf{y}$ linearly through a vector of coefficients $\boldsymbol{\beta}$:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \mathbf{y} \in \mathbb{R}^n, \quad \mathbf{X} \in \mathbb{R}^{n\times k}, \quad \boldsymbol{\beta} \in \mathbb{R}^k, \quad \boldsymbol{\epsilon} \overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}) \tag{2.7}$$

Regression models are primarily deployed for two purposes:

- To predict output values for new, unseen inputs: predictive models are fitted to a training set consisting of input-output pairs. Once the model's parameters have been computed, they can be used to predict outputs for new, unseen input values.
- To determine the explanatory power of the model's features: the coefficients of the linear regression model can be used to identify which input variables are pertinent in predicting the output.

Linear regression models are typically fitted through the Ordinary Least Squares method (OLS), in which coefficients are derived by minimising the sum of the squared residuals $\boldsymbol{\epsilon}$ in the linear regression model defined in Equation (2.7). Formally, we define a loss function $L(\boldsymbol{\beta})$ that captures the mismatch between the observed outputs $\mathbf{y}$ and the model's predictions $\hat{\mathbf{y}}$:

$$\hat{\mathbf{y}} = \mathbf{X}\boldsymbol{\beta}, \quad \hat{\mathbf{y}} \in \mathbb{R}^n, \quad \mathbf{X} \in \mathbb{R}^{n \times k}, \quad \boldsymbol{\beta} \in \mathbb{R}^k \tag{2.8}$$

$$\begin{aligned} L(\boldsymbol{\beta}) &= (\mathbf{y} - \hat{\mathbf{y}})^2 \\ &= \boldsymbol{\epsilon}^2 \\ &= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \end{aligned} \tag{2.9}$$

We minimise $L(\boldsymbol{\beta})$ by differentiating w.r.t. $\boldsymbol{\beta}$ and setting to zero, deriving the OLS coefficients of the linear regression model:

$$\boldsymbol{\beta}_{\text{OLS}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} \tag{2.10}$$

2.2.2 Regularised Linear Regression Models

A common pitfall in statistical modelling arises from the design of models that correspond too closely to a particular dataset, and fail to generalise to future observations. In the context of parametric models, choosing the appropriate number of parameters is a non-trivial task. Too few parameters will result in a model lacking the complexity to learn the data's underlying patterns, a problem termed *underfitting*. Conversely, a model with too many parameters may track its dataset too well, fitting its noise and generalising poorly by virtue of *overfitting*. To consider an extreme case: if the number of parameters matches or exceeds the number of observations in a training set, a parametric model could achieve perfect predictive accuracy by memorising the data in its entirety. Such models typically underperform severely on unseen data, as their parameters have adapted excessively to noise during training.

A common means of overcoming this challenge is through regularisation, a technique that consists of adding a regularisation term to the loss function $L(\boldsymbol{\beta})$. Regularisation imposes Occam's razor on the solution by penalising complex solutions involving extreme coefficients. Formally, the loss function defined in Equation (2.9) is amended to include a regularisation term $\lambda R(\boldsymbol{\beta})$, where λ controls the importance of a regularised solution, and the functional form of $R(\boldsymbol{\beta})$ is typically a ℓ_p norm of the vector of coefficients $\boldsymbol{\beta}$ that enforces desirable characteristics (e.g. simplicity, sparsity) on the solution:

$$L(\boldsymbol{\beta}) = (\mathbf{y} - \hat{\mathbf{y}})^2 + \lambda R(\boldsymbol{\beta}) \tag{2.11}$$

Two of the most popular forms of regularised parametric models, ridge regression and LASSO (Least Absolute Shrinkage and Selection Operator) regression, emerge from specific choices in the functional form of $R(\boldsymbol{\beta})$.

2.2.2.1 Ridge Regression

One of the most common forms of regularisation uses the ℓ_2 norm of $\boldsymbol{\beta}$ as a regularisation term.

$$L(\boldsymbol{\beta}) = (\mathbf{y} - \hat{\mathbf{y}})^2 + \lambda \left\|\boldsymbol{\beta}\right\|_2^2 \tag{2.12}$$

Minimising $L(\boldsymbol{\beta})$ under ridge regression leads to a different set of parameters from Equation (2.10):

$$\boldsymbol{\beta}_{\text{ridge}} = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y} \tag{2.13}$$

This formulation of the loss function penalises large values for the coefficients in $\boldsymbol{\beta}$, favouring simpler solutions. A further benefit of ridge regression arises from the challenges posed by an ill-conditioned design matrix $\mathbf{X}$. If $\mathbf{X}$ is rank deficient (for example due to perfect multicollinearity in the data), then $\mathbf{X}^\top \mathbf{X}$ will not be invertible. The addition of a scalar value λ to the diagonal of $\mathbf{X}^\top \mathbf{X}$ prior to inversion, per Equation (2.13), overcomes this problem.

2.2.2.2 LASSO Regression

A widespread alternative to ridge regression uses instead the ℓ_1 norm of $\boldsymbol{\beta}$ for regularisation.

$$L(\boldsymbol{\beta}) = (\mathbf{y} - \hat{\mathbf{y}})^2 + \lambda |\boldsymbol{\beta}| \tag{2.14}$$

This approach, known as LASSO regression, differs from ridge regression by pushing solutions to reside on a simplex (Figure 2.1), facilitating the pruning of a dataset's least relevant features. LASSO induces sparsity in parametric solutions and provides one of the most popular methods for principled feature selection in machine learning. Incremental increases in the value of the regularisation parameter λ generate the LASSO path, corresponding to a sequence of increasingly sparse solutions for $\boldsymbol{\beta}_{\text{lasso}}$.

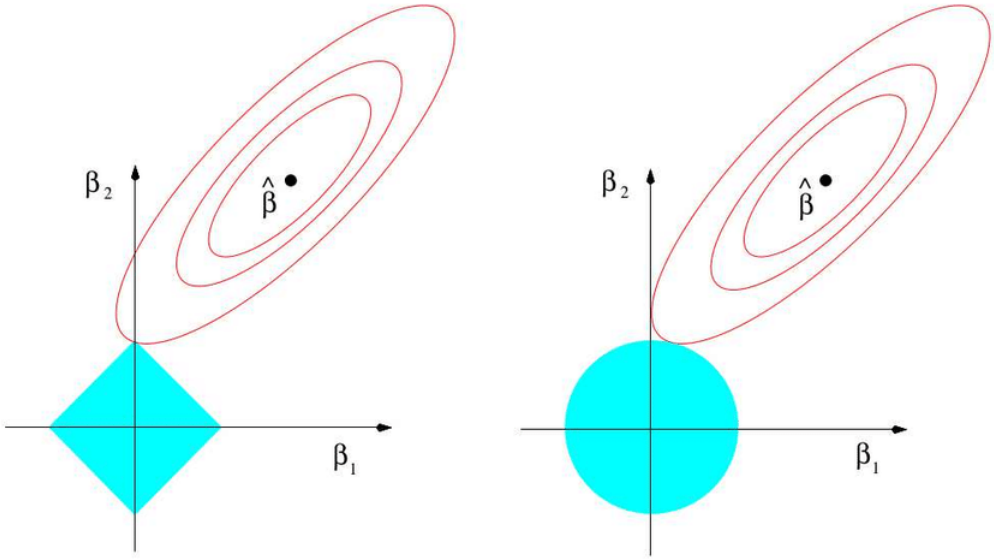

Figure 2.1: Comparison of the loss contours generated by ℓ_1 norm (left) and ℓ_2 norm (right) regularisation. The ℓ_1 norm regularisation of LASSO regression encourages solutions that reside on a simplex, with some coefficients (β_1 in this case) being forced to zero. Image from *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Hastie et al (2009).

Unlike ridge regression, there is no closed form solution for $\boldsymbol{\beta}_{\text{lasso}}$. Though originally solved via quadratic programming techniques from convex optimisation, algorithms have been devised to compute the LASSO path efficiently, e.g. forward stagewise and least angle regression (Efron et al, 2004).

2.3 Ultra-Parametric Models

Linear regression is conceptually simple. Given a dataset with input features $\mathbf{X}$ and target variable $\mathbf{y}$, we minimise a loss function (typically, the Residual Sum of Squares as in Equation (2.9)) to find the parameters that optimally map a linear combination of the input variables to the output variable. Neural networks are a logical evolution of the concept: they rely on linear regression as building blocks of computation, gated by non-linear transformations termed *activation functions* to enable the discovery of

more complex relationships in data.

2.3.1 Introduction to Neural Networks

Over the last decade, neural networks have risen dramatically in popularity, propelled by the success of deep learning in a wide range of practical applications. Formally, neural networks map inputs to outputs through a collection of non-linear computation nodes, called *neurons*, stacked into *hidden layers*. Inputs and outputs are connected by potentially many such hidden layers, leading to so-called deep learning architectures. Neural networks straddle the boundary between parametric and non-parametric models. We opt to interpret them as an ultra-parametric extension of linear regression models, wherein each neuron computes a weighted linear combination of its inputs (as per the parametric regression models of Section 2.2), applies an activation function to the newfound value and forwards its output to the next layer's neurons.

2.3.2 Universal Approximation via Neural Networks

The effectiveness of neural networks finds its theoretical foundation in the *universal approximation theorem*, which states that a feed-forward[2] network with a single hidden layer comprised of a sufficiently large (but finite) number of neurons can approximate any continuous function on compact subsets of $\mathbb{R}^n$, given an appropriate choice of activation function (Cybenko, 1989).

Universal approximation theorem. *Let $\phi(\cdot)$ be a non-constant, bounded and continuous function. Let I_n denote the n-dimensional unit hypercube $[0,1]^n$, and $C(I_n)$ denote the space of continuous functions on I_n. Then, given any $\epsilon > 0$ and any function $f \in C(I_n)$, there exists an integer K, real constants a_i, $b_i \in \mathbb{R}$ and real vectors $\mathbf{w_i} \in \mathbb{R}^n$, where $i = 1, \ldots, K$, such that we may define:*

[2]Feed-forward neural networks are neural networks in which the connections between nodes do not form a cycle. Information moves in a single direction, forward through each layer sequentially, from the input layer to the output layer via potentially several intermediate hidden layers.

$$F(\mathbf{x}) = \sum_{i=1}^{K} a_i \phi(\mathbf{w_i}^\top \mathbf{x} + b_i) \tag{2.15}$$

as an approximate realisation of f. That is,

$$|F(\mathbf{x}) - f(\mathbf{x})| < \epsilon \tag{2.16}$$

holds for all $\mathbf{x}$ *in* I_n.

2.3.3 Perceptron

The perceptron is a simple neural network comprised of a single hidden layer with one neuron between its input and output layer (Figure 2.2). As such, it is very similar to the earlier parametric models:

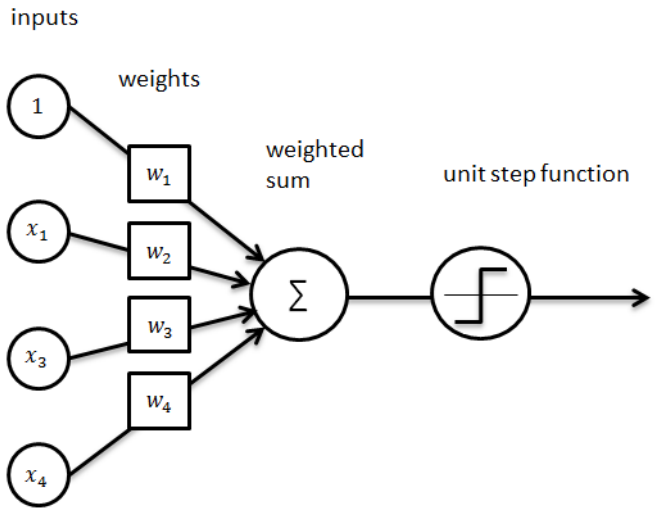

Figure 2.2: The perceptron is a simple neural network comprised of a single hidden layer, joining the input layer to the output layer through a weighted linear combination of the inputs followed by a Heaviside activation function.

- It constructs a single linear combination of the model's inputs: the sole neuron of the perceptron learns a set of weights with which to linearly combine the

original features, and can be thought of as a single, constituent parametric model.

- It applies a non-linear, Heaviside transformation $H(o)$ to the hidden layer's output o, enabling the perceptron to learn a non-linear mapping between the input and output layer.

$$H(o) = \begin{cases} 1, & \text{if } o > 0. \\ 0, & \text{otherwise.} \end{cases} \tag{2.17}$$

2.3.4 Multilayer Perceptron

The benefit of activation functions is particularly pronounced as architectures are extended in depth: without non-linearity, additional layers would not confer any incremental value, as the linear combinations of linear combinations would themselves just be linear combinations with different weights. In other words, multiple hidden layers without non-linear transformations would be equivalent to a single hidden layer with appropriately chosen weights. The inclusion of activation functions between the hidden layers allows neural networks to learn more complex functional mappings than the linear models of Section 2.2. The multilayer perceptron (MLP) harnesses this potential, by including multiple layers between input and output and allowing each layer to possess many neurons (Figure 2.3). In the case of MLPs, the activation functions employed are commonly the hyperbolic tangent function $\tanh(o)$, logistic function $\sigma(o)$ and rectified linear unit $\text{ReLU}(o)$.

$$\tanh(o) = \frac{\sinh(o)}{\cosh(o)} = \frac{\mathrm{e}^{o} - \mathrm{e}^{-o}}{\mathrm{e}^{o} + \mathrm{e}^{-o}} \tag{2.18}$$

$$\sigma(o) = \frac{1}{1 + \mathrm{e}^{-o}} \tag{2.19}$$

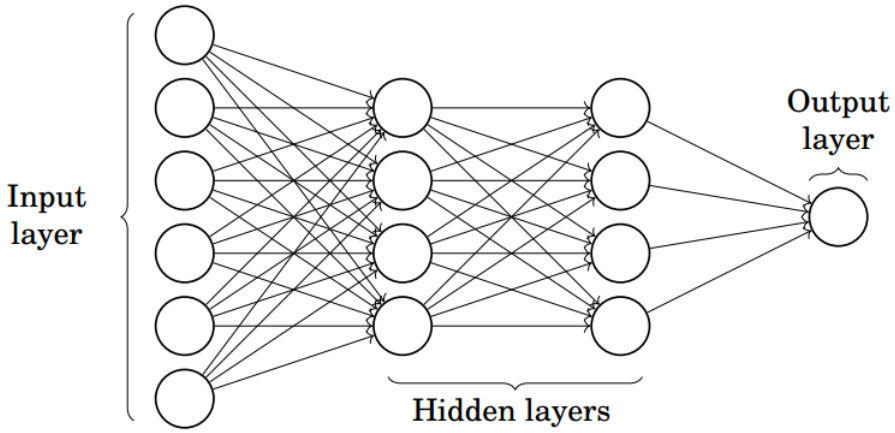

Figure 2.3: A multilayer perceptron with 2 hidden layers. Each neuron within a layer computes a linear combination of its inputs followed by a non-linear transformation, much like the perceptron.

$$\text{ReLU}(o) = \max(0, o) \tag{2.20}$$

The weights in each neuron of the multilayer perceptron are learned through backpropagation, an algorithm whereby the mismatch between a model's predictions and actual observations is distributed back to the weights of each neuron, layer by layer.

2.3.5 Convolutional Neural Network

Convolutional neural networks extend multilayer perceptrons, by adding one or several additional layers at the beginning of the architecture. These layers, termed *convolutional layers*, consist of a set of learned filters. These filters are typically much smaller than the input, and measure local similarity (calculated by sliding dot or Hadamard product). The output of a convolutional layer is a feature map, identifying regions where the input to the layer was similar to the learned filter. In effect, convolution functions as bespoke feature extractors for neural network architectures, enabling in the process vastly superior model performance (Figure 2.4).

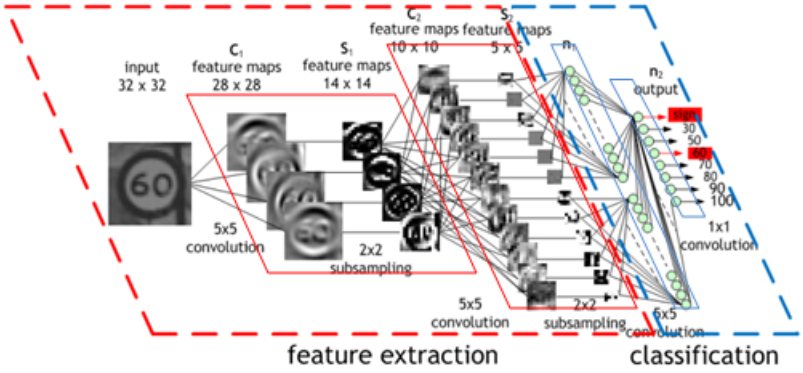

Figure 2.4: A convolutional neural network with 2 convolutional layers. The red hashed outline contains the feature-extractive convolutional layers, and the blue hashed outline is effectively a single layer perceptron. In this architecture, convolutional outputs over a local area are reduced to a single value via *subsampling* or *pooling*, an operation designed to improve the model's memory footprint and invariance to translations/rotations. Image from the Nvidia webpage on Convolutional Neural Networks (https://developer.nvidia.com/discover/convolutional-neural-network).

2.4 Non-Parametric Models

A significant shortcoming of parametric models arises from the need to define in advance the desired complexity of the model, through the choice of which features $\mathbf{x}$ (and potentially, transformations $\phi(\mathbf{x})$ thereof) to combine linearly. Ultra-parametric models are not immune to this concern: the choice of layer size and number of layers directly influences the neural network's learning prowess. Though regularisation techniques help partially address this issue, other non-parametric techniques overcome the problem by instead allowing data itself to govern model complexity.

2.4.1 Kernel Density Estimation

Kernel density estimation (KDE) is a simple non-parametric method for estimating a random variable's probability density function. Formally, given a univariate independent and identically distributed sample $(x_1, x_2, ..., x_n)$ drawn from a distribution with unknown density f, the kernel density estimator of f is given by:

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{2.21}$$

where K is the estimator's *kernel*, a non-negative function of x, and $h > 0$ is a smoothing parameter called the *bandwidth*. Though similar to histograms, kernel density estimates can be assigned properties such as smoothness or continuity through an appropriate choice of kernel K.

2.4.2 Introduction to Gaussian Processes

Gaussian Processes provide a rich and flexible framework for non-parametric modelling. We outline the fundamentals of Gaussian Processes for regression; for a more comprehensive treatment of Gaussian Processes, we refer the interested reader to Rasmussen and Williams (2006).

A Gaussian Process is a collection of random variables, any finite subset of which has a joint Gaussian distribution. Gaussian Processes are fully parametrised by a mean function and covariance function, or kernel. We write a Gaussian Process $f(\mathbf{x})$ as:

$$f(\mathbf{x}) \sim \ \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')) \tag{2.22}$$

where functions $m(\mathbf{x})$ and $k(\mathbf{x},\mathbf{x}')$ are respectively the mean and covariance functions:

$$m(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})] \tag{2.23}$$

$$k(\mathbf{x}, \mathbf{x}') = \mathbb{E}[(f(\mathbf{x}) - m(\mathbf{x})) \times (f(\mathbf{x}') - m(\mathbf{x}'))] \tag{2.24}$$

Inputs are commonly centered during pre-processing, implying the process has zero mean. For a given training set $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$ with corresponding output variables $\mathbf{y} = \{y_1, ..., y_n\}^\top$ and Gaussian Process f, the distribution of $\mathbf{f} = [f(\mathbf{x}_1), ..., f(\mathbf{x}_n)]^\top$ will be multivariate Gaussian:

$$\mathbf{f} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}) \tag{2.25}$$

where $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$. Conditional on $\mathbf{f}$, we have a Gaussian observation model given by:

$$y_i | f(\mathbf{x}_i) \sim \mathcal{N}(\mathbf{0}, \sigma_n^2) \tag{2.26}$$

where σ_n^2 parametrises noise. Gaussian distribution conjugacy allows us to marginalise out $\mathbf{f}$ to find the distribution:

$$y_i \sim \mathcal{N}(\mathbf{0}, \mathbf{K} + \sigma_n^2 \mathbf{I}) \tag{2.27}$$

and conditioning on the training data yields the following predictive distribution y^* for an unseen test datapoint $\mathbf{x}^*$:

$$y^* | \mathbf{x}^*, \mathbf{X}, \mathbf{y} \sim \mathcal{N}(\mathbf{k}^* (\mathbf{K} + \sigma_n^2 \mathbf{I})^{-1} \mathbf{y}, k^{**} - \mathbf{k}^* (\mathbf{K} + \sigma_n^2 \mathbf{I})^{-1} \mathbf{k}^{*\top}) \tag{2.28}$$

where $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$, $\mathbf{k}^* = [k(\mathbf{x}_1, \mathbf{x}^*), ..., k(\mathbf{x}_n, \mathbf{x}^*)]$ and $k^{**} = k(\mathbf{x}^*, \mathbf{x}^*)$. This methodology combines prior knowledge over $\mathbf{f}$, encoded in the covariance function $k(\mathbf{x}, \mathbf{x}')$, with observation data to produce a posterior distribution for forecasting.

2.4.3 Universal Approximation via Gaussian Processes

As was the case with the neural networks of Section 2.3, Gaussian Processes can be applied to approximate any function on compact subsets of $\mathbb{R}^n$ given an appropriate choice of kernel (Micchelli et al, 2006), for the purposes of both regression and classification.

Formally, let X be a compact subset of $\mathbb{R}^n$, let $k(\cdot,\cdot)$ be a continuous covariance function defined on $X \times X$, and let $C(X)$ be the space of continuous functions on X. Then any function $f \in C(X)$ can be approximated to a tolerance ϵ by a function F in

the Reproducing Kernel Hilbert Space $K(X)$ associated to k.[3] For kernels possessing the universal approximation property (e.g. the Squared Exponential kernel defined in Section 2.4.5), $K(X)$ is dense in $C(X)$, meaning that for any $f \in C(X)$ and for any tolerance ϵ there is an F in $K(X)$ such that:

$$|F(\mathbf{x}) - f(\mathbf{x})| < \epsilon \tag{2.29}$$

for all $\mathbf{x}$ in X.

2.4.4 Regularisation for Gaussian Processes

Generalisation is an important consideration in non-parametric regression, just as it was for the parametric models of Section 2.2. In the context of Gaussian Processes, overfitting can still occur as a result of learning hyperparameters that are too closely fitted to a training set. One approach to regularisation commonly used for GP regression is k-fold cross-validation, a model validation methodology that involves partitioning the original training set into k complementary subsets. We train the model on *k-1* subsets and test it on the one remaining subset. After rotating through the k choices for this validation set, the results are then averaged across all tests and provide insight into the model's ability to generalise well. Cross-validation provides the means of identifying (or even constructing) the appropriate kernel for a given problem domain.

2.4.5 Automatic Relevance Determination Kernels

Covariance functions typically employ an isotropic norm as the similarity measure between two vectors in input space. The Squared Exponential kernel, arguably the most widely used, parametrises the covariance between two inputs $\mathbf{x}$ and $\mathbf{x}'$ as follows:

[3] A Reproducing Kernel Hilbert Space $K(X)$ is the closure of the vector space formed by all possible finite linear combinations of functions $f_{\mathbf{y}}(\mathbf{x}) = k(\mathbf{x},\mathbf{y})$ where $\mathbf{y} \in X$.

$$k(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\left[-\frac{(\mathbf{x} - \mathbf{x}')^2}{2l^2} \right] \tag{2.30}$$

The Squared Exponential kernel possesses 2 hyperparameters:

- An input length scale l that determines the speed at which changes in input space reflect on the output.
- An output length scale σ_f that acts as a scaling factor.

Hyperparameter optimisation for Gaussian Processes amounts to finding the values for hyperparameters $\{l,\ \sigma_f,\ \sigma_n\}$ that maximise the log-likelihood (in effect, the 'fit') of the training dataset. Formally, given a dataset comprised of inputs $\mathbf{X} \in \mathbb{R}^{n \times k}$ and outputs $\mathbf{y} \in \mathbb{R}^n$, we maximise the log-likelihood function[4] defined as:

$$\log p(\mathbf{y}|\mathbf{X}) = -\frac{1}{2}\mathbf{y}^\top (K + \sigma_n^2 I)^{-1}\mathbf{y} - \frac{1}{2}\log|K + \sigma_n^2 I| - \frac{n}{2}\log 2\pi \tag{2.31}$$

where $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$ and σ_n parametrises noise.

The kernel described by Equation (2.30) assumes that a single, global characteristic length scale l can appropriately evaluate proximity in all input dimensions. Even with all inputs normalised to the same scale during pre-processing, it is not uncommon for a model's covariates to contain varying levels of information on the response, motivating the use instead of input-specific characteristic length scales.

In Automatic Relevance Determination (ARD) kernels, the scalar input length scale l of Equation (2.30) is replaced with a diagonal hyperparameter covariance matrix over inputs, with different matrix entries l_i for each input dimension i allowing for variable distance measures. For example, the ARD Squared Exponential kernel for a D-dimensional input differs from the original isotropic Squared Exponential kernel of Equation (2.30) as follows:

[4]Equivalently, this is often described as minimising the negative log-likelihood function.

$$k(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\left[-\frac{1}{2} \sum_{i=1}^{D} \left(\frac{x_i - x_i'}{l_i} \right)^2 \right] \tag{2.32}$$

The hyperparameters l_i will adapt to any given dataset: inputs with large length scales cause only marginal variations in the covariance function, whereas inputs with small length scales effectively magnify those variations.

2.4.6 Ranking Relevance in ARD Kernels

ARD algorithms have been successfully used in research ranging from bioinformatics (Campbell and Tipping, 2002) to seismography (Oh et al., 2008), providing an effective tool for pruning large numbers of irrelevant features. A limitation of the methodology as presented is that the input length scale vector only provides a relative ranking between the features of a model. Two equally meaningless inputs will have length scales of similar magnitude, as would two equally meaningful features. On their own, these values provide little basis for performing dimensionality reduction. To overcome this, we introduce the notions of *Relevance Score* and *Relevance Ratio*. We define the relevance score of each feature to be the reciprocal of its input length scale, and rank the salience of inputs by descending relevance.

$$\text{Relevance Score}_i = l_i^{-1} \tag{2.33}$$

In each regression, we include a baseline feature composed of standard Gaussian noise. We assert that a meaningful input should have a relevance score that is at least two orders of magnitude greater than noise. By computing the ratio between the relevance scores of a feature and Gaussian noise, we can determine which features are objectively informative.

$$\text{Relevance Ratio}_i = \frac{\text{Relevance Score}_i}{\text{Relevance Score}_{noise}} \tag{2.34}$$

This forms the basis for a novel approach to hypothesis testing tailored to the Gaussian Process framework. Relevance ratios objectively determine a feature's relevance while still permitting for that salience to manifest non-parametrically, unlike the regularised linear regression models of Section 2.2.2.

2.5 Significance Testing

In addition to the range of parametric and non-parametric models outlined in Sections 2.2-2.4, we also rely on several widely utilised statistical tests for identifying significant relationships or changes in data. We provide a brief summary of these methodologies next.

2.5.1 Correlation Analysis

In determining whether a particular covariate exhibits a strong relationship with a target variable, our first step will often be to examine the correlation between the two. Standard t-tests are employed to evaluate whether an observed sample correlation is significant. Given two equally-sized, standardised independent random variables $\mathbf{x}$ and $\mathbf{y}$ of length N with sample correlation r, the statistic

$$t = \frac{r \times \sqrt{N-2}}{\sqrt{1-r^2}} \tag{2.35}$$

is t-distributed with *N-2* degrees of freedom. Values for the *(r,N)* pair that land outside the 95% confidence interval of the t-distribution violate the null hypothesis of independence, providing a methodology via the Student's t-test for identifying significant correlations in a dataset. p-values are derived from t-distribution tables and measure the probability that uncorrelated sample data will yield a t-statistic as or more extreme than the value of t obtained from Equation (2.35). Common

significance thresholds in applied statistics are p-values of 0.05 or 0.01, but in the case of very large datasets we will look for p-values below 0.001.

2.5.2 Kolmogorov-Smirnov Two-Sample Tests

We may also wish to assess whether a conditional subset of our data exhibits a meaningfully different distribution from the unconditional distribution of the full, original dataset. Kolmogorov-Smirnov two-sample tests allow us to determine whether conditioning on a precise criterion results in a statistically significant change in the data distribution. The two-sample Kolmogorov-Smirnov (K-S) test (Massey, 1951) evaluates the null hypothesis that the distributions generating both samples (conditional and unconditional) have identical cumulative distribution functions, by evaluating the K-S statistic:

$$\gamma_N = \left(\frac{n_1 n_2}{n_1 + n_2} \right)^{1/2} \sup_{-\infty < z < \infty} |F_N(z) - F(z)| \tag{2.36}$$

The limiting distribution of γ_N provides percentile thresholds above which we reject the null hypothesis. When this occurs, we may infer that the conditioning criterion significantly impacts the distribution, and therefore warrants further attention.

Chapter 3

Heterogeneous Data Fusion

3.1 Introduction

One of the central challenges in financial forecasting is determining where to look. A financial instrument's time series history, comparables and derivatives, news articles and opinion pieces all have the potential to influence price evolution. Developing a robust framework for knowledge extraction from disparate, jointly informative datasets remains a major challenge for finance and machine learning researchers.

In this chapter, we attempt to forecast daily returns on the S&P500 index, a broad market benchmark for US equities commonly viewed as a gauge of financial stability. The S&P500 is a market capitalisation-weighted index of the 500 largest corporations in the US, covering the full range of technology, consumer goods, utilities and financial services companies. It is one of the most visible benchmarks in the world, actively traded by buy-and-hold mutual funds and high-frequency hedge funds alike.

We begin by postulating four broad categories in which to search for salient explanatory variables (Section 3.2). *Market technicals* include lagged returns to measure autocorrelation, as well as chartist signals used in industry like the Moving Average Convergence Divergence (MACD). *Sentiment analysis* covers the impact of newsflow, measured by optimism or pessimism in social media. *Options market metrics* provide a glimpse into the positioning of market experts and give us a principled, data-driven method for modelling price space as an inhomogeneous dimension with regions of

directional bias and return compression. *Broker recommendations* collate the wisdom of equity analysts and allow us to measure the predictive value, if any, of their upgrades and downgrades.

We show that predictive performance improves when combining signals from each domain, and provide a principled framework for the triage of inputs by implementing Automatic Relevance Determination (ARD) in the covariance parametrisation of an adaptive Gaussian Process model (Section 3.3). The ranking that emerges from this analysis defies expectations, and encourages further investigation of options markets and the price space representation.

3.2 Data

We detail the features considered for each of the four domains under consideration, all of which will be used to predict *Return(t+1)*, the next-day log-return on the S&P500. Details on the sourcing of each dataset are provided in Appendix A.1.

3.2.1 Technical Indicators

Technical analysis was one of the earliest forms of financial forecasting, first appearing in merchant accounts of the Dutch markets in the 17th century. Formalised as a discipline in the 1940s (Edwards and Magee, 1946), it involves the use of price and volume time series to make directional forecasts. It has been extensively studied in seminal regression analyses (Lo et al., 2000) demonstrating the incremental gains in predictive performance provided by identifying specific patterns in price history. More sophisticated, non-parametric modelling of technical analysis soon followed: technicals-driven Gaussian Process regression has been applied to forecasting time-series in a wide range of asset classes, including stock market prices (Farrell and Correa, 2007), stock market volatility (Ou and Wang, 2009) and commodity spreads (Chapados and Bengio, 2007).

Market technicals are metrics derived directly from the price history p(t) of a financial instrument. We consider four features commonly watched in industry (Taylor and Allen, 1992): the *previous daily log-return* on the S&P500, its *50-day Simple Moving Average*, as well as the *Moving Average Convergence Divergence (MACD)* and *Signal Line*, constructed from Exponential Moving Averages (EMA) of the time series as follows:

$$\text{MACD(t)} = \text{12-day EMA}\big[\text{p(t)}\big] - \text{26-day EMA}\big[\text{p(t)}\big] \tag{3.1}$$

$$\text{Signal Line(t)} = \text{MACD(t)} - \text{9-day EMA}\big[\text{MACD(t)}\big] \tag{3.2}$$

We do not believe that the formulation of these metrics is inherently meaningful, but rather that standardised definitions provide precise, measurable thresholds at which chartist market participants will react. Including these features will allow our model to identify those thresholds and thereby anticipate technically-led order flow.

3.2.2 Sentiment Analysis

Literature on financial prediction using text data has proliferated in recent decades, closely tracking advances in the field of natural language processing. The methodology in this domain has typically involved converting words or phrases into numerical gauges of sentiment with which to predict stock market direction (Nikfarjam et al., 2010). Modelling techniques have ranged from simple Naive-Bayes or Support Vector Machine classifiers to more advanced algorithms built on deep learning. More recent work in sentiment composition has sought to predict economic indicators like the U.S. Non-Farm Payrolls using newsflow data. These studies show evidence that accurate parsing of news articles can produce state-of-the-art forecasts for market-moving announcements (Levenberg et al., 2013, 2014).

While factual newsflow is significant, it is specifically the polarity of its interpretation by markets - as beats or disappointments - that drives market movement. Market sentiment was captured using indicators derived from both Twitter and Stocktwits, a social media site dedicated to real-time discussions of financial markets and actively frequented by S&P500 retail investors. Sentiment data time series were drawn from a publicly available repository produced by PsychSignal, a data analytics firm focused on parsing social media text in the context of stock prices. In addition to their sentiment 'state' variables, we derived two further metrics by tracking the daily changes in the sentiment indices to capture the potential impact of momentum shifts in opinion data.

3.2.3 Options-based Modelling of Price Space

Research on the interactions between stock and options market prices has been scarce, though early attempts were made to assess correlation in the volume data. Studies indicated that call options flow leads underlying shares flow with a one-day lag, lending credence to the hypothesis of a sequential flow of information between the options and stock markets (Anthony, 1998). As a province reserved for more sophisticated traders, options market open interest volumes offer a window into the expectations of the most experienced, well-capitalised participants. As strike-sensitive instruments[1], options data also allow us to gauge how these expectations vary at different price levels, motivating the representation of price as an inhomogeneous space with identifiable regions of high directional bias or variance. Illustratively, high open interest (OI) in call options coupled with low open interest in put options indicates experts pre-positioning for a rally. By contrast, high open interest in straddles[2] at a given

[1]Call and put option prices are calculated via the Black-Scholes formula and depend on a 'strike' level that defines the price at which the option owner may buy or sell the underlying asset.

[2]A long straddle position refers to the ownership of both a call and a put option at the same strike price and expiry date: it does not express a directional view, and benefits so long as the underlying asset deviates sufficiently from the strike before expiry.

strike implies low consensus among experts about *directionality* at that price, and hints at evenly matched, competing forces that will compress returns locally. We term this phenomenon *viscosity*, appealing to the visual analogy of price space as an inhomogeneous fluid that enables price gaps in regions of low viscosity and prevents it in regions of high viscosity.

To capture the directionality and viscosity implied by open interest data, we constructed two metrics. Directionality measures the daily change in call minus put open interest at strike s with time-to-expiry τ, summed across all strikes S and expiries T. It proxies for expert optimism as evidenced by bullish option positioning, and by construction correlates positively with S&P500 next-day returns. The scaling factors $\exp(-\gamma_D \tau)$ account for the time sensitivity of options traders, and serve to scale up the weight of nearby expiries by mimmicking the exponential decay of gamma risk as time-to-expiry lengthens.

$$\begin{aligned}\text{Directionality(t)} = &\sum_{s\in S,\tau\in T}\Big[(\text{OI(s,t,}\tau\text{)}_{Call} - \text{OI(s,t,}\tau\text{)}_{Put}) \times \exp\left(-\gamma_D\tau\right)\Big] \\ &- \sum_{s\in S,\tau\in T}\Big[(\text{OI(s,t-1,}\tau\text{)}_{Call} - \text{OI(s,t-1,}\tau\text{)}_{Put}) \times \exp\left(-\gamma_D\tau\right)\Big] \quad (3.3)\end{aligned}$$

The parameter γ_D measures the rate at which directionality decays as a function of time-to-expiry, and is optimised over the training data by solving:

$$\gamma_D = \arg\max_{\gamma_D} \text{corr}\big(\text{Directionality, Return}\big) \quad (3.4)$$

where $\text{corr}(\cdot,\cdot)$ is the linear correlation between its two arguments.

In parametrising viscosity, we make three modelling assumptions. Firstly, the pinning effect of high straddle open interest is at its greatest for options very near their expiry date. Secondly, this effect decays as live prices move away from the straddle's strike s. Thirdly, we claim that open interest volumes follow a log-normal

distribution, evolving over time through the compounding of normally-distributed exponential factors and restricted to non-negative values.[3] These claims jointly motivate the following representation:

$$\text{Viscosity(t)} = \sum_{s\in S,\tau\in T}\Big[\exp\left(-\lambda_V|\text{price(t)-s}|\right)\times\exp\left(-\gamma_V\tau\right) \times\log[\min(\text{OI(s,t,}\tau)_{Call},\text{OI(s,t,}\tau)_{Put})+1]\Big] \quad (3.5)$$

We expect a significant negative correlation between viscosity and the magnitude of S&P500 next-day returns; as such tuning λ_V and γ_V equates to solving the following optimisation problem:

$$\lambda_V,\gamma_V = \arg\min_{\lambda_V,\gamma_V}\text{corr}\big(\text{Viscosity},|\text{Return}|\big) \quad (3.6)$$

3.2.4 Broker Recommendations

Multiple studies have been conducted to ascertain the influence of buy and sell recommendations on stock prices. Research on equity analyst reports shows a significant, systematic but asymmetric drift in the aftermath of broker actions, with short-lived, modest gains following upgrades but durable, material sell-offs following downgrades (Womack, 1996). The magnitude of these changes depended not only on the action (upgrade vs. downgrade), but also on the reputation of the analyst, the size of their brokerage firm and the size of the recommended firm (Stickel, 1995).

Market analysts issue recommendations on individual stocks rather than on the broad market - partly a reflection of the incentive structure for brokerage firms: commissions are substantially larger for actively managed portfolios than for passive

[3]Like many natural growth processes, open interest volumes are driven by the compounding of many small percentage changes, which become additive on a log scale. Per the Central Limit Theorem, the distribution of their sum is approximately normal. Back-transformed onto the original scale, this yields an approximately log-normal distribution. This multiplicative version of the Central Limit Theorem is also known as Gibrat's law.

index-trackers. To overcome this, we construct an index of broker opinions, based on a weighted sum of broker recommendations across the top 100 stocks in the S&P500. These account for 63% of the index's market capitalisation, and broker actions on these household names have a disproportionate effect on the index as a whole. Two indices were built from these weighted sums, to track both changes in analyst opinion (upgrades and downgrades) and the consensus state (buy, hold or sell).

3.3 Results

In this section we outline the findings of our analysis. We begin by discovering relevance hierarchies in the data using ARD, before proceeding with model testing and benchmarking. Model performance metrics were derived using market data from Jan-13 to Dec-14 for training and Jan-15 to Apr-15 for testing. The price history of the S&P500 Index for this period is provided in Figure 3.1.

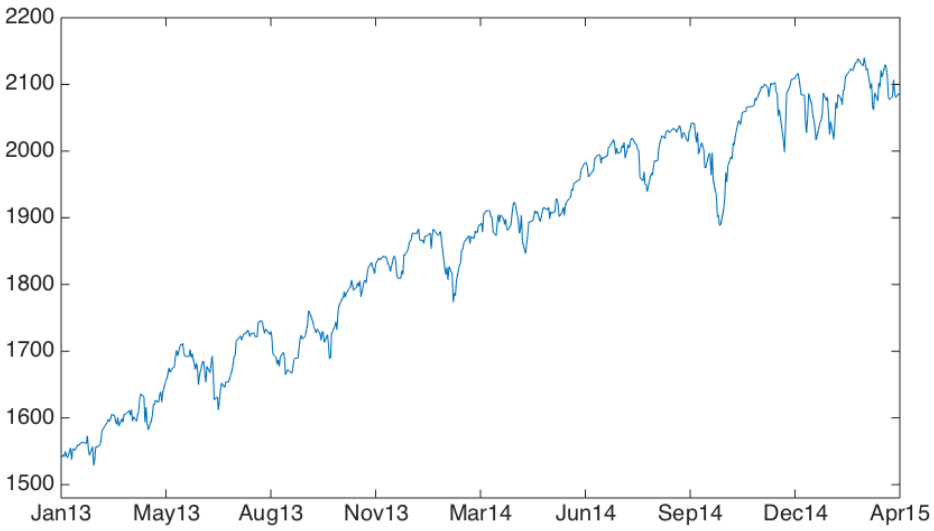

Figure 3.1: S&P500 Index price history between Jan-13 and Apr-15.

Table 3.1: Input-Output Correlation Analysis measured on the training set, N=503 (Jan-13 to Dec-14). We highlight in bold the input features exhibiting Pearson p-values below 0.05.

	Correlation		p-value	
Feature	Pearson	Spearman	Pearson	Spearman
Return(t)	−0.0336	−0.0862	0.4524	0.0534
50dSMA	−0.0451	−0.1123	0.3130	0.0117
MACD	**−0.1403**	**−0.1576**	**0.0016**	**0.0004**
Signal Line	−0.0170	−0.0365	0.7034	0.4138
Stocktwits	**−0.1103**	**−0.1247**	**0.0133**	**0.0051**
Twitter	−0.0287	−0.0539	0.5201	0.2275
Stocktwits Change	−0.0581	−0.0658	0.1933	0.1406
Twitter Change	+0.0269	+0.0215	0.5474	0.6305
Directionality	**+0.1011**	**+0.1135**	**0.0234**	**0.0108**
Viscosity*	**−0.2262**	**−0.1831**	**0.0001**	**0.0001**
Broker State	+0.0348	+0.0159	0.4361	0.7220
Broker Change	+0.0024	+0.0263	0.9564	0.5562

* Correlation for Viscosity measures the correlation between the Viscosity variable and absolute next-day log-returns. As explained in Section 3.2.3, the dampening effect of Viscosity is non-directional, and therefore requires the adoption of absolute log-returns. Correlation for all other features measures the correlation between that feature and next-day log-returns.

3.3.1 Correlation Analysis

We begin by running a correlation analysis on each feature of the training set, grouped by domain and collect the findings in Table 3.1. In most cases, rank correlations are stronger than linear correlations, though the variations are too marginal to alter the analysis. For brevity, in all ensuing sections we have adopted the linear definition of correlation. Applying t-tests to our dataset, every domain apart from broker recommendations held at least one feature bearing significant correlation with next-day returns, signalled by p-values under 0.05 in Table 3.1.

The use of 4 distinct domains stemmed from the belief that, by virtue of tracking different market agents, these datasets will exhibit low correlation with each other and

therefore enhance the predictive power of a combined model. In Table 3.2 we measure the correlation between input pairs in the training set, and find indeed that *intra-domain* correlations are generally stronger than *inter-domain* correlations, inspiring the pursuit of information gain across diverse, heterogeneous datasets.

Table 3.2: Input-Input Correlation Analysis measured on the training set, N=503 (Jan-13 to Dec-14).

	Technicals				Sentiment		Price Space		Broker	
	yret	50dMA	MACD	SL	Twtr	ST	Dir	Visc	State	Change
yret	1.00	0.34	-0.03	0.18	0.29	0.29	0.02	0.01	0.05	0.01
50dMA	0.34	1.00	0.13	0.08	0.11	0.12	-0.11	0.02	0.05	0.00
MACD	-0.03	0.13	1.00	0.49	0.11	0.24	-0.49	0.37	-0.01	-0.15
Signal	0.18	0.08	0.49	1.00	0.27	0.27	-0.18	0.18	0.10	-0.12
Twtr	0.29	0.11	0.11	0.27	1.00	0.44	-0.14	0.21	0.15	-0.01
ST	0.29	0.12	0.24	0.27	0.44	1.00	-0.18	0.11	0.05	0.02
Dir	0.02	-0.11	-0.49	-0.18	-0.14	-0.18	1.00	-0.35	-0.07	0.03
Visc	0.01	0.02	0.37	0.18	0.21	0.11	-0.35	1.00	0.02	-0.10
State	0.05	0.05	-0.01	0.10	0.15	0.05	-0.07	0.02	1.00	0.11
Change	0.01	0.00	-0.15	-0.12	-0.01	0.02	0.03	-0.10	0.11	1.00

3.3.2 Feature Relevance

Through 10-fold cross-validation, we determine the optimal covariance function for our dataset from a range of options (Squared Exponential, Rational Quadratic, Matérn 1/2, Matérn 3/2 and Matérn 5/2; Rasmussen and Williams, 2006), and settle on the Matérn 3/2 kernel, a once-differentiable function exhibiting the low smoothness typical of financial time series.

$$k(\mathbf{x}, \mathbf{x}') = \sigma_f^2\Big(1 + \frac{\sqrt{3}|\mathbf{x} - \mathbf{x}'|}{l}\Big) \times \exp\Big(-\frac{\sqrt{3}|\mathbf{x} - \mathbf{x}'|}{l}\Big) \tag{3.7}$$

Using training data from Jan-13 to Dec-14, we implement separate Gaussian Process regressions for each data domain. Per Section 2.4.5, this amounts to finding the hyperparameters $\{l, \sigma_f, \sigma_n\}$ that maximise the log-likelihood of the training set:

$$\log p(\mathbf{y}|\mathbf{X}) = -\frac{1}{2}\mathbf{y}^\top (K + \sigma_n^2 I)^{-1}\mathbf{y} - \frac{1}{2}\log|K + \sigma_n^2 I| - \frac{n}{2}\log 2\pi \tag{3.8}$$

where $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$ and σ_n parametrises noise.

This allows both a ranking of feature relevance within each domain, and bivariate visualisations of the mean surfaces learned from the two top-ranked features in each model. The heatmaps in this chapter, and all subsequent chapters of this thesis, depict the mean function of the ARD GP without any smoothness imposed on the visualisation: their relative smoothness is an inherent property of the GP kernel and lengthscales inferred from data. Relevance is ranked on the basis of Relevance Score and Relevance Ratio, as defined in Section 2.4.6.

3.3.2.1 Market Technicals

The results for market technicals are provided in Table 3.3.

Table 3.3: Relevance of Market Technicals.

	Relevance	
Feature	Score	Ratio
Return(t)	0.0637	4.3×10^2
50dSMA	0.5620	3.8×10^3
MACD	0.1783	1.2×10^3
Signal Line	0.0883	6.0×10^2
Noise	0.0002	1

Whilst the MACD-derived Signal Line and previous day's return explained little of the variation in output, the 50-day Simple Moving Average was salient, as was the MACD. Figure 3.2 provides a heatmap of return variation based on the two top features of the technical domain, MACD and 50dMA(t), indexed by percentile score. As a first approximation, MACD and next-day returns move inversely: cheapness with respect to recent history correlates with next-day gains. Given MACD's formal definition as the difference between a 12-day EMA and 26-day EMA, its inclusion

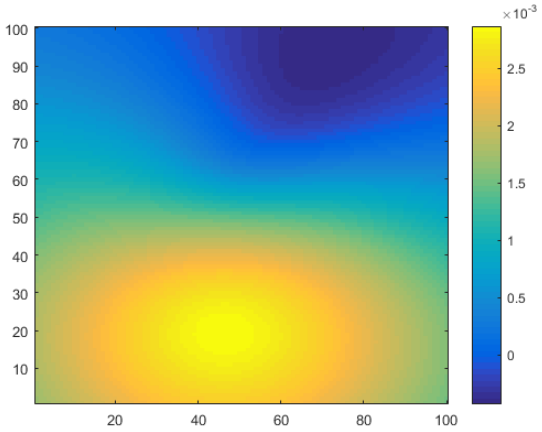

Figure 3.2: S&P500 Daily Return Variation as a function of 50-day Moving Average (x-axis) and MACD (y-axis).

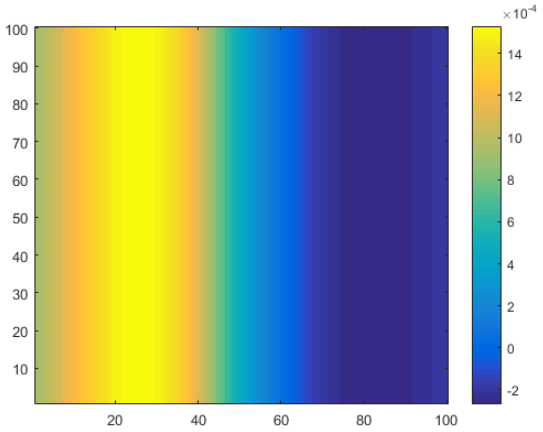

Figure 3.3: S&P500 Daily Return Variation as a function of Stocktwits Sentiment (x-axis) and Twitter Sentiment (y-axis).

in the ARD GP model is effectively teaching it that equity markets mean-revert on a 1-month timescale, automating the rediscovery of time series properties that were manually identified in the parametric literature on financial markets (Jegadeesh, 1990).

3.3.2.2 Sentiment Data

Table 3.4 provides an analysis of sentiment feature relevance. Stocktwits sentiment data is significantly more informative than Twitter data, to the point where the Twitter feature is irrelevant and can be discarded. As a social media site focused on finance, it is likely that Stocktwits's polarity reflects solely market sentiment, whereas Twitter's captures public opinion on a wide range of market-irrelevant issues (e.g. celebrity gossip, local politics). The 1-day change variables were also meaningless and were discarded from subsequent analysis. Notably, the mean function learned through GP regression calls into question the wisdom of crowds: as Figure 3.3 indicates, optimism on Stocktwits foreshadows broad market declines, and conversely. Sentiment analysis lends credence to the Warren Buffett adage: "be greedy when others are fearful, fearful when others are greedy."

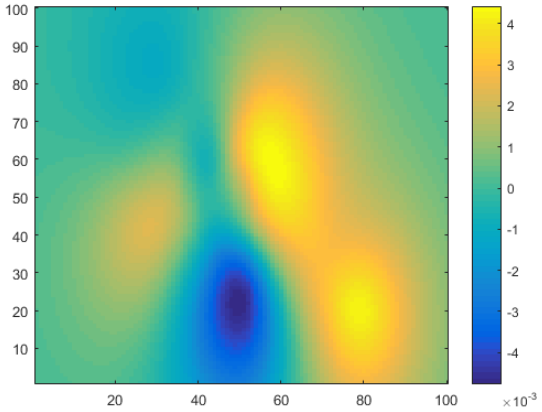

Figure 3.4: S&P500 Daily Return Variation as a function of Directionality (x-axis) and Viscosity (y-axis).

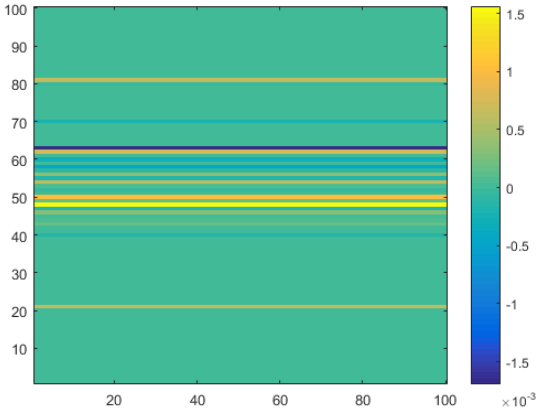

Figure 3.5: S&P500 Daily Return Variation as a function of Broker State (x-axis) and Broker Change (y-axis).

Table 3.4: Relevance of Sentiment Analysis.

	Relevance	
Feature	Score	Ratio
Stocktwits Index	0.2087	$2,8 \times 10^3$
Stocktwits Change	0.0001	0.9
Twitter Index	0.0001	0.9
Twitter Change	< 0.0001	0.3
Noise	0.0001	1

3.3.2.3 Options Market Metrics

Relevance for options-derived metrics is provided in Table 3.5. Directionality and viscosity were almost equally relevant, with positive directionality - that is, experts pre-positioning for rallies via call options - anticipating positive next-day returns. Viscosity instead tracked areas of return compression, and acted as a form of friction. This manifests in Figure 3.4 as areas of peak return coinciding with high directionality and low viscosity.

Table 3.5: Relevance of Price Space.

	Relevance	
Feature	Score	Ratio
Directionality	0.5656	4.7×10^3
Viscosity	0.3844	3.2×10^3
Noise	0.0001	1

3.3.2.4 Broker Recommendations

The relevance of broker actions is assessed in Table 3.6. Broker upgrades and downgrades are infrequent occurrences, resulting in a sparse Broker Change input. The Matérn 3/2 kernel is capable of learning the non-smooth behaviour exhibited in Figure 3.5, but with relevance metrics indistinguishable from Gaussian noise, it is unlikely this domain will provide meaningful improvements to a combined model. In particular, the horizontal lines of Figure 3.5 imply that the Broker State variable is irrelevant, much like the Twitter Sentiment variable of Section 3.3.2.2. This suggests that analyst opinions have little predictive power, and merely reflect market changes after they've occurred.

Table 3.6: Relevance of Broker Recommendations.

	Relevance	
Feature	Score	Ratio
Broker State	0.0157	2.0×10^{-2}
Broker Change	0.2649	0.3×10^{-1}
Noise	0.4523	1

Retaining only the salient features, we run a high-dimensional Gaussian Process regression on relevant inputs from all domains simultaneously. The results, compiled in Table 3.7, broadly mirror our expectations from the correlation analysis, highlighting the ARD framework's ability to discover structure in the data.

Table 3.7: Relevance across all Domains measured on the training set, N=503 entries (Jan-13 to Dec-14).

	Relevance		Pearson	
Feature	Score	Ratio	Correlation	p-value
Directionality	0.3698	7.5×10^3	+0.1011	0.0234
Viscosity*	0.3332	6.7×10^3	−0.2262	0.0001
Stocktwits	0.0738	1.5×10^3	−0.1103	0.0133
50dMA	0.6660	1.3×10^4	−0.0451	0.3130
MACD	0.3159	6.4×10^3	−0.1403	0.0016
Broker Change	< 0.0001	1.58	+0.0024	0.9564
Noise	< 0.0001	1	−0.0238	0.5948

* Correlation for Viscosity measures the correlation between the Viscosity variable and absolute next-day log-returns.

3.3.3 Model Performance

Having established a method for identifying salient features, we now turn our attention to the predictive performance of ARD Gaussian Processes using each data domain. We separately test the predictive value of each domain before fusing them into a combined model, and measure performance according to the Pearson correlation between forecasts and observations, Median Absolute Deviation (MAD) and Normalised Root Mean Square Error (NRMSE), where the normalisation constant is the standard deviation of the observations. The results are provided in Table 3.8.[4]

[4]For clarity: correlation in this table refers to the Pearson correlation between ARD GP forecasts and observations in the test window. Similarly, MAD and NRMSE track the error between ARD GP forecasts and observations.

Table 3.8: ARD GP Performance measured on the test set, N=75 (Jan-15 to Apr-15). We highlight in bold the features or domains for which the correlation between observations and forecasts yields Pearson p-values below 0.05.

	Pearson		Performance	
Feature	Correlation	p-value	MAD (bp)	NRMSE
MACD	**+0.2387**	**0.0392**	**58.22**	**0.9834**
Stocktwits	+0.1779	0.1268	52.51	0.9888
Directionality	**+0.2412**	**0.0371**	**54.07**	**0.9769**
Viscosity	−0.1635	0.1611	51.59	0.9880
Broker Change	−0.1206	0.3026	51.73	0.9941
Technicals (all)	**+0.3079**	**0.0072**	**56.99**	**0.9796**
Sentiment (all)	+0.1779	0.1268	52.51	0.9888
Price Space (all)	**+0.3315**	**0.0037**	**60.76**	**0.9477**
Broker Data (all)	−0.1343	0.2505	51.73	0.9941
Combined	**+0.3803**	**0.0008**	**51.53**	**0.9298**

The model registers monotonic improvements in performance when additional features are included, with the options market data providing the greatest gain. This was the most surprising discovery of the entire study: industry and academia alike have historically focused significant resources on mining technical patterns for predictive signals. As such, we had fully expected technical data to surpass other domains for informativeness, yet found options positioning to be a superior predictor of market fluctuations. Given the relative recency of our dataset (2013-2015), this may merely be a reflection that technical datamining by practitioners has reached a saturation point, its signals so well anticipated that they no longer exhibit the significance that led to their original prominence.

3.3.4 Offline Benchmarks

We benchmark ARD Gaussian Process regression against a range of parametric alternatives. It strictly outperforms traditional single-factor financial models such as look-ahead AR Processes on measures of ground-truth correlation and NRMSE (Ta-

ble 3.9). We also include in our benchmarks the full linear regression model including all features, as well as regularised linear regression using ridge regression and LASSO. The optimal penalty parameter λ in the ridge and LASSO loss function of Equations (2.12) and (2.14) are determined through 10-fold cross validation on the training set. LASSO prunes similar features to ARD: under LASSO, the most salient features are (in order) MACD, Stocktwits, Broker Data, Directionality, Viscosity. This presents a significant overlap to ARD's selection (50dMA, Directionality, Viscosity, MACD, Stocktwits, per the Relevance Ratios in Table 3.7), yet yielded a substantially less accurate model (0.9706 NRMSE for LASSO vs 0.9298 for ARD).

Table 3.9: Benchmark Performance.

Model	Correlation	MAD (bp)	NRMSE
AR(1)	+0.0018	51.34	0.9937
AR(3)	+0.2538	51.35	0.9917
AR(10)	+0.2328	51.33	0.9878
LinReg	+0.2136	58.61	0.9736
Ridge	+0.2298	60.07	0.9724
LASSO	+0.2926	60.42	0.9706

3.3.5 Adaptive ARD Gaussian Process Regression

Over timeframes much larger than our study's 28-month window, supervised batch algorithms in finance run the risk of failing to recognise significant changes in the landscape. For example, Stocktwits sentiment's relevance would have been very low when the site was launched in 2009, and grew in tandem with the size of its user base. A solution to this challenge involves adaptively learning the kernel hyperparameters from recent history only, removing the impact of old, potentially irrelevant data. The evolution from offline to online, adaptive learning follows straightforwardly: we define a window w over which to train an adaptive ARD Gaussian Process for next-day predictions. Rolling the window forward, we infer updated hyperparameters for each day, initialising them to the values obtained in the preceding window to speed

up computation. We generate forecasts in our test set using the optimally combined feature set, measuring model performance as before. Performance metrics for the Adaptive ARD Gaussian Process model are included in Table 3.10.

Table 3.10: Adaptive ARD GP Performance measured on the test set, N=75 (Jan-15 to Apr-15). We highlight in bold the timeframe corresponding to peak model performance, as measured by the Pearson correlation between observations and forecasts.

	Pearson		Performance	
Window Length	Correlation	p-value	MAD (bp)	NRMSE
$w = 150$	+0.2922	0.0110	50.96	0.9990
$w = 175$	+0.3181	0.0054	44.12	0.9769
$w = 200$	+0.3019	0.0085	49.08	0.9756
$w = 225$	+0.3147	0.0060	53.29	0.9692
w = 250	**+0.3797**	**0.0007**	**43.33**	**0.9377**
$w = 275$	+0.3551	0.0018	48.64	0.9579
$w = 300$	+0.3368	0.0031	52.06	0.9686
$w = 325$	+0.2966	0.0098	61.31	0.9789
$w = 350$	+0.3111	0.0066	61.62	0.9620
$w = 375$	+0.3236	0.0046	57.80	0.9438
$w = 400$	+0.3526	0.0019	54.84	0.9313
$w = 425$	+0.3359	0.0032	63.02	0.9369
$w = 450$	+0.3584	0.0016	58.86	0.9286
$w = 475$	+0.3508	0.0020	57.77	0.9313
$w = 500$	+0.3636	0.0013	57.93	0.9275

Predictive performance dips to impractical levels below the $w = 250$ threshold corresponding to one full year's data, highlighting the need for a critical mass of data for Gaussian Process regression and hinting at seasonal variance in stock market returns, in line with a long history of empirical studies on the topic of annual cyclicality (Lakonishok and Smidt, 1989; Agrawal and Tandon, 1994). Factoring in correlation and Mean Absolute Deviation measures, the best adaptive performance was obtained using exactly one full year of the most recent data.

3.3.6 Online Benchmarks

In Table 3.11 we provide performance metrics on benchmark adaptive models such as one-step-ahead AR with varying lags, as well as 150-day, 250-day and 500-day rolling-window variants of the regression models of Section 3.3.4, and find the Adaptive ARD GP yields both superior results and the benefit of automatic, reliable feature selection. A notable analogue to the adaptive GPs of Section 3.3.5: using less than a full year's data for the rolling window produces a sharp degradation in the performance of parametric models, much as it did for the non-parametric models of Table 3.10.

Table 3.11: Adaptive Benchmark Performance.

Model	Correlation	MAD (bp)	NRMSE
AR(1)	+0.0539	51.41	0.9926
AR(3)	+0.0134	52.21	0.9943
AR(10)	+0.1581	53.33	0.9816
LinReg (w=150)	+0.0730	63.60	0.9987
Ridge (w=150)	+0.1154	56.91	0.9942
LASSO (w=150)	+0.0613	51.85	0.9915
LinReg (w=250)	+0.2469	55.01	0.9637
Ridge (w=250)	+0.2576	60.74	0.9648
LASSO (w=250)	+0.2877	57.56	0.9659
LinReg (w=500)	+0.1994	62.16	0.9762
Ridge (w=500)	+0.2102	64.97	0.9752
LASSO (w=500)	+0.2925	59.87	0.9644

3.4 Summary

Extracting information from multiple domains presents the dual challenge of identifying both what to pick and how to mix. This chapter provides a principled framework for reducing input dimensionality through iterative ARD GP regression. We show measurable gains in predictive performance from fusing multiple data streams together in an online setting. Technical analysis proved highly informative for forecasting market movements, and reminds us to pay heed to the metrics that agents on

financial markets will watch - even if their derivation may sometimes appear arbitrary or arcane. Our work also draws attention to the relevance of options market data and the implicitly inhomogeneous representation of price space. As an untapped, feature-rich, strike-dependent dataset shaped by the interactions of informed players, options market salience provides a strong mandate for further research into data-driven modelling of price space and its implications for financial forecasting.

Chapter 4

Deep Learning for Technical Patterns

4.1 Introduction

The ranking of datasets, performed in Chapter 3 via the implementation of Automatic Relevance Determination in our Gaussian Process kernels, allowed us to identify domains that warrant closer inspection. In light of their high relevance and extensive documentation in the literature, we now turn our attention to market technicals, and in particular a branch of technical analysis involving visual pattern-seeking.

Known as *chartism*, this form of financial analysis relies solely on historical price and volume data to produce forecasts, on the assumption that specific graphical patterns hold predictive information for future asset price fluctuations (Blume et al, 1994). Early research into genetic algorithms devised solely from technical data (as opposed to e.g. fundamentals or sentiment analysis) showed promising results, sustaining the view that there could be substance to the practice (Neely et al, 1997; Allen and Karjailainen, 1999). Unlike the *regression* models of Chapter 3, the output of chartist methods does not typically provide a numerical estimate of the return expected by its visual patterns. The forecasts are a purely directional *classification* of outcomes: a specific pattern implies an upwards or downwards price evolution of unspecified magnitude.

Fuelled by advances in computational processing power and data availability in the past decade, the rising popularity of neural networks as state-of-the-art classifiers renewed interest in their applicability to the domain of finance. Krauss et al (2017) applied multilayer perceptrons (MLPs) to find patterns in the daily returns of the S&P500 stock market index. Dixon et al (2017) further demonstrated the effectiveness of neural nets on intraday data, deploying MLPs to classify returns on commodity and FX futures over discrete 5-minute intervals. Architectures comprised of 4 dense hidden layers were sufficient to generate annualised Sharpe ratios in excess of 2.0 on their peak performers. In each instance, patterns were sought in the time series of returns rather than in the price process itself.

Seminal findings by Lo et al (2000) employed instead the visuals emerging from line charts of stock closing prices, relying on kernel regression to smooth out the price process and enable the detection of salient trading patterns. An equally common visual representation of price history in finance is the candlestick. Candlesticks encode the opening price, closing price, maximum price and minimum price over a discrete time interval, visually represented by a vertical bar with lines extending on either end. Much as with line charts, technical analysts believe that specific sequences of candlesticks reliably foreshadow impending price movements. A wide array of such patterns are commonly watched for (Taylor and Allen, 1992), each with their own pictogram and associated colourful name ('inverted hammer', 'abandoned baby', etc).

Though recurrent neural networks - and in particular Long Short-Term Memory (LSTM) models (Hochreiter and Schmidhuber, 2017) - have been the most popular choice for deep learning on time series data (Tsantekidis et al, 2017; Fischer and Krauss, 2017; Dixon, 2017), promising results have begun to appear from the application of convolutional neural networks to financial data. Neural networks have frequently been labeled as black boxes, limiting their deployment in domains where interpretability is sought. Convolutional neural networks partially overcome this, by

extracting locally interpretable features in their early layers. Furthermore, recent research suggests that these models bear the capacity to generalise not merely across time but across assets as well, identifying universal features of stock market behaviour (Zhang et al, 2018).

Drawing on this intuition for locally interpretable visual pattern recognition, we reframe candlestick patterns as a form of *feature engineering* intended by chartists to extract salient features, facilitating the classification of future returns with higher fidelity than the raw price process would otherwise allow. After defining the data used throughout the chapter (Section 4.2), we motivate the pursuit of new, better visual heuristics for finance by assessing the predictiveness of candlestick formations (Section 4.3). Feeding candlestick data through a neural network involving separate filters for each technical pattern, we classify next-day returns with the filters implied by chartist doctrine (Section 4.4.1-4.4.2) and set this cross-correlational approach as a baseline to improve upon (Romaszko, 2015). We then compare the model's accuracy when filters are not preset but instead learned by convolutional neural networks (CNNs) during their training phase (Section 4.4.3), and benchmark deep learning against alternative methods drawn from both traditional finance and machine learning (Section 4.4.4). We enhance the accuracy of CNNs through the addition of thresholding and ensembling (Section 4.4.5), and finish with two practically-minded extensions: the backtested performance of the model (Section 4.4.6) and the visual interpretation of the features extracted by the CNN (Section 4.4.7).

The contributions of this chapter are threefold: firstly, we rigorously evaluate the practice of candlestick chartism, and find little evidence to support it. We agree with Lo et al (2000) that the distribution of future returns conditioned on observing technical patterns diverges significantly from the unconditional distribution, but upon close inspection the resulting classifier barely outperforms guesswork. Secondly, we show that filters learned and tested on 22 years of S&P500 price data in a CNN

architecture can yield modest gains in accuracy over both technical methods and machine learning alternatives, including MLPs unsupported by convolution's feature-extractive capabilities. Thirdly, we demonstrate that considerable gains in forecasting capability are achievable through ensemble methods and confidence thresholds.

4.2 Data

4.2.1 Definition of Candlestick Data

Both the financial time series data and the candlestick technical filters used by chartists take the same form. Asset price data for a discrete time interval is represented by four features: the opening price (price at the start of the interval), closing price (price at the end of the interval), high price (maximum over the interval) and low price (minimum over the interval). The candlestick visually encodes this information (Figure 4.1): the bar's extremities denote the open and close prices, and the lines protruding from the bar (the candle's 'wicks' or shadow) denote the extrema over the interval. The colour of the bar determines the relative ordering of the open and close prices: a white bar denotes a positive return over the interval (close price > open price) and a black or shaded bar denotes a negative return (close price < open price).

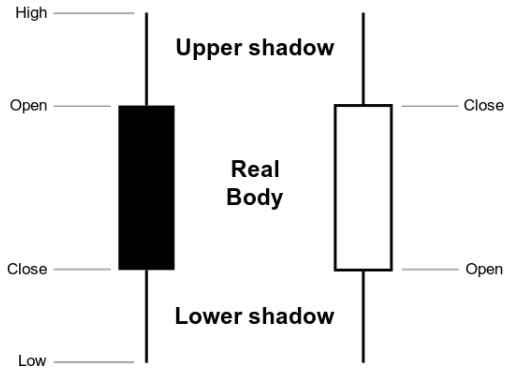

Figure 4.1: Candlestick representation of financial time series data.

We can therefore summarise the candlestick representation of a financial time series of length n timesteps as a $4 \times n$ price signal matrix F capturing its four features. Throughout this chapter we rely on daily market data, but the methods can be extended to high-frequency pattern recognition on limit order books - an active area for current research (Sirignano and Cont, 2018; Zhang et al, 2018).

4.2.2 Definitions of Technical Patterns

We include major candlestick patterns cited by practitioners of technical analysis at three timescales: 1-day, 2-day and 3-day patterns. The simple 1-day patterns include the hammer (normal and inverted), hanging man, shooting star, dragonfly doji, gravestone doji, and spinning tops (bullish and bearish, where bullish implies a positive future return and bearish implies a negative future return). Our 2-day patterns cover the engulfing (bullish and bearish), harami (bullish and bearish), piercing line, cloud cover, tweezer bottom and tweezer top. Finally our 3-day patterns cover some of the most cited cases in chartist practice: the abandoned baby (bullish and bearish), morning star, evening star, three white soldiers, three black crows, three inside up and three inside down. Figure 4.2 provides both the visual template associated with each pattern, as well as the future price direction it is meant to presage. We summarise a technical pattern P of length m timesteps as a $4 \times m$ matrix T_{P_m}, standardised for comparability to have zero mean and unit variance.

4.2.3 Empirical Data

Throughout this chapter, we use daily technical (i.e. open, close, high and low price) data from the S&P500 stock market index constituents for the period Jan 1994 - Dec 2015, corresponding to n = 2,439,184 entries of financial data in the price signal

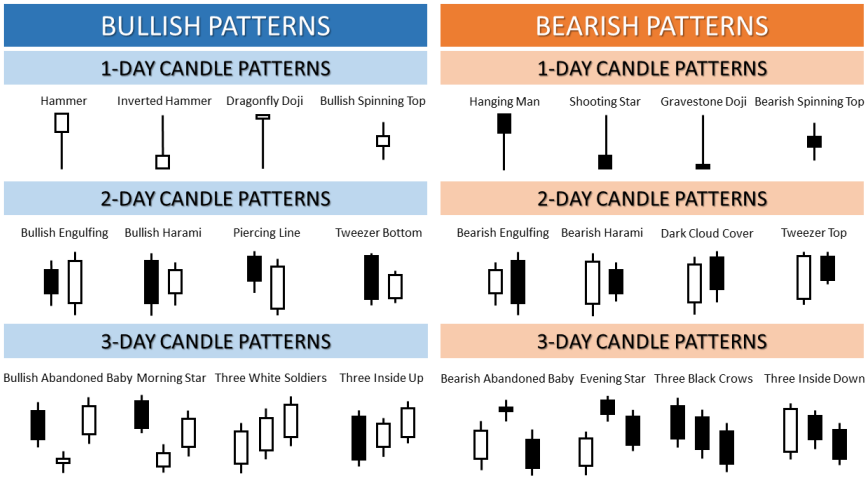

Figure 4.2: For each timescale (1-day, 2-day and 3-day), we specify 8 chartist patterns and the future direction they predict ('bullish' for positive returns, 'bearish' for negative returns).

F.[1] This dataset covers a representative cross-section of US companies across a wide timeframe suitable for learning the patterns, if any, of both expansionary and recessionary periods in the stock market. The means of generating this dataset is provided in Appendix A.2.

4.3 Motivation

As a preliminary motivation for the adoption of machine learning for technical forecasts, we assess the merits of candlestick chartism in finance. We run several diagnostics to assess separately the informativeness and predictiveness for each technical pattern.

[1]We include 500 individual stocks from the S&P500 index as of 31st December 2015. Each stock's daily open, close, high and low is recorded over a period of up to 22 years, with each year including 252 business days on average.

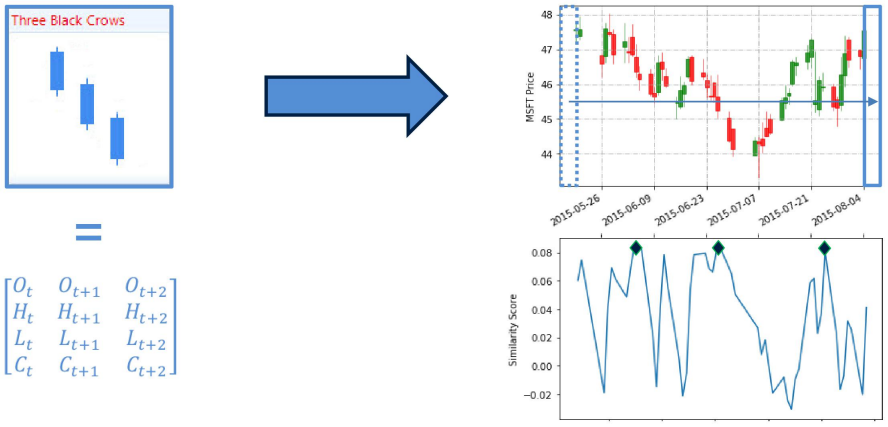

Figure 4.3: The pattern 'Three Black Crows' is transformed into a standardised 4 × 3 matrix corresponding to the Open, High, Low and Close prices on each of the three days of the pattern. The methodology calculates the inner product between the pattern's matrix form and standardised 3-day intervals of the Microsoft price history, generating a time series of similarity scores. High values can be interpreted as detections of the pattern.

4.3.1 Conditioning Returns on the Presence of a Pattern

The $4 \times m$ matrix representation T_{P_m} for pattern P of length m and equal-length, standardised rolling windows F_n of the full price signal F at timestep n can be cross-correlated together to generate a time series S_P measuring the degree of similarity between the price signal and the pattern. For a given pattern P, at each timestep n:

$$S_{P,n} = \left\langle \frac{T_{P_m}}{\|T_{P_m}\|}, \frac{F_n}{\|F_n\|}, \right\rangle \tag{4.1}$$

where $\langle \cdot, \cdot \rangle$ is the inner product of the two matrices and $\| \cdot \|$ is the L^2 norm.

For each pattern P, we produce a conditional distribution of next-day returns by extracting the top quantile (in our study, decile and centile) of similarity scores S_P. Figure 4.3 provides an example of this methodology, applying the pattern 'Three Black Crows' to the share price of Microsoft over the period May-August 2015.

4.3.2 Informativeness

For our purposes we define a technical pattern to be *informative* if its presence significantly alters the distribution of next-day returns in a Kolmogorov-Smirnov two-sample (K-S) test (Massey, 1951), comparing the unconditional distribution of all next-day returns to the distribution conditioned on having just witnessed the pattern. Denoting by $\{R_{P}{}_{t=1}^{n_1}\}$ the subset of returns conditioned on matching pattern P and $\{R_{t=1}^{n_2}\}$ the full set of unconditional returns, we compute their empirical cumulative distribution functions $F_1(z)$ and $F_2(z)$. The K-S test evaluates the null hypothesis that the distributions generating both samples have identical cdfs, by computing the K-S statistic:

$$\gamma = \left(\frac{n_1 n_2}{n_1 + n_2}\right)^{1/2} \sup_{-\infty<z<\infty} |F_1(z) - F_2(z)| \tag{4.2}$$

The limiting distribution of γ provides percentile thresholds above which we reject the null hypothesis. When this occurs, we infer that conditioning on the pattern does materially alter the future returns distribution.

4.3.3 Predictiveness

Whilst these patterns may bear some information, it does not follow that their information is actionable, or even aligns with the expectations prescribed by technical analysis. Notched boxplots of both unconditional returns and returns conditioned on each of the filters (Figure 4.4) allow us to gauge whether the pattern's occurrence does in fact yield significant returns in the intended direction.

A closer examination suggests several of the 1-day patterns are in fact relevant, but that the more elaborate 2-day and 3-day formations are not. Conditioning on 14 of the 16 multi-day patterns produces no significant alteration in the median of next-day returns distributions (Figure 4.5): only the 'Bearish Engulfing' and 'Three

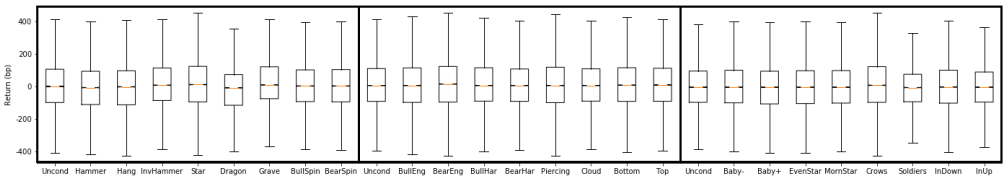

Figure 4.4: Notched boxplots of the distributions of returns in basis points (one hundredth of a percent), conditional on observing each of the technical patterns (similarity score S_P in its top centile). Whiskers cover twice the interquartile range. At a glance, none of the conditional distribution medians diverge substantively from the unconditional baseline, and the distributions' standard deviations dwarf their medians by two orders of magnitude.

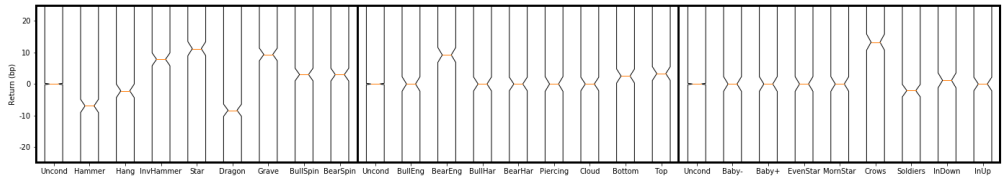

Figure 4.5: Close-up of boxplot notches for the distributions of returns in basis points (one hundredth of a percent), conditional on observing each of the technical patterns (similarity score S_P in its top centile). Absence of overlap between the boxplot notches of a conditional distribution and the unconditional distribution provides evidence at the 95% confidence threshold that the medians of the 2 distributions differ (Chambers et al, 1983). Surprisingly, several single-day patterns do in fact correlate with abnormal next-day returns. Almost all of the multi-day patterns exhibit notches that overlap with the unconditional distribution's, implying that the distribution medians are not meaningfully changed by conditioning. Only 'Bearish Engulfing' and 'Three Black Crows' seem to be significant - as harbingers of better times, despite their names.

Black Crows' patterns produce a conditional distribution for which the 95% confidence interval of the median (denoted by the notch) differs markedly from its unconditional counterpart.

4.3.4 Findings

We report the empirical results of the K-S goodness of fit tests and top decile and centile (Tables 4.1 and 4.2 respectively) conditional distribution summary statistics, using daily stock data from the S&P500. Though several of the patterns do indeed bear information altering the distribution of future returns, their occurrence is neither a reliable predictor of price movements (high standard deviation relative to both the median per Figure 4.4 and the mean per Tables 4.1 and 4.2) nor even, in many instances, an accurate classifier of direction. Elaborate multi-day patterns systematically perform worse than their single-day counterparts. Surprisingly, 6 of the 8 single day patterns do in fact produce meaningful deviations from the unconditional baseline, with the dragonfly and gravestone doji standing out as significant outliers (-25.81 bp[2] and +22.41 bp respectively when conditioning on the top centile of similarity score, Table 4.2). But even in those instances, technical analysis forecasts incorrectly, as prices move in the direction opposite to chartism's predictions. McLean and Pontiff (2016) showed that predictor variables lose on average 58% of their associated return, post-publication. In a similar vein, we hypothesise that these patterns may have once been predictive on a historical dataset, but that their disclosure and subsequent overuse has long since negated their value. Conceptually, the notion of using filters in financial data to extract informative feature maps may bear merit - but the chartist filter layer is demonstrably an improper specification today.

[2] A basis point (bp) corresponds to one hundredth of one percentage point.

Table 4.1: Summary statistics for the next-day return distributions conditioned on matching technical patterns. A match on pattern P is deemed to have occurred when the cross-correlational similarity score S_P is in its top *decile*. K-S statistics γ above 1.86 are significant at the 0.001 level. Mean return μ for each pattern is expressed as a difference from the unconditional baseline. The incremental mean returns are dwarfed by their standard deviation, and do not even always move in the direction prescribed by chartism.

Pattern	γ	$\mu(bp)$	$\sigma(bp)$
Unconditional		4.26	229.40
Hammer	11.17	-10.20	214.20
Inverted Hammer	8.37	+9.83	216.49
Hanging Man	9.84	-11.67	220.59
Shooting Star	10.90	+9.99	222.52
Dragonfly Doji	10.65	-11.37	217.57
Gravestone Doji	9.46	+10.53	220.37
Bullish Spinning Top	4.97	-0.42	209.17
Bearish Spinning Top	5.22	-1.47	213.24
Bullish Engulfing	2.78	+1.30	226.39
Bearish Engulfing	5.65	+0.83	228.06
Bullish Harami	2.65	+4.11	224.22
Bearish Harami	4.39	-0.04	214.61
Piercing Line	2.28	+0.44	232.04
Cloud Cover	2.56	-0.72	220.12
Tweezer Bottom	3.45	+3.26	230.10
Tweezer Top	2.78	+0.93	218.13
Abandoned Baby-	4.03	-4.24	224.60
Abandoned Baby+	2.88	+4.87	227.16
Evening Star	2.53	-2.32	223.24
Morning Star	2.79	+4.86	228.15
Three Black Crows	14.28	+5.62	265.14
Three White Soldiers	12.97	-7.98	208.90
Three Inside Down	2.91	+0.45	231.62
Three Inside Up	3.27	+0.71	220.71

Table 4.2: Summary statistics for the next-day return distributions conditioned on matching technical patterns more stringently. A match on pattern P is deemed to have occurred when the cross-correlational similarity score S_P is in its top *centile*.

Pattern	γ	$\mu(bp)$	$\sigma(bp)$
Unconditional		4.26	229.40
Hammer	5.13	-15.80	223.04
Inverted Hammer	5.00	+13.75	211.62
Hanging Man	3.71	-14.92	222.06
Shooting Star	4.78	+12.01	232.42
Dragonfly Doji	14.73	-25.81	219.99
Gravestone Doji	12.93	+22.41	223.57
Bullish Spinning Top	2.64	-0.72	214.70
Bearish Spinning Top	1.67	+0.94	213.30
Bullish Engulfing	1.61	-0.28	236.50
Bearish Engulfing	4.16	+5.75	238.47
Bullish Harami	1.07	+5.5	222.17
Bearish Harami	1.51	-0.96	219.43
Piercing Line	2.29	+0.78	241.42
Cloud Cover	1.06	-0.75	218.24
Tweezer Bottom	1.76	+4.19	223.23
Tweezer Top	1.22	+2.97	221.93
Abandoned Baby-	3.29	-4.04	232.45
Abandoned Baby+	1.27	+2.94	232.28
Evening Star	2.89	-0.27	231.76
Morning Star	1.80	+2.59	231.89
Three Black Crows	6.85	+13.09	229.40
Three White Soldiers	6.30	-11.77	203.26
Three Inside Down	1.63	+2.72	233.12
Three Inside Up	2.50	+0.13	220.75

4.4 Results

In Chapter 3, we applied ARD GPs to learn a mapping function between known features (e.g. moving average crossovers, sentiment indices on social media, etc.) and next-day returns on the S&P500. Our challenge in this chapter runs deeper: not only to learn the functional relationship between technical features and next-day returns on single stocks, but also to extract those features from raw candlestick price data. In effect, our model must be capable of teaching itself the shape of technical indicators worth tracking - and then of learning the complex, non-linear relationship between those indicators and future returns. To achieve this dual purpose, we rely on convolutional neural networks, as described in Section 2.3.5.

The approach of searching for informative intermediate feature maps in classification problems has seen widespread success in domains ranging from acoustic signal processing (Hinton et al, 2012) to computer vision (Krizhevsky et al, 2012). Where technical analysis uses filters that are arbitrarily pictographic in nature, we turn instead to *convolutional layers* to extract salient features. We evaluate the performance of passing the raw data both with and without chartist filters, and subsequently measure the incremental gain from learning optimal feature maps by convolution. The findings are then benchmarked against widely recognised approaches to time series forecasting including autoregression, recurrent neural networks, nearest neighbour classifiers, support vector machines (SVM) and random forests.

In the experimental results that follow, we split our S&P500 time series data into training and test sets corresponding to single stock prices from 1994-2004 and 2005-2015 respectively.[3,4]

[3]An extension to this work could include an alternative form of data partitioning, e.g. an online learning scheme similar to the adaptive GP models of Section 3.3.5, featuring rolling in- and out-of-sample periods for training and testing. The simple train-test partition we've adopted is by no means definitive, but serves as a proof of concept for the application of deep learning in this space.

[4]Our classes are best be defined as 'negative return' and 'strictly positive return'. As zero return days occur (albeit infrequently) in assets with low denomination, we address the issue of class imbalance in two ways. Firstly, we add Gaussian noise with variance 10^{-6} to all returns, evenly

4.4.1 Multi-Layer Perceptron

To address issues of scale and stationarity, we process the original $4 \times n$ price signal matrix F into a new $80 \times n$ price signal matrix F^* where each column is a standardised encoding of 20 business days of price data. This encoding provides 4 weeks of price history, a context or 'image' within which neural network filters can scan for the occurrence of patterns and track their temporal evolution. We pass F^* through a multilayer perceptron (MLP) involving fully-connected hidden layers. Preliminary 5-fold cross-validation experiments with financial time series determined the network topology required for the model to learn from its training data.[5] Insufficient height (neurons per hidden layer) and depth (number of hidden layers) led to models incapable of learning their training data. We settled on 2 fully-connected layers of 64 neurons with ReLU activation functions, followed by a softmax output layer to classify positive and negative returns. Regularisation was achieved via the inclusion of dropout (Srivastava, 2014) in the fully-connected layers of the network, limiting the model's propensity towards excessive co-adaptation across layers. A heavily-regularised (dropout = 0.5) 2-layer MLP is already able to identify some structure in its data (out-of-sample accuracy of 50.6% after 100 epochs, Table 4.3).

4.4.2 Technically-Filtered MLP

Reframing technical patterns as pre-learned cross-correlational filters, we consider for each pattern length m the 8 pattern matrices T_{P_m} defined visually in Figure 4.2. Each such formation, of form $4 \times m$, is stacked along the depth dimension, producing a $4 \times m \times 8$ tensor T whose inner product with standardised windows of the raw price signal F yields a new $8 \times n$ input matrix F_T,

spreading the zero return days across both classes. Secondly, we select the boundary value that separates our training set into two equally-balanced classes (+0.0000005%). The resulting training set is perfectly balanced, against a mildly positive skew in the test set (51.1% strictly positive return days, 48.9% negative return days).

[5]For optimisation, we employed Adam, an extension of stochastic gradient descent that dynamically adapts per-parameter learning rates to deal with sparse and noisy datasets (Kingma, 2015).

Table 4.3: Accuracy (%) obtained after training a 2-layer MLP on single stock data from the S&P500, using open-close-high-low price data.

Epochs	In Sample	Out-of-Sample
1	50.3	50.3
5	50.7	50.4
10	51.4	50.5
50	52.2	50.6
100	52.5	50.6

Table 4.4: Accuracy (%) obtained after training a technically-filtered MLP (filter length $m = 1$) on single stock data from the S&P500, using open-close-high-low price data. Multi-day filters produced similarly lacklustre results. The technical analysis filters produce feature maps with less discernible structure than the original input.

Epochs	In Sample	Out-of-Sample
1	50.0	50.2
5	50.2	49.8
10	50.2	49.7
50	50.4	49.9
100	50.5	49.8

$$F_T = \langle T, F \rangle. \tag{4.3}$$

This new input is the result of cross-correlating the raw price signal F with the technical analysis filter tensor T, and can be interpreted as the feature map generated by technical analysis. We now use F_T as the input to the same MLP as before and look for improvements in model forecasts. The results we find are consistent with Section 4.3: using technical analysis for feature extraction hinders the classifier, slightly degrading model performance (out-of-sample accuracy of 49.8% after 100 epochs using the 1-day patterns, Table 4.4).

Table 4.5: Details of the architecture for a CNN scanning patterns of length m. The number of filters in the convolution layer was deliberately kept low (8) and their dimensions ($4 \times m$) match the technical patterns used in Section 4.4.2, to enable like-for-like comparability with the technical filter approach.

#	Layer	Units	Activation Function	Dropout	Filter Shape	Outgoing dimensions
1	Input	-	-	-	-	(input) [4 × 20]
2	Convolutional	8	ReLU	0.5	[4 × m]	[8 × 20]
3	FC	64	ReLU	0.5	-	[64]
4	FC	64	ReLU	0.5	-	[64]
5	FC	2	Softmax	-	-	(output, 2 classes) [2]

Table 4.6: Accuracy (%) obtained In-Sample (IS) and Out-of-Sample (OoS) after training a deep neural network with a single convolutional layer learning 1-day, 2-day and 3-day patterns.

Filter Length	1-day		2-day		3-day	
Epochs	IS	OoS	IS	OoS	IS	OoS
1	50.3	50.4	50.2	50.3	50.1	50.2
5	50.6	50.3	50.7	50.4	50.5	50.3
10	50.9	50.7	51.0	50.7	51.1	50.6
50	51.4	51.1	51.5	50.9	51.4	51.0
100	51.7	51.3	51.8	51.2	51.7	51.2

4.4.3 Convolutional Neural Network

We now deepen the neural network by adding a single convolutional layer with 8 filters (so chosen to match the number of technical filters at each timescale, per Figure 4.2) to our earlier MLP (architecture detailed in Table 4.5). Separate experiments are run for convolutional filters of size 4, 8 and 12, corresponding to scanning for 1-day, 2-day and 3-day patterns. Their performance is reported in Table 4.6. The CNN finds much greater structure in its training data than the MLP could, and generalises better. Accounting for the size of the test set (n = 1,332,395), the leap from the MLP's out-of-sample accuracy of 50.6% to the 1-day CNN's out-of-sample accuracy of 51.3% is considerable.

4.4.4 Model Evaluation

To investigate whether the predictive performance of the neural network classifiers is not merely considerable but statistically significant, we derive the area under the curve (AUC) of each model's receiver operating characteristic curve (ROC), and exploit an equivalence between the AUC and Mann-Whitney-Wilcoxon test statistic U (Mason and Graham, 2002):

$$AUC = \frac{U}{n_P n_N} \tag{4.4}$$

where n_P and n_N are the number of positive and negative returns in the test set, respectively. In our binary classification setting, the Mann-Whitney-Wilcoxon test evaluates the null hypothesis that a randomly selected value from one sample (e.g., the subset of test data classified as positive next-day returns) is equally likely to be less than or greater than a randomly selected value from the complement sample (the remaining test data, classified as negative next-day returns). Informally, we are testing the null hypothesis that our models have classified at random. The

test statistic U is approximately Gaussian for our sample size, so we compute each model's standardised Z-score and look for extreme values that would violate this null hypothesis.

$$Z = \frac{U - \mu_U}{\sigma_U} \tag{4.5}$$

where:

$$\mu_U = \frac{n_P n_N}{2} \tag{4.6}$$

and

$$\sigma_U = \sqrt{\frac{n_P n_N (n_P + n_N + 1)}{12}} \tag{4.7}$$

We benchmark our CNNs against traditional linear models in finance (AR(1) and AR(5) models), a buy-and-hold strategy and a range of machine learning alternatives detailed below.

4.4.4.1 Recurrent Neural Networks (RNN)

Deep learning for time series analysis has typically relied on recurrent architectures capable of learning temporal relations in the data. Long Short-Term Memory (LSTM) networks have achieved prominence for their ability to memorise patterns across vast spans of time by addressing the vanishing gradient problem. A thorough RNN architecture search (Jozefowicz et al, 2015) identified a small but persistent gap in performance between LSTMs and the recently introduced Gated Recurrent Unit (GRU, Chung et al, 2014) on a range of synthetic and real-world datasets. Our benchmark RNNs involve a preliminary recurrent layer (LSTM and GRU, in separate experiments) of 8 neurons followed by 2 dense layers of 64 neurons with dropout, comparable in architectural complexity to the CNN models of Section 4.4.3.

4.4.4.2 k-Nearest Neighbours (k-NN)

We evaluate a range of nearest neighbour classifiers, labelling each day of the test set with the most frequently observed class label (positive or negative next-day return) in the k training points that were closest in Euclidean space.

4.4.4.3 Support Vector Machines (SVM)

SVMs have been applied to financial time series forecasting in prior literature, and achieved moderate success when the input features were not raw price data but hand-crafted arithmetic derivations like the market technicals of Chapter 3, e.g. Moving Averages and MACD (Kim, 2003). We report SVM performance under different kernel assumptions (linear and RBF), where the model hyperparameters (regularisation parameter C to penalise margin violations, RBF kernel coefficient γ to control sensitivity) were selected by cross-validation on a subset of the training data.

4.4.4.4 Random Forests (RF)

In their study of European financial markets, Ballings et al (2015) evaluated the classification accuracy of ensemble methods against single classifiers. Their empirical work highlighted the effectiveness of random forests in classifying stock price movements and motivates their inclusion in our list of benchmarks, under varying assumptions for the number of trees hyperparameter n.

4.4.4.5 Benchmark Findings

Table 4.7 provides the AUC, Z-score and significance of each model, where significance measures the area of the distribution below Z. We disregard significance for negative Z-scores (as is the case for the technically-filtered neural network) as they imply classifiers that performed (significantly) worse than random chance. The results underscore the scale of the challenge for pattern recognition in finance: deep learning achieved the best results by a significant margin, and most alternative methods

Table 4.7: Benchmark performance across a range of models trained on S&P500 technical data for Jan 1994 - Dec 1994 and tested on Jan 2005 - Dec 2015. Precision and recall are computed as weighted averages across both classes. Significance refers to the p-value of the Mann-Whitney-Wilcoxon test for each model.

Model	Acc	Prec	Rec	F1	AUC	Z	Significance
MLP	50.6	49.7	49.6	49.6	51.1	23.766	< 0.0001
Technical NN	49.8	49.1	49.3	49.2	49.9	-1.878	–
1-day CNN	51.3	50.9	51.2	51.0	51.8	36.546	< 0.0001
2-day CNN	51.2	51.0	51.0	51.0	51.5	31.291	< 0.0001
3-day CNN	51.2	50.8	51.0	50.9	51.5	31.423	< 0.0001
RNN-LSTM	50.8	50.6	51.0	50.8	51.0	19.616	< 0.0001
RNN-GRU	50.9	50.3	50.8	50.6	51.2	24.880	< 0.0001
1-NN	50.0	50.0	50.0	50.0	50.1	1.087	0.1386
10-NN	49.9	49.9	49.9	49.9	49.8	-3.317	–
100-NN	49.7	49.6	49.9	49.8	49.6	-7.651	–
Linear SVM	49.9	49.9	49.8	49.8	49.8	-0.962	–
RBF SVM	49.9	49.8	49.8	49.8	49.8	-2.416	–
10-RF	50.0	49.9	49.8	49.9	50.0	0.256	0.3991
50-RF	49.8	49.9	49.8	49.8	49.7	-5.986	–
100-RF	49.8	49.8	49.7	49.7	49.6	-7.628	–
AR(1)	49.9	50.1	49.9	50.0	49.9	-2.129	–
AR(5)	49.8	50.1	49.8	49.9	49.8	-3.323	–
Buy-and-Hold	51.1	26.2	51.1	34.7	51.2	23.701	< 0.0001

yielded accuracies that were not statistically distinguishable from guesswork.[6] Convolution also outperforms recurrence in our experiments, suggesting that a 20-day window may be sufficient to capture temporal dependencies in markets.

4.4.5 Methodological Extensions to the ConvNet Framework

Learning neural network filter specifications via convolution yields a significant boost to predictive prowess over the baseline model of Section 4.4.1 and technically-filtered variant of Section 4.4.2. The CNNs' outperformance of autoregression and machine learning alternatives further confirms the aptitude of convolutional feature extraction on technical data, and spurs us to target domain-specific enhancements to our deep learning models.

4.4.5.1 Confidence Thresholding

In contrast to mission-critical application domains like autonomous navigation, finance does not require an algorithmic agent to be accurate at all times. It is acceptable (and factoring in friction costs, preferable) for a model to be sparse in making decisions, only generating 'high conviction' calls, if this results in greater accuracy. Furthermore, the output values in the final layer of the CNN can be assigned a probabilistic interpretation, enabling a filtered, nuanced approach to classification. We replicate this by adding a confidence threshold α to the classification output of the final softmax layer of Table 4.5: test points where neither class is assigned a probability greater than α are deemed uncertain, and disregarded by the thresholded

[6]The best alternative was also the simplest: buy-and-hold outperformed many systematic alternatives. This is in part a reflection of the test window (2005-2015) and the upwards bias of equity markets. Note however that compared to buy-and-hold, which involves risk exposure to the entire period, the CNN models deliver long-short portfolios whose returns provide diversification from the broad market performance. Furthermore, the Thresholded CNNs described in Section 4.4.5.1 are more parsimonious in their deployment of risk capital, being exposed only to single days' returns on the top centile of calls. The returns per unit of risk, captured by measures such as the Sharpe ratio of Table 4.9, are considerably different: buy-and-hold on the S&P500 would've yielded a Sharpe ratio of 0.75 over the period 2004-2014, compared with a Sharpe ratio of 4.08 for the TCNN Ensemble with friction costs.

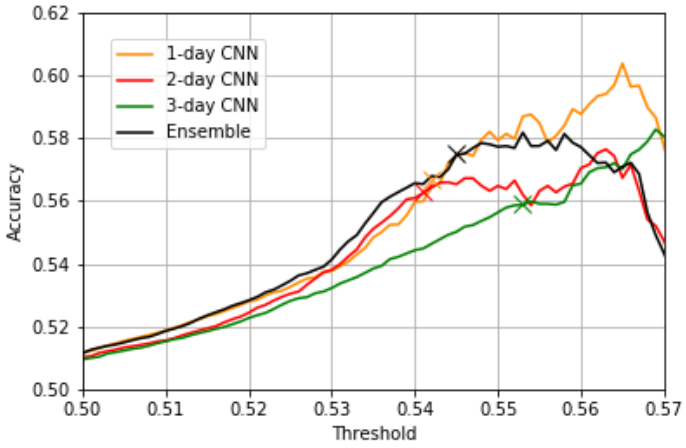

Figure 4.6: Model accuracy as a function of softmax threshold α. For each model, we indicate by a cross the threshold level that retains the 1% of test data for which the model's output probabilities imply the highest confidence.

convolutional neural network (TCNN). For each model (1-, 2- and 3-day TCNN), the confidence threshold α is tuned through 5-fold cross-validation. Accuracy as a function of confidence threshold α is presented in Figure 4.6, and demonstrates in all 3 cases that a substantial increase in model prowess can be achieved by thresholding the softmax output to only consider class assignments with high certainty. We also highlight the α threshold which retains the top centile of test outputs, corresponding to the model's most confident assignments. These vary by model (54.2%, 54.1% and 55.3% for the 1-, 2- and 3-day TCNNs respectively), but in each case form a reliable heuristic for balancing model confidence and sample size. A notable analogue to the study of technical analysis in Section 4.3: models searching for more elaborate multi-day patterns tend to underperform the single-day TCNN.

4.4.5.2 Ensembling TCNNs

An effective technique in image processing involves homogeneous ensembling of multiple copies of the same CNN architecture, averaging across the class assignments of the constituent models (Krizhevsky et al, 2012; Antipova et al, 2016). Combining this

Table 4.8: Performance comparison between MLP, CNN and TCNN models trained on S&P500 technical data for Jan 1994 - Dec 1994 and tested on Jan 2005 - Dec 2015. Precision and recall are computed as weighted averages across both classes. Significance refers to the p-value of the Mann-Whitney-Wilcoxon test for each model.

Model	Acc	Prec	Rec	F1	AUC	Z	Significance
MLP	50.6	49.7	49.6	49.6	51.1	23.766	< 0.0001
Technical NN	49.8	49.1	49.3	49.2	49.9	-1.878	–
1-day CNN	51.3	50.9	51.2	51.0	51.8	36.546	< 0.0001
2-day CNN	51.2	51.0	51.0	51.0	51.5	31.291	< 0.0001
3-day CNN	51.2	50.8	51.0	50.9	51.5	31.423	< 0.0001
CNN Ensemble	51.2	51.0	51.2	51.1	51.7	35.628	< 0.0001
1-day TCNN	56.7	56.5	56.6	56.5	57.2	14.533	< 0.0001
2-day TCNN	56.3	56.1	56.7	56.4	56.5	13.017	< 0.0001
3-day TCNN	55.9	57.1	55.9	56.5	56.2	12.493	< 0.0001
TCNN Ensemble	57.5	56.9	57.0	56.9	57.5	15.301	< 0.0001

probabilistic interpretation of the softmax layer with model averaging, we construct a heterogeneous ensemble out of our 1-day, 2-day and 3-day TCNNs. The ensemble benefits from learning patterns manifesting at different timescales, and achieves a higher accuracy (57.5%) on its top-confidence centile than any of the individual learners (56.7%, 56.3% and 55.9% for the 1-day, 2-day and 3-day TCNN respectively, Figure 4.6).

Performance metrics of both the TCNNs and TCNN ensemble are provided in Table 4.8. Whilst the Z-scores of the TCNN models are lower than those of unthresholded CNN models, this is primarily the consequence of sample size on statistical significance tests - AUC improves markedly under thresholding.

4.4.6 Practical Implementation

Through thresholding, we enforce sparsity in the model's decision making. In a real-world deployment, infrequent activity keeps friction costs low - a desirable property for trading algorithms. We track the activity level of the various models over time, as well as the cumulative profit they would generate over the 11-year test window.

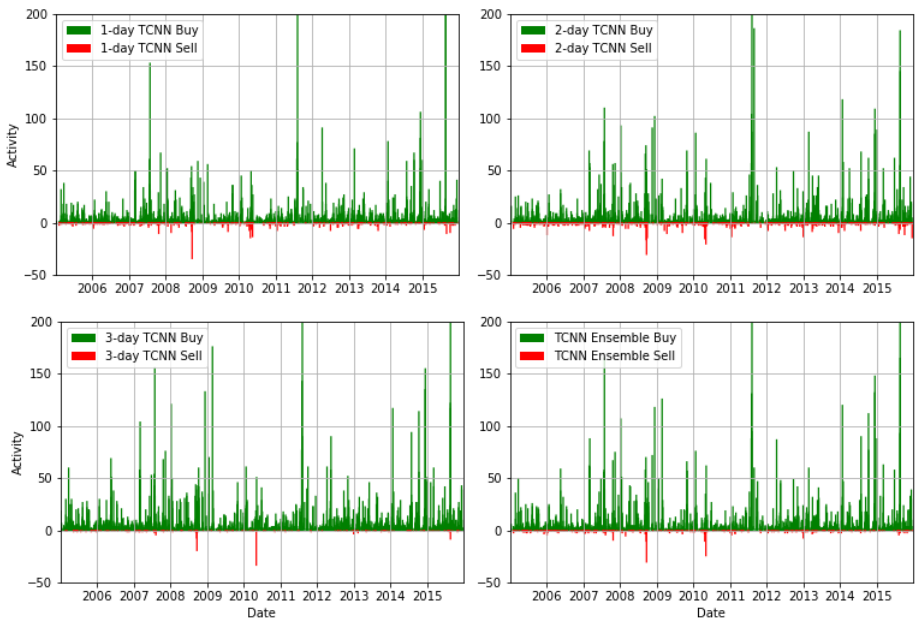

Figure 4.7: Activity level of the various TCNN models through a 11-year period. As we retain the top centile from the 1,408,679 test points, each model is generating 14,087 trading decisions over 2868 business days, or on average 4.91 trades per day. Though the model is active throughout the window, discernible spikes in activity occur around major events, most notably the US debt ceiling crisis in August 2011.

We assume the model fully captures the 1-day return associated with the top centile of its thresholded class assignments, additively for positive class predictions and subtractively for negative class predictions.

The models are heavily skewed towards buying activity, with accurately-timed spikes centred around major world events (Figure 4.7). The 2 largest single-day buy orders (in the sense of number of buys across all constituents of the index) occur on the 9th of August 2011 (328 buys), at the tail end of the US debt ceiling crisis which caused the S&P500 to drop 20% in 2 weeks, and on the 24th of August 2015 (241 buys), following a flash crash in which US markets erased 12% of their value before recovering. The largest sell volume occurs on the 22nd of September 2008 (31 sells), a full week after the collapse of Lehman Brothers. This coincides with market-

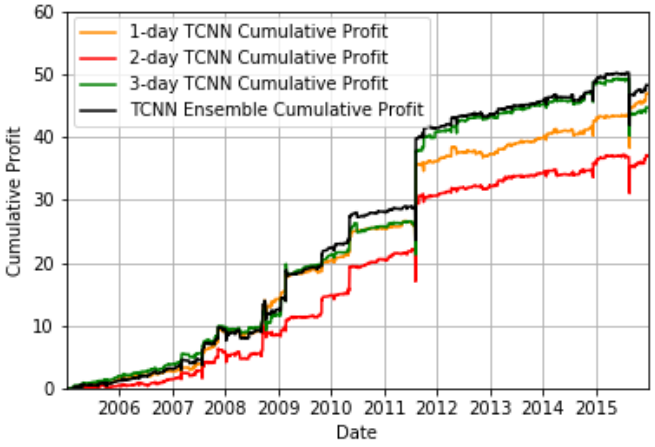

Figure 4.8: Cumulative profit (as a multiple of starting wealth, per Table 4.9) generated by the various TCNN models between Jan-2005 and Dec-2015, in the absence of friction costs. The models are steadily profitable, with occasional spikes related to major events. Drawdowns are infrequent and of limited scale. We do note however that the performance of the strategy has tapered off somewhat since mid-2011. Unlike the physical sciences, lucrative statistical abnormalities in financial markets are arbitraged away once discovered. We are likely not alone in uncovering such patterns through deep learning.

wide relief over Nomura's decision to buy Lehman's operations - and presented the last opportunity to sell before the nosedive of the Great Financial Crisis in late 2008. Despite having no information about world news in their technical dataset, the models were capable of both inferring crucial moments in history, and timing trading decisions around them.

Figure 4.8 presents the model's profitability over time to highlight the relative steadiness of convolution's performance in identifying stock market patterns, when the decisions are generated by TCNNs and their ensemble. Table 4.9 translates this performance into compounded annual returns and Sharpe ratios under various assumptions for friction. Even in the absence of tight execution (average trading cost of 0.25% from the mid-market price), the models remain highly profitable. This sensitivity analysis does nevertheless highlight the importance of good execution in

Table 4.9: Compound Annual Growth Rate (CAGR, in %) and Sharpe ratio of the TCNN models under various assumptions for the cost of trading. The TCNN ensemble and 1-day TCNN are optimal choices for return and risk-adjusted return maximisation, respectively.

Friction Cost	No friction			0.10% per transaction			0.25% per transaction		
Model	Profit	CAGR	Sharpe	Profit	CAGR	Sharpe	Profit	CAGR	Sharpe
1-day TCNN	46.9	42.15	**8.04**	32.8	37.72	**7.16**	11.7	25.98	**4.75**
2-day TCNN	36.9	39.16	7.81	22.8	33.41	6.61	1.7	9.52	1.65
3-day TCNN	44.6	41.50	5.95	30.5	36.84	5.59	9.4	23.71	3.49
TCNN Ensemble	**48.2**	**42.50**	6.57	**34.1**	**38.20**	5.86	**13.0**	**27.13**	4.08

any real-world deployment of algorithmic trading: the TCNN ensemble can only just break even if the per-transaction cost rises to 0.35%.

4.4.7 Interpretable Feature Extraction

The convolutional filters learned by the network provide a basis for feature extraction. In particular, the convolutional layer's filters define patches whose cross-correlation with the original input data was informative in minimising both in-sample and out-of-sample categorical cross-entropy. We produce a mosaic of these filters as Hinton diagrams[7] (Figure 4.9) and visualise them in the language of technical analysis as candlestick patterns (Figures 4.10 and 4.11), cross-correlational templates whose occurrence is informative for financial time series forecasting. Unlike technical patterns however, these templates have no set meaning: the purpose of individual neurons in a convolutional layer is not readily interpretable.

4.5 Summary

Our results present, to our knowledge, the first rigorous statistical evaluation of candlestick patterns in time series analysis, using normalised signal cross-correlation to identify pattern matches. We find little evidence of predictive prowess in the stan-

[7]Hinton diagrams provide a means of visualising numerical values in a matrix. The area occupied by a square is proportional to the value's magnitude, and the sign of the matrix entry is colour-coded (white for positive values, black for negative values).

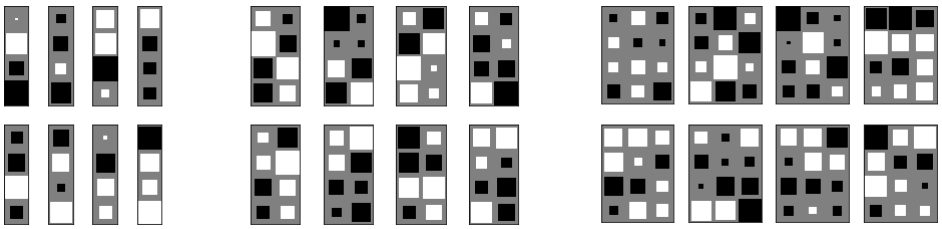

Figure 4.9: Weight-space visualisation as Hinton diagrams for the 24 cross-correlational filters learned from the first layer of each CNN (8 per constituent model).

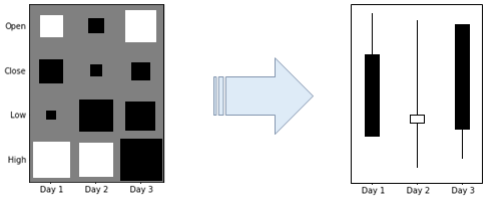

Figure 4.10: Hinton diagram of the sixth cross-correlational filter learned in the first layer of the 3-day CNN. The relative values of the standardised open, close, low and high for each column in the filter define, in a chartist sense, a specific candlestick sequence (or patch thereof, in instances where the filter's open or close is incompatible with the high-low range) which the neural network extracted as informative for time series forecasting.

dard chartist pictograms, and suspect that the enduring quality of such practices owes much to their subjective and hitherto unverified nature. Our findings in Chapter 3 did nevertheless suggest that price history does contain some predictive information, and indeed much of quantitative finance practice relies on elements of technical pattern recognition (e.g., momentum-tracking) for its success. Through a deep learning lens, technical analysis is merely an arbitrary and incorrect specification of the feature-extractive early layers of a neural network. Within relatively shallow architectures, learning effective filters for technical data improves accuracy significantly while also providing an interpretable replacement for chartism's visual aids. Thresholding and deep ensembles yield a robust framework for systematic decision making in financial

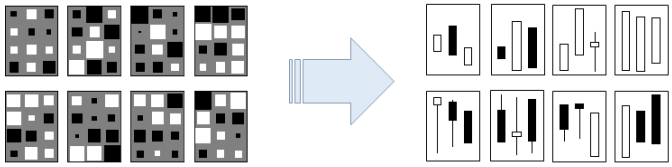

Figure 4.11: Candlestick pattern translation of the cross-correlational filter mosaic for the 3-day CNN. Unlike the chartist patterns of Figure 4.2, we can not draw any immediate directional conclusions from the occurrence of these patterns in time series data. The convolutional layer of the network merely found these patterns to be the most informative.

markets, further enhancing performance. The emergence of hybrid architectures - possessing both convolutional layers for feature-extraction and recurrent layers for memorising long-term dependencies - presents fresh opportunities for research in this space.

Chapter 5

Market Making in the Presence of Adverse Selection

5.1 Introduction

The models we've explored thus far have tackled the complex tasks of feature selection and extraction, in the context of forecasting the future direction of asset prices. The Automatic Relevance Determination kernels of Chapter 3 provided a basis for ranking features and pruning irrelevant inputs, whereas the deep learning approach of Chapter 4 extracted patterns from raw data, learning price data's salient features through convolution. In both instances, the purpose is fundamentally predictive: the aim is to infer what the future price of an asset will be, based on currently available data. Guided by this knowledge, the liquidity taker can devise their strategy, determining whether to place buy or sell orders at prevailing market prices, known as *market orders*.

The completion of every trade relies on a matching process: the buy or sell *market order* of a liquidity taker is matched against the offer or bid *limit order* of a market maker. The market maker's *limit orders* identify the prices at which they are willing to buy and sell a fixed quantity of an asset. The *bid-offer spread* between these prices, locked in as profit for every contemporaneous buy and sell order of identical size, exists to compensate dealers for their two main sources of risk: *inventory risk*

and *adverse selection risk* (Glosten and Harris, 1988).

Inventory risk stems from variations in the value of the asset held, and adverse selection refers to instances where knowledgeable counterparties take advantage of information asymmetry to build an (eventually) lucrative position, to the detriment of the market maker. Stochastic control models pioneered by Ho and Stoll (1981) and further developed by Avellaneda and Stoikov (2008) and Guéant et al (2013) provide a robust framework for dynamically adjusting bid and offer prices to manage asset-specific inventory risk. Adverse selection, however, is an exogenous client-specific risk. As such, its study benefits from the adoption of data-driven methods drawn from machine learning and pattern recognition.

Over-the-counter (OTC) markets such as foreign exchange differ materially from the on-exchange equity markets studied in the stochastic control literature. Liquidity providers in an OTC setting can show different quotes to each client, relying on their knowledge of the counterparty's price elasticity to optimise bid and offer levels. This chapter explores Bayesian representations of client behaviour, in an effort to quantify adverse selection and thereby adjust bid and offer prices punitively for counterparties whose trades are deemed opportunistic.

We begin by describing the inventory control model (Section 5.2) and datasets (Section 5.3) employed throughout this chapter. After some preliminary data exploration using kernel density estimation (Section 5.4.1) and correlation analysis (Section 5.4.2), we produce multivariate Automatic Relevance Determination (ARD) Gaussian Process representations of each counterparty (Section 5.4.3), pairing trade activity with market data metrics such as time of day, volatility, short-term returns and bid-offer spread. We show evidence of predictability in client responses to market conditions, and construct a matrix of Bhattacharyya distances to measure the similarity of connections to each other. In the interest of scalability, we reduce the problem's dimensionality by clustering counterparties with a community detection technique

built on Bayesian non-Negative Matrix Factorisation (Section 5.4.4). Finally, our approach is modular: it can readily be integrated into an augmented Hamilton-Jacobi-Bellman (HJB) framework to control for both inventory risk and adverse selection risk simultaneously. We provide simulations (Section 5.4.5-5.4.6) to estimate the performance of three strategies: static bid-offer, inventory-adjusted bid-offer and inventory- and counterparty-adjusted bid-offer, and demonstrate the benefits of learning the behaviour of liquidity takers.

5.2 Model

We begin with a summary of the Avellaneda and Stoikov model for managing inventory risk (Avellaneda and Stoikov, 2008), defining the value function of the market maker before drawing on results in control theory literature to solve for utility maximisation.

5.2.1 The Market Maker Value Function

We define the market maker's bid-mid and mid-offer price gaps as

$$\delta^b = s - p^b \tag{5.1}$$

$$\delta^a = p^a - s \tag{5.2}$$

where s is an exogeneously-defined mid-market price, and p^b and p^a are the market maker's bid and ask prices. Their model follows Garman (1976) in the assumption that a market maker's buy and sell limit orders are filled at a Poisson rate $\lambda^b(\delta^b)$ and $\lambda^a(\delta^a)$, decreasing functions of their respective arguments.

Using the utility framework of Ho and Stoll, they parametrise the following value function for the market maker:

$$v(x, s, q, t) = E_t[-\exp(-\gamma(x + qS_T))] \tag{5.3}$$

where x is the market maker's wealth, s is the mid price, q is the current inventory (which can be negative, when shorting), γ is a risk aversion factor and t is time. They then define the market maker's *reservation* or *indifference* bid and ask prices r^b and r^a using the relations

$$v(x - r^b(s,q,t), s, q+1, t) = v(x,s,q,t) \tag{5.4}$$

$$v(x + r^a(s,q,t), s, q-1, t) = v(x,s,q,t) \tag{5.5}$$

The average of r^b and r^a in Equations (5.4) and (5.5) provides the market maker's indifference price, and constitutes a divergence from the fair mid that decreases monotonically with inventory level.

5.2.2 The Hamilton-Jacobi-Bellman Equation

The dealer's objective is given by the value function:

$$u(x,s,q,t) = \max_{\delta^b,\delta^a} E_t[-\exp(-\gamma(x+qS_T))] \tag{5.6}$$

where δ^b and δ^a can be thought of as feedback controls. A key finding in the works of Ho and Stoll was the application of the dynamic programming principle to show that the function u solves the following Hamilton-Jacobi-Bellman equation:

$$\begin{cases} u_t + \frac{1}{2}\sigma^2 u_{ss} + \max\limits_{\delta^b} \lambda^b(\delta^b)\big[u(s,x-s+\delta^b,q+1,t) - u(s,x,q,t)\big] \\ + \max\limits_{\delta^a} \lambda^a(\delta^a)\big[u(s,x+s+\delta^a,q-1,t) - u(s,x,q,t)\big] = 0 \\ u(s,x,q,T) = -\exp(-\gamma(x+qs)) \end{cases} \tag{5.7}$$

Avellaneda and Stoikov's choice of an exponential utility allows a simplification of the problem. They postulate that there exists a function θ such that

$$u(x,s,q,t) = -\exp(-\gamma x)\exp(-\gamma\theta(s,q,t)) \tag{5.8}$$

Direct substitution of Equation (5.8) into Equation (5.7) yields the following system for θ:

$$\begin{cases} \theta_t + \frac{1}{2}\sigma^2\theta_{ss} - \frac{1}{2}\sigma^2\gamma\theta_s^2 + \max\limits_{\delta^b}\left[\frac{\lambda^b(\delta^b)}{\gamma}\left[1-\exp(\gamma(s-\delta^b-r^b))\right]\right] \\ + \max\limits_{\delta^a}\left[\frac{\lambda^a(\delta^a)}{\gamma}\left[1-\exp(-\gamma(s+\delta^a-r^a))\right]\right] = 0 \\ \theta(s,q,T) = qs \end{cases} \tag{5.9}$$

Furthermore, applying Equation (5.8) to the reservation bid and ask prices given by Equations (5.4) and (5.5) yields the relations

$$\delta^b_{HJB} = s - r^b(s,q,t) + \frac{1}{\gamma}\log\left(1 - \gamma\frac{\lambda^b(\delta^b)}{(\partial\lambda^b/\partial\delta^b)(\delta^b)}\right) \tag{5.10}$$

and

$$\delta^a_{HJB} = r^a(s,q,t) - s + \frac{1}{\gamma}\log\left(1 - \gamma\frac{\lambda^a(\delta^a)}{(\partial\lambda^a/\partial\delta^a)(\delta^a)}\right) \tag{5.11}$$

The HJB solution to inventory risk management can be viewed as a two-step process: first solve for $r^b(s,q,t)$ and $r^a(s,q,t)$ in the PDE given by Equation (5.9), then solve Equations (5.10) and (5.11) to obtain the optimal distance $\delta^b_{HJB}(s,q,t)$ and $\delta^a_{HJB}(s,q,t)$ between the mid price and the optimal bid and ask quotes.

Control theory is inherently suited to the feedback dependency of inventory management. In contrast to the endogenous nature of inventory risk, adverse selection produces exogenous shocks to the market maker's business activity, driven by the behaviour of individual counterparties. As such, we believe its treatment benefits from the adoption of data-driven methods drawn from machine learning and pattern recognition. In the following sections of this chapter, we present the data and probabilistic framework used to detect likely instances of adverse selection and demonstrate its effectiveness in simulations.

5.3 Data

In this section we describe both the market maker dataset and the market metrics derived from EURUSD tick data for the same time period. Further details on data provenance are included in Appendix A.3.

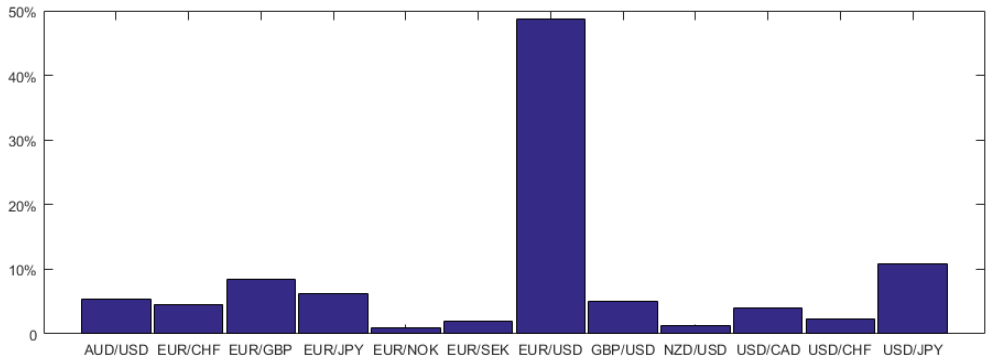

Figure 5.1: Histogram of tradecount by currency in the dataset for January 2013, N=74,958. 48.7% of trade orders are in EURUSD, motivating our currency choice.

5.3.1 Raw Datasets

The market maker dataset is a list of FX trades executed by a leading market maker in 2013. The complete dataset counts 916,683 trades split across 12 currencies, with EURUSD by far the most actively traded pair. For each trade, the raw dataset provides the price, the timestamp, the currency pair, deal size, whether it was a buy or a sell, as well as both client ID and broker ID numbers. In the analysis that follows, we restrict our attention to EURUSD trades taking place in January 2013 - this dataset's dealcount is 35,001, large enough to remain representative. Figure 5.1 provides the full currency breakdown for the January 2013 subset.

We supplement the market maker data with a EURUSD tick dataset providing high-frequency bid and offer prices for 2013. The tick data for January 2013 alone contains 10,133,449 entries.

5.3.2 Derived Features

Combining these two datasets allows us to compute return and volatility metrics for each transaction involving the market maker, and use this information to build profiles of their counterparties. We produce 7 market-derived features to better understand

client sensitivity.

- The time of day feature returns the decimal part of each timestamp, to track intraday seasonality effects.

- Under existing models of adverse selection, the volatility response ranks as a significant determinant of adverse selection (Van Ness et al, 2001). Two volatility metrics were constructed to account for tradeflow reacting to volatility ('lagging volatility') and tradeflow anticipating volatility ('leading volatility'). In particular, a significant bias in the latter may indicate an information advantage as the client's activity foreshadows rapid change in market conditions.

  1-minute Lagging Volatility(t) =

$$\sqrt{\frac{1}{N_t}\sum_{t\in[t-1m,t]}\Big(\text{EURUSD(mid)}_t - \mu_{t\in[t-1m,t]}\Big)^2} \tag{5.12}$$

  1-minute Leading Volatility(t) =

$$\sqrt{\frac{1}{N_t}\sum_{t\in[t,t+1m]}\Big(\text{EURUSD(mid)}_t - \mu_{t\in[t,t+1m]}\Big)^2} \tag{5.13}$$

  where $\mu_{t\in[t_0,t_1]}$ refers to the mean midprice of EURUSD over the interval $[t_0, t_1]$, derived from our EURUSD tick dataset.

- Three return metrics were devised to search for systematic bias in a client's returns at various timescales. A systematic positive bias in an individual connection's trade returns would provide strong evidence of adverse selection.

$$\text{1-minute Return(t)} = \frac{\text{EURUSD(mid)}_{t+1m} - \text{EURUSD(mid)}_t}{\text{EURUSD(mid)}_t} \tag{5.14}$$

$$\text{1-hour Return(t)} = \frac{\text{EURUSD(mid)}_{t+1h} - \text{EURUSD(mid)}_t}{\text{EURUSD(mid)}_t} \tag{5.15}$$

$$\text{1-day Return(t)} = \frac{\text{EURUSD(mid)}_{t+1d} - \text{EURUSD(mid)}_t}{\text{EURUSD(mid)}_t} \tag{5.16}$$

- Finally, we include bid-offer spreads as an inverse proxy for market-wide volumes, a metric we could not obtain without an exhaustive (and prohibitively costly) analysis of every EURUSD transaction occuring on every FX trading platform.

$$\text{Bid-Offer Spread(t)} = \text{EURUSD(offer)}_t - \text{EURUSD(bid)}_t \tag{5.17}$$

5.4 Results

In this section, we detail the results of our analysis. After preliminary client data exploration using kernel density estimators (KDEs), we determine significant correlations within the dataset and evaluate the ARD framework's ability to identify salience. With Gaussian Processes constructed for each connection, we define the Bhattacharyya distance between GPs and derive a similarity matrix on which to apply community detection techniques, looking for counterparty clusters bearing similarities. Finally, we measure the returns persistence in each counterparty's dynamics and compare our findings to the solution of Avellaneda and Stoikov via numerical simulations.

5.4.1 Kernel Density Estimation

Univariate KDEs offer a first glimpse into aggregate client behaviour, by viewing tradeflow as a point process and assigning a Gaussian density to each datum. The resulting distribution provides a probabilistic interpretation of client activity, and yields intuitive findings concerning intraday seasonality and short-term return bias.

- Aggregate client activity peaks twice in the day, at 10:00 GMT and 15:00 GMT, reflecting high points in tradeflow for London and New York respectively. These spikes appear in Figure 5.2.

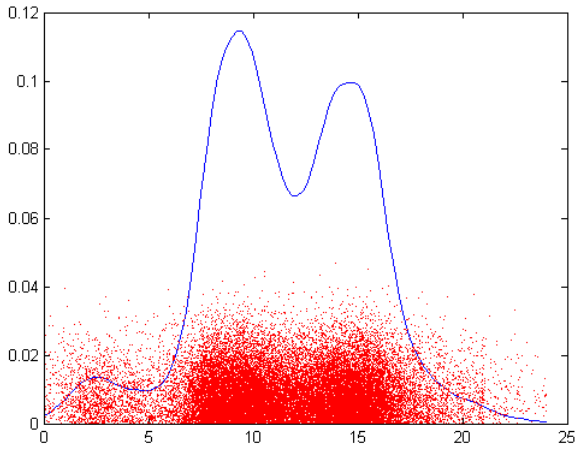

Figure 5.2: Kernel density estimator of January 2013 trades in EURUSD as a function of time of day, with scatter of individual trades, n=35,001. The x-axis follows Greenwich Mean Time. For graphical clarity, each datapoint was jittered vertically with Gaussian noise.

- Client 1-min returns tend to be slightly negative, reflecting that price takers must pay half the spread to transact. Their trades execute at a small disadvantage to mid-market, and are therefore biased to remain that way in the very short term (Figure 5.3).

5.4.2 Correlation Analysis

A significant drawback of using kernel density estimation lies in its application of equal Gaussian densities for each trade, regardless of size. In reality, meaningful learning of client behaviour would require forecasting the circumstances in which their largest orders are placed. We conduct two separate correlation analyses, to account for the fact that certain features are sensitive to the sign of trade volumes (that is, whether the order is a buy or a sell).

- Correlation between each market return metric and signed deal size. The results are provided in Table 5.1.

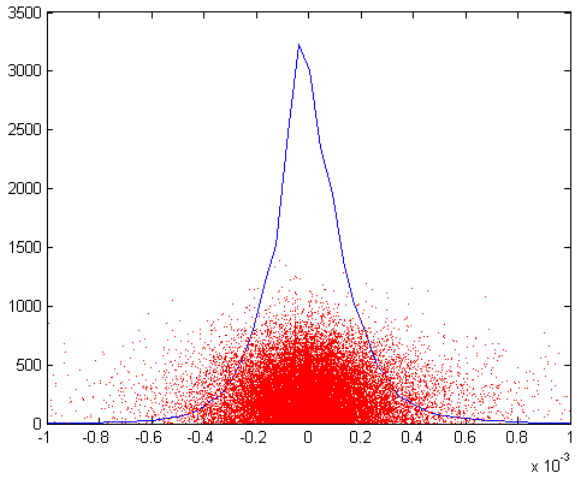

Figure 5.3: Kernel density estimator of January 2013 trades in EURUSD as a function of 1-minute returns, with scatter of individual trades, n=35,001.

Table 5.1: Correlation between Return Metrics and Signed Trade Volume, N=35,001. We highlight in bold the input features exhibiting Pearson p-values below 0.001.

	Correlation		p-value	
Feature	Pearson	Spearman	Pearson	Spearman
1-min Return	**−0.0379**	**−0.0405**	**< 0.0001**	**< 0.0001**
1-hour Return	−0.0061	−0.0178	0.2516	0.0009
1-day Return	+0.0032	−0.0024	0.5494	0.6589

- Correlation between volatility, time of day, bid-offer spread and absolute deal size. The results are provided in Table 5.2.

Applying t-tests for significant correlations in our dataset, it quickly emerges that client activity is hugely sensitive to market movements, signalled by numerous p-values under 0.001 by several orders of magnitude in Tables 5.1 and 5.2.

5.4.3 ARD-GP Representation of Counterparty Behaviour

The discovery of multiple significant correlations in our dataset motivates the search for a robust, multivariate probabilistic technique for modelling each connection's dy-

Table 5.2: Correlation between Time of Day, Volatility, Bid-Offer Spread and Absolute Trade Volume, N=35,001. We highlight in bold the input features exhibiting Pearson p-values below 0.001.

	Correlation		p-value	
Feature	Pearson	Spearman	Pearson	Spearman
Time of Day	**+0.0366**	**+0.0134**	**< 0.0001**	**< 0.0001**
1-min Lagging Volatility	**+0.0345**	**+0.0801**	**< 0.0001**	**< 0.0001**
1-min Leading Volatility	**+0.0478**	**+0.0570**	**< 0.0001**	**< 0.0001**
Bid-Offer Spread	−0.0047	+0.0247	0.3835	< 0.0001

Table 5.3: Relevance across all polarity-sensitive features measured on the January 2013 dataset, N=35,001.

	Mean Relevance		Spearman	
Feature	Score	Ratio	Correlation	p-value
1-min Return	0.2	4.1	−0.0405	< 0.0001
1-hour Return	0.2	6.4	−0.0178	0.0009
1-day Return	0.2	4.5	−0.0024	0.6589

namics. ARD Gaussian Processes learn the respective relevances of each feature, and should therefore rank highest those features whose correlation is most significant.

We produce an ARD Gaussian Process for each individual connection, and provide in Tables 5.3 and 5.4 the mean Relevance Score and Relevance Ratio for each feature, alongside its Spearman correlation and p-value. Separate tables were necessary to reflect that certain features were regressed on, and correlate with, signed trade volume (features we term polarity-sensitive) whereas others were measured against absolute trade volume (which we term polarity-insensitive features).

The ARD ranking tallies broadly with the results from the earlier correlation analysis on absolute deal sizes, but defies our expectations for the 1-min, 1-hour and 1-day feature set. It nevertheless allows us to produce meaningful bivariate visualisations of client sensitivity to its dominant features. As an example of the

Table 5.4: Relevance across all polarity-insensitive features measured on the January 2013 dataset, N=35,001. We highlight in bold the input features exhibiting Relevance Ratios great than 10^2, per Section 2.4.6.

	Mean Relevance		Spearman	
Feature	Score	Ratio	Correlation	p-value
Time of Day	5.8	11	+0.0134	< 0.0001
1-min Lagging Volatility	$\mathbf{1.2 \times 10^6}$	$\mathbf{2.3 \times 10^6}$	**+0.0801**	$\mathbf{< 0.0001}$
1-min Leading Volatility	$\mathbf{7.4 \times 10^6}$	$\mathbf{1.3 \times 10^7}$	**+0.0570**	$\mathbf{< 0.0001}$
Bid-Offer Spread	0.56	1.0	+0.0247	< 0.0001

framework's ability to represent behaviour, we provide heatmaps of broker connection #1's response to its most salient features in Figures 5.4 and 5.5. Heatmaps that diverge materially from the market norm - for example, where signed deal volume correlates positively with 1-hour and 1-day returns - would be indicative of outlier behaviour and warrant further inspection.

5.4.4 Community Detection

Having constructed Gaussian Processes for each connection, we wish to measure the similarity between behaviours in order to cluster counterparties and apply different adjustments to the inventory-optimal bid-offer price, based on the cluster's track record. In light of the intuitive salience of returns history in identifying adverse selectors, we have adopted the 1-hour and 1-day return features as our basis for community detection.

5.4.4.1 Bhattacharyya Distance

Detection of clusters requires the creation of a matrix of distances that capture the pairwise overlap between any two connections (Psorakis et al, 2012). GPs by definition have a probabilistic interpretation, as a draw from a probability distribution over functions. Amidst the wide range of available symmetric distance measures, we

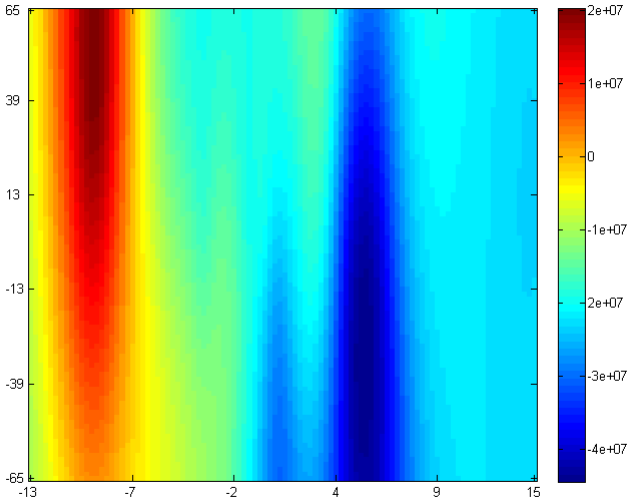

Figure 5.4: Signed Deal Volume for broker connection #1 as a function of ensuing 1-minute Returns (x-axis, in basis points) and 1-hour Returns (y-axis, in basis points). Large buy orders correlate with negative 1-min returns (peak red area corresponding to -0.09% return on the x-axis) and large sell orders correlate with positive 1-min returns (peak blue area corresponding to +0.05% on the x-axis). This matches the earlier finding of a strong negative correlation between signed deal size and 1-minute returns.

choose the Bhattacharyya distance as our measure of similarity between two probability distributions f and g, defined as

$$D_B(f(\mathbf{x}), g(\mathbf{x})) = -\log(BC(f(\mathbf{x}), g(\mathbf{x}))) \tag{5.18}$$

where $BC(f(\mathbf{x}),g(\mathbf{x}))$ is the Bhattacharyya Coefficient given by

$$BC(f(\mathbf{x}), g(\mathbf{x})) = \int \sqrt{f(\mathbf{x})g(\mathbf{x})}d\mathbf{x} \tag{5.19}$$

In our case, f and g are GPs defined by their mean and covariance functions. Detrending during data preprocessing ensures that the distributions have zero mean, $\mu_f = \mu_g = 0$. This simplifies the distance calculation to

$$D_B(f(\mathbf{x}), g(\mathbf{x})) = \frac{1}{2}\log\left(\frac{|K|}{\sqrt{|K_f||K_g|}}\right) \tag{5.20}$$

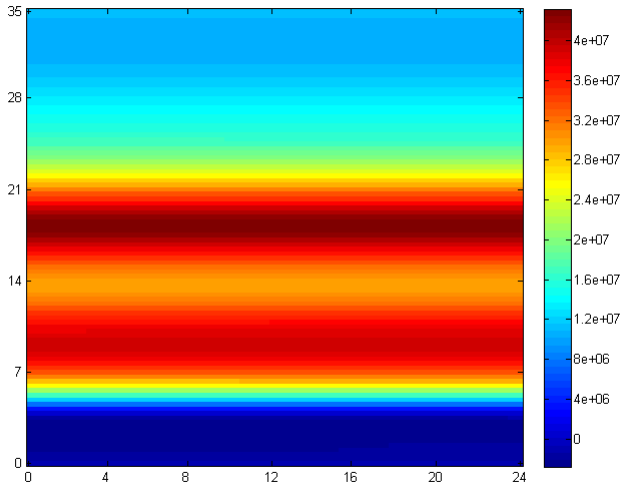

Figure 5.5: Absolute Deal Volume for broker connection #1 as a function of Time of Day (x-axis, in hours) and 1-minute Leading Volatility (y-axis, in percentage points). In the case of this connection, time of day was insignificant (correlation of +0.0128 with a p-value of 0.4137), with leading volatility dominating the forecast distribution (correlation of +0.0435 with a p-value of 0.0054, below the significance level of 0.01).

where K_f and K_g are the covariance matrices for the two GPs f and g, and

$$K = \frac{K_f + K_g}{2} \tag{5.21}$$

The Bhattacharyya distance for GPs is therefore only dependent on the choice of kernel and optimised hyperparameters. Estimation of each connection's GP hyperparameters is achieved by maximising the marginal likelihood of the data, and allows us to compute covariance matrices for each connection on a common input scale. Formation of a Bhattacharyya similarity structure follows straightforwardly from Equation (5.20), calculating the inverse distance between each GP pair.

5.4.4.2 Bayesian Non-Negative Matrix Factorisation

Inverse Bhattacharyya distances between GPs allows us to map the client dataset to a relational space, where every pair i, j of connections is given a similarity value s_{ij}.

The resulting similarity matrix $\mathbf{S}$ is input as an adjacency structure to a Bayesian non-negative matrix factorisation (NMF) scheme (Psorakis et al, 2011), a method with a history of successful application to a diverse set of community detection challenges ranging from social network structure in ecological systems to maritime anomaly detection (Smith et al, 2014).

In Bayesian NMF, communities are viewed as explanatory latent variables for the observed link weights in the adjacency structure. The latent clustering is produced by a factorisation $\mathbf{S} \simeq \mathbf{WH}$, $\mathbf{S} \in \mathbb{R}^{N \times N}, \mathbf{W} \in \mathbb{R}^{N \times K}, \mathbf{H} \in \mathbb{R}^{K \times N}$ where $\mathbf{W}$ and $\mathbf{H}$ are non-negative. The inner rank K and factor elements w_{ik}, h_{kj} are derived by Maximum a Posteriori inference, and provide a soft, probabilistic membership score for clients to each community. The framework is therefore capable of recognising communities that overlap.

We apply Bayesian NMF to the January 2013 EURUSD dataset comprised of 25 broker connections, and provide the resulting classification in Table 5.5. Bayesian NMF finds 2 classes: Cluster 1 comprises the majority (20 of 25) of counterparties, and Cluster 2 contains the outliers.

5.4.4.3 Comparison to Benchmark Classifiers

For completeness, we compare the results to a hierarchical clustering dendrogram approach and k-medoids, with k set equal to the number of communities found by Bayesian NMF.

We define briefly the method of construction for dendrograms:

- The initial partition P_0 assigns each node to its own class.
- The final partition P_{n-1} (the conjoint partition) is one all-inclusive node class.
- P_{k+1} is defined from P_k by uniting a single pair of subsets in P_k. Union is pursued on the basis of closeness, where we minimise the maximum distance between any pair within the joined subsets (we minimise the subset 'diameter').

Table 5.5: Bayesian NMF clustering on the January 2013 dataset across 25 broker connections.

Bayesian NMF Cluster	Nodes Assigned
Cluster 1	1,2,3,4,5,6,7,8,9,10,11,12,15,16,17,19,20,21,23,24
Cluster 2	13,14,18,22,25

Table 5.6: 2-medoids clustering on the January 2013 dataset across 25 broker connections.

2-medoids Cluster Centroid	Nodes Assigned
Centroid 20	1,2,3,4,5,6,7,8,9,10,11,12,15,16,17,19,20,21,23,24
Centroid 13	13,14,18,22,25

The resulting tree is provided in Figure 5.6. In line with Bayesian NMF's discovery of 2 communities, we test our matrix of Bhattacharyya distances on a 2-medoids clustering algorithm, and report its results in Table 5.6. In both cases, Bayesian NMF finds identical classes with two additional benefits. Firstly, it determines the appropriate number of classes k, obviating the need to hand-tune. Secondly, its soft partitioning allows the identification of nodes at the intersection of the classes. Summary statistics defining each cluster's behaviour are provided in Table 5.7. The mean 1-hour and 1-day profitability of trades from Cluster 1 is close to zero (-0.2 bp gains on 1-hour returns, 0.5 bp gains on 1-day returns), indicative of normal counterparties with no systematic edge over the market maker. In contrast, trades from Cluster 2 register on average 3.2 bp gains within an hour and 30.1 bp gains within a day, highlighting counterparties with an abnormal propensity towards successful trades.[1]

[1]We can never be certain that individual successes are the result of knowledge rather than luck. In an extension of this framework, we would apply community detection to lengthier rolling windows of the market maker's dataset, and measure not just the mean return μ_d of a counterparty, but also the standard deviation σ_d of those returns. Clustering on the basis of the ratio $\frac{\mu_d}{\sigma_d}$ would permit us to even more reliably separate informed counterparties from the merely lucky.

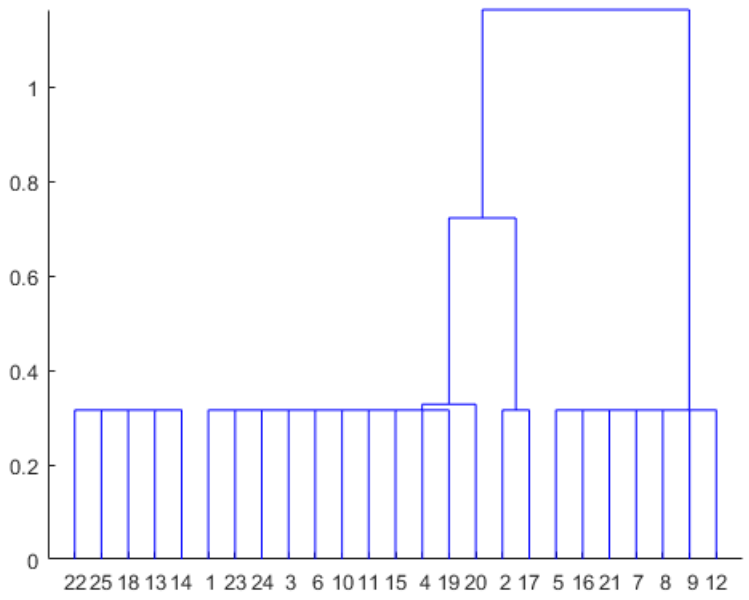

Figure 5.6: Dendrogram of the broker connections. Connections 1,2,3,4,5,6,7,8,9,10, 11,12,15,16,17,19,20,21,23 and 24 are all part of the same subset. These connections correspond exactly to Cluster 1 from Bayesian NMF. The remaining nodes form a separate partition in the dendrogram, and can be thought of as a single alternate cluster, matching Cluster 2 from Bayesian NMF.

5.4.5 Integration with Inventory Control Frameworks

Adverse selection represents opportunistic arbitrage of information asymmetry. We assume that by definition adverse selectors will produce persistently high returns. The timeframe over which an adverse selector generates positive returns has shortened considerably in the high-frequency trading era: once measured in weeks, it is now more commonly sought intraday, even over fractions of a second (O'Hara, 2015). As a conservative estimate, we consider the mean 1-day percentage return μ_d of trades for each connection. Any counterparty with systematic, significant 1-day returns may be benefitting from information asymmetry. We therefore rank our list of counterparties by 1-day return, and postulate that the inventory-optimal bid-mid and mid-offer solutions δ^b_{HJB} and δ^a_{HJB} from Equations (5.10) and (5.11) should be adjusted on

Table 5.7: Summary statistics for each cluster.

Bayesian NMF Cluster	Mean 1-hour Return	Mean 1-day Return
Cluster 1	-0.2 bp	0.5 bp
Cluster 2	3.2 bp	30.1 bp

a per-connection basis as a monotonically increasing function f of 1-day returns, yielding counterparty-specific solutions

$$\delta^b(i) = f\big(\delta^b_{HJB}, \mu_d(i)\big) \tag{5.22}$$

and

$$\delta^a(i) = f\big(\delta^a_{HJB}, \mu_d(i)\big) \tag{5.23}$$

where the form of f is subjectively defined by the market maker to reflect the severity of the price penalty they wish to impose on adverse selectors (Glosten and Milgrom, 1985).

In the event of an extremely large number of counterparties in the dataset, dimensionality can be first reduced by the community detection methods outlined in Section 5.4.4.

5.4.6 Numerical Simulations

Having devised a framework for adjusting bid-offer spreads to account for adverse selection, we now test the performance of our strategy by measuring the summary statistics of the P&L profile on a virtual market simulation generated identically to Avellaneda and Stoikov's. For ease of comparison, we adopt the same parametrisation as they did for our simulations: initial mid-price $s = 100$, time horizon $T = 1$, volatility $\sigma = 2$, timestep $dt = 0.005$ (i.e. 200 timesteps per simulation), initial inventory $q = 0$, risk aversion $\gamma = 0.01$, and trading intensity parameters $k = 1.5$, $A = 140$.

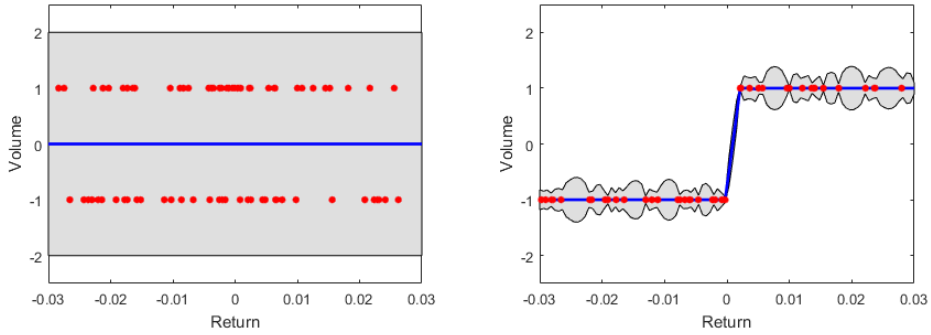

Figure 5.7: Trading functions as learned by GP regression for the normal counterparty (left) and the adverse selector (right) at the end of one simulation run, showing trading volumes (y-axis) as a function of future return (x-axis). Blue lines trace the GP mean function, red dots are individual trade occurrences. Shaded regions correspond to a 95% confidence interval around the GP mean function, and widen out in areas with few observations. The uninformed client's trading volumes do not correlate with ulterior market movements, whereas the adverse selector systematically places correct trades of constant size.

We update the model's state variables at every timestep, and model the arrival rate of buy and sell orders as Poisson Processes with intensities given by:

$$\lambda^a(\delta^a) = A\exp(-k\delta^a) \tag{5.24}$$

and

$$\lambda^b(\delta^b) = A\exp(-k\delta^b) \tag{5.25}$$

For every buy order filled with probability $\lambda^b(\delta^b)dt$, we increase inventory q by 1 and reduce wealth by $s - \delta^b$. For every sell order filled with probability $\lambda^a(\delta^a)dt$, we decrease q by 1 and increase wealth by $s + \delta^a$. At every timestep, the mid-price s is adjusted by $\pm\sigma\sqrt{dt}$.

Where our simulations differ from the approach of Avellaneda and Stoikov is in the inclusion of a normal counterparty and adverse selector with distinctive trading functions f_{norm} and f_{adv}, where we define these functions to be the client-specific mapping of 1-day returns to trading volumes. Normal counterparties present a zero-mean

constant function f_{norm}, reflecting that their trading activity shows no systematic correlation with market movements. By contrast, adverse selectors will tend to buy ahead of price increases and sell ahead of declines, generating a positive bias in their P&L profile to the detriment of the market maker's. An idealisation of their equivalent function f_{adv} will therefore assign positive volumes (buy orders) to regions involving positive returns, and negative volumes (sell orders) to regions involving negative returns. Figure 5.7 shows these functions as learned by our GP-based counterparty modelling. f_{norm} is a constant zero-mean function and f_{adv} has been assigned a piecewise constant relationship, whereby the adverse selector places directionally accurate trades with a fixed volume. The adoption of constant trade volumes (1 unit per buy or sell order) replicates the simulation environment of Avellaneda and Stoikov, and ensures consistency with their results.

We assume that a proportion $\alpha = 0.10$ of trades originate from the adverse selector, with the remaining $1 - \alpha = 0.90$ drawn from the normal counterparty. Adverse selection modelling amounts to sequentially updating the Gaussian Process prior after each trade occurs, building over time a profile for each counterparty from which to tailor client-specific bid and ask prices from the historical performance of their respective distributions.

- Strategy #1: The 'symmetric' strategy applies a static bid-offer spread centred on the mid-price s.

- Strategy #2: The 'inventory' strategy updates the market maker's reservation price r as a function of q at every time-step, and applies a bid-offer centred on r and determined by:

$$r(s,t) = s - q\gamma\sigma^2(T-t) \tag{5.26}$$

$$\delta^a + \delta^b = \gamma\sigma^2(T-t) + \frac{2}{\gamma}\log(1+\frac{\gamma}{k}) \tag{5.27}$$

This approach assumes all trades contain the same level of information, and does not attempt to learn the distinction between counterparties.

- Strategy #3: The 'inventory and counterparty' strategy further refines the 'inventory' strategy by learning counterparty dynamics and measuring the track record of their return distributions. A simple example of counterparty-specific bid-mid and mid-offer spread is specified in Equations (5.28) and (5.29) as monotonically increasing functions of mean 1-day percentage returns $\mu_d(i)$, floored at the Avellaneda and Stoikov HJB solution. These prices serve to deflate the Poisson intensity of trading activity with adverse selectors, mitigating the losses they engender.

$$\delta^b(i) = \delta^b_{HJB} \times \big[\max(1, 100 \times \mu_d(i)\big)] \tag{5.28}$$

$$\delta^a(i) = \delta^a_{HJB} \times \big[\max(1, 100 \times \mu_d(i)\big)] \tag{5.29}$$

The scalar-valued mean 1-day percentage return $\mu_d(i)$ for counterparty i in Equations (5.28) and (5.29) is derived as the return-weighted expectation of the Gaussian Process mean function $m(\mathbf{r})$ in Figure 5.7 mapping 1-day returns to trading volumes. Formally,

$$\mu_d(i) = \mathbb{E}[m(\mathbf{r}) \odot \mathbf{r}] \tag{5.30}$$

where $\odot$ is the Hadamard product.

We run 1000 simulations of each strategy to compare our 'inventory and counterparty' strategy to the 'inventory' and 'symmetric' benchmarks set in Avellaneda and Stoikov (2008). The adverse selection trades still occur in Strategy #3 but arrive at a lower Poisson rate due to the widened bid-offer spread, resulting in greater mean profitability. The summary statistics of each strategy are included in Table 5.8, and

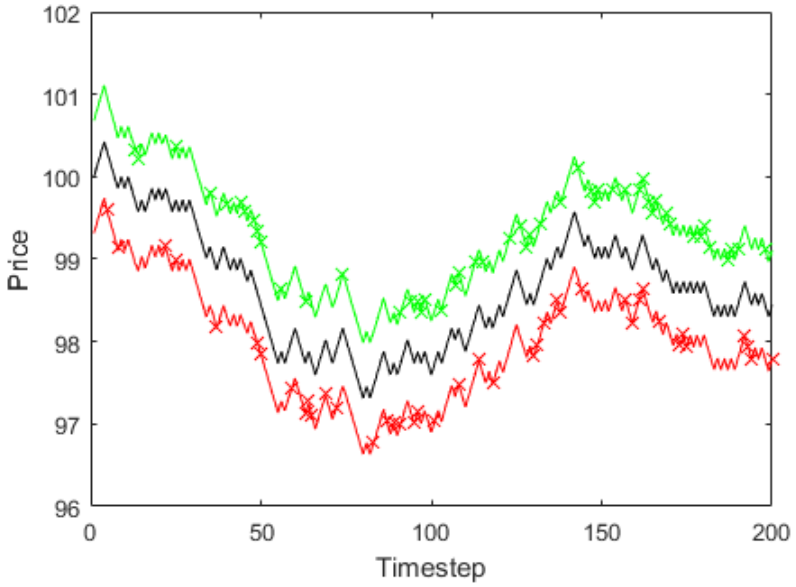

Figure 5.8: Simulation of the mid-price (black line) and inventory-optimal bid and ask quotes (red and green lines respectively). Buy and sell order fills are represented by red and green crosses.

an example of the mid-price and inventory-optimal bid and ask quotes is provided in Figure 5.8.

The main gain in managing inventory risk (Strategy #2 versus Strategy #1, performance histogram and boxplot provided in Figure 5.9) is a reduction in the standard deviation of the P&L profile, emanating from feedback control mechanisms that avoid inventory accumulation and the consequent P&L volatility arising from price fluctuations.

Table 5.8: P&L Profile by Strategy for 1000 simulations.

Strategy	Profit	Std(Profit)	Final q	std(Final q)
Symmetric	46.41	12.80	-0.23	8.64
Inventory	45.02	8.65	0.08	5.09
Inventory & Counterparty	50.89	9.23	0.17	4.71

By contrast, monitoring counterparty risk (Strategy #3 versus Strategy #2, performance histogram and boxplot provided in Figure 5.10) shifts the mean P&L upwards by reducing the frequency of negative shocks caused by adverse selection. In a small number of instances, a lucky streak of successful trades from the normal counterparty early on can lead to the market maker erroneously classifying them as an adverse selector, widening the bid-offer spread to both counterparties and resulting in virtually no interaction for the rest of the simulation. These cases account for the incremental left-tail density of the 'inventory and counterparty' strategy in Figure 5.10: those simulations involve an early misclassification of the normal counterparty, resulting in no market making activity with them and therefore no profitability. We note however that on balance, the benefits of successfully detecting adverse selectors outweigh the opportunity cost of occasionally shunning normal counterparties.

For computational efficiency, we do not update the Gaussian Process hyperparameters after each trade, but rather in batches once per 20 timesteps. This results in 10 Gaussian Process updates per simulation, and allows us to measure the evolution in profitability over the course of each simulation.[2] Figure 5.11 compares the mean profitability of simulations between Strategy #2 and Strategy #3. The two curves begin identically: the market maker initially assigns a zero-mean uninformative prior to both counterparties. As more trades are observed, the Gaussian Process posterior for the adverse selector is updated to reflect their systematic edge. This widens the client-specific bid-mid and mid-ask price gaps shown to the adverse selector, deflating the Poisson intensity of their trades and thereby limiting the losses caused by their order flow.

Changes in client behaviour can readily be addressed by switching from static, batch Gaussian Process regression to an online approach, much as we did for the data-fusing ARD GPs of Section 3.3.5. By defining a rolling window size w over

[2]Our findings are not sensitive to increases in this parameter: a higher update frequency would not produce a noticeable alteration in the evolution of profitability depicted in Figure 5.11.

which to train an adaptive Gaussian Process regression, the market maker will only include data from the recent past, and thereby learn changes in a counterparty's trading function.

A final consideration: there may be value to the market maker in 'paying' to see the behaviour of informed counterparties (Dolgopolov, 2004). If the adverse selector's trades are reliably indicative of future movements (i.e., high $\mu_d(i)$ but also low standard deviation in returns $\sigma_d(i)$), the second-order benefits to the market maker's other trading activities may circumstantially outweigh the first-order losses incurred from dealing with that counterparty. The functional form of f in Equations (5.22) and (5.23) can include return dispersion to model this trade-off, enabling further value extraction from the pursuit of counterparty modelling.

5.5 Summary

Adverse selection poses a client-specific risk hitherto insufficiently addressed by the literature on optimal market making. This chapter proposes a scalable basis for representing client behaviour and quantifying adverse selection via returns persistence. Designed to fit with existing methods for dynamic inventory risk management, our result bridges the state-of-the-art in stochastic control with recent techniques from the machine learning community.

While we have focused in this chapter on countering the threat of adverse selection via machine learning methods, a substantial body of research has looked instead at improving inventory control through purely data-driven algorithms. Indeed while inventory models such as Avellaneda and Stoikov's - and by extension ours - are built on simulations with a limited parameter space (e.g. the risk aversion γ and trading intensity parameters k and A of Section 5.4.6), the latest literature on algorithmic market making applies temporal-difference reinforcement learning on full limit order book data to learn optimal market-making policies (Spooner et al, 2018). The feed-

back mechanisms of control theory are a recurrent theme of reinforcement learning, where ongoing machine learning research provides further perspective on the challenge of inventory management.

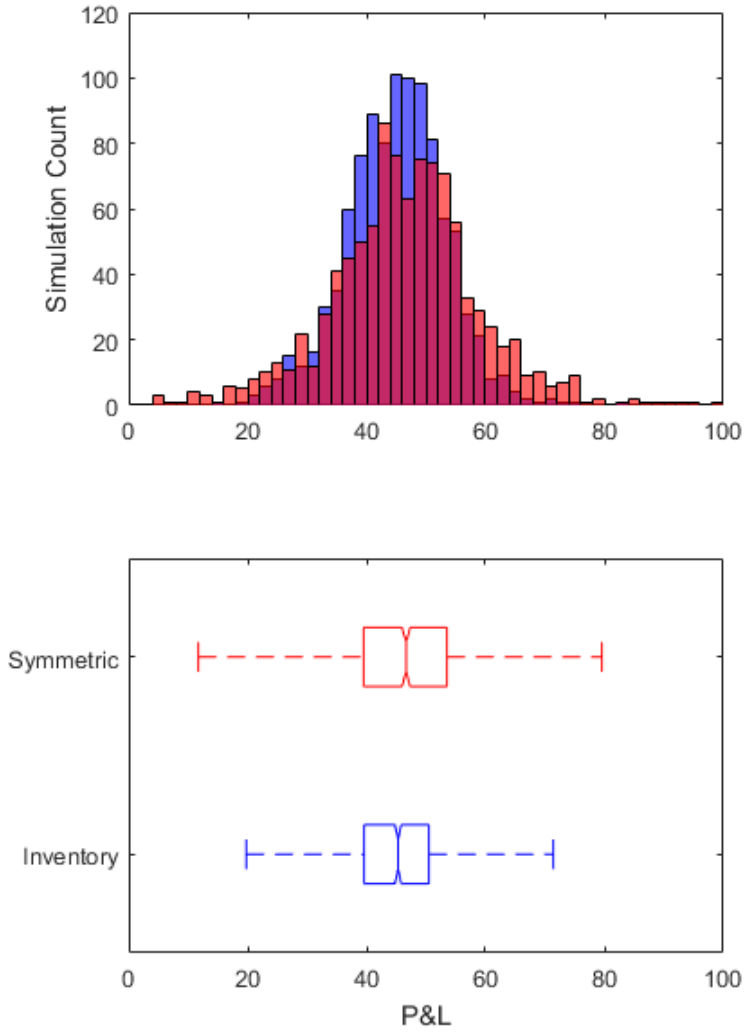

Figure 5.9: Top: Histogram of P&L performance across 1000 simulations for the 'symmetric' (red) and 'inventory' (blue) strategies. Mean performances are similar (46.41 vs. 45.02), but inventory control significantly reduces P&L volatility (standard deviation of 12.80 vs. 8.65). Bottom: Boxplot of P&L performance for the 'symmetric' and 'inventory' strategies; whiskers cover twice the interquartile range. Overlap in the significance notches provides no evidence, at the 95% confidence threshold, that the medians of the 2 distributions are different (Chambers et al, 1983).

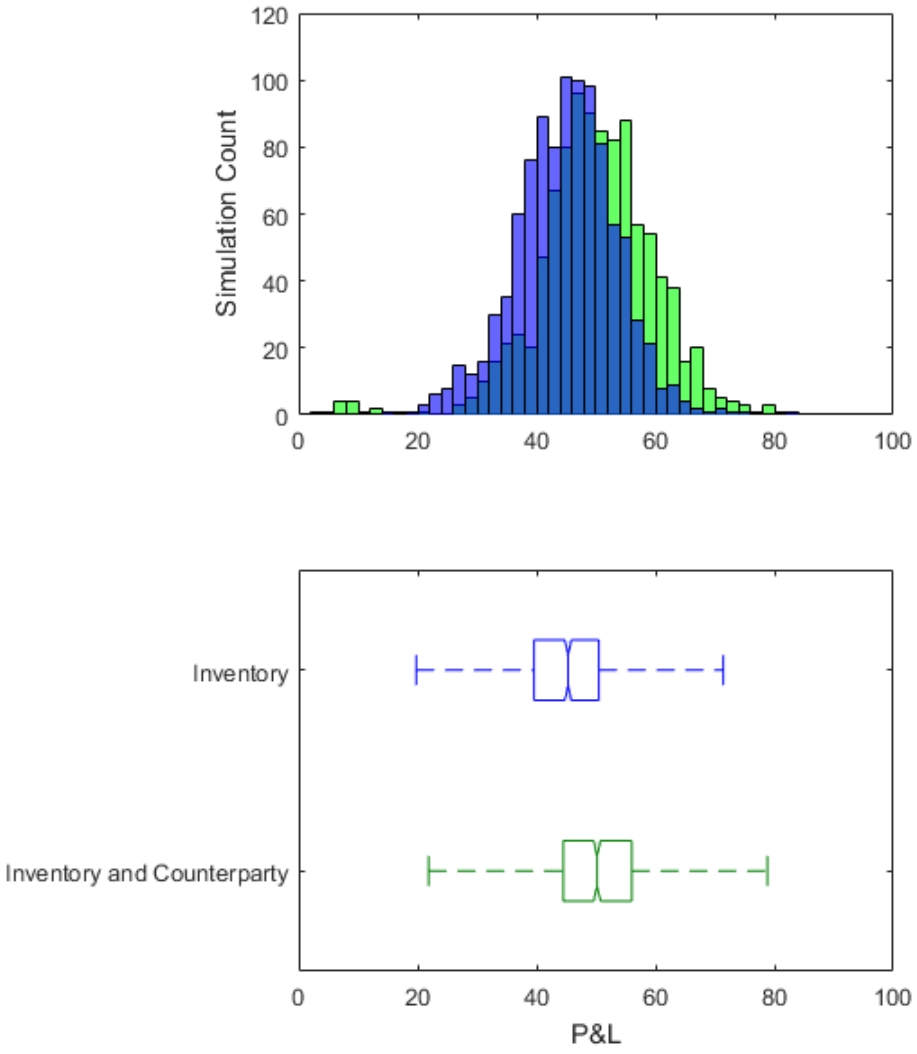

Figure 5.10: Top: Histogram of P&L performance across 1000 simulations for the 'inventory' (blue) and 'inventory and counterparty' (green) strategies. The discouragement of adverse selectors through punitive bid-offer spreads provides a tangible boost to mean profitability (45.02 vs. 50.89), in exchange for a small increase in P&L volatility (8.65 vs 9.23), owing to infrequent cases of erroneous counterparty modelling. Bottom: Boxplot of P&L performance for the 'inventory' and 'inventory and counterparty' strategies; whiskers cover twice the interquartile range. The absence of overlap in the boxplot notches provides evidence at the 95% confidence threshold that the medians of the 2 distributions differ.

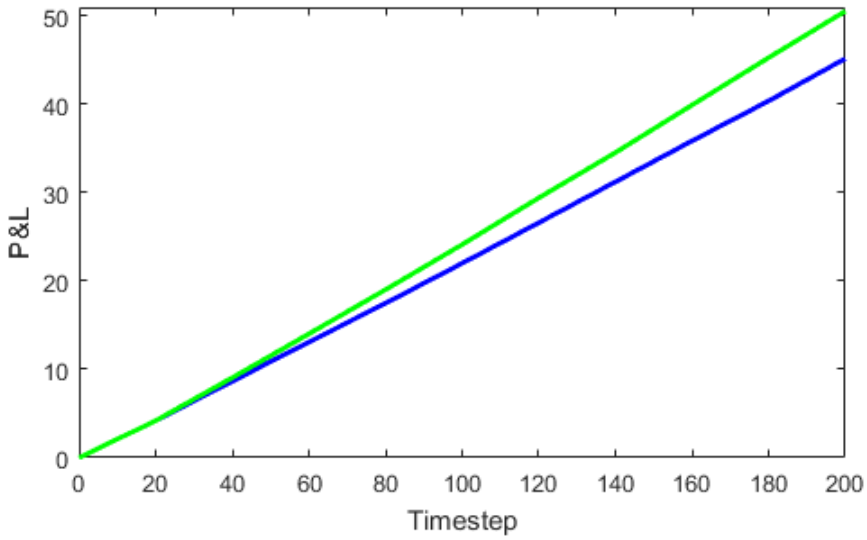

Figure 5.11: Comparison of the mean profitability of Strategy #2 (blue line) and Strategy #3 (green line) over the course of the simulation. The two curves begin identically, but over time diverge as the counterparty-modelling approach identifies the systematic return bias of the adverse selector and widens the bid-offer spread punitively, limiting future costly interactions. By the end of the run, the difference in profitability is significant: 45.02 if bid-offer prices are counterparty-insensitive, 50.89 if bid-offer prices are counterparty-specific.

Chapter 6

Modelling Regulatory Impact

At the heart of the price discovery process lies the notion that, through due diligence, liquidity takers can evaluate the fair worth of an asset, and place market orders that will adjust its price accordingly. The information at their disposal for this process is however incomplete: they can only make use of publicly available information, as the acquisition of material nonpublic information is subject to strict regulation. The ability of regulators to police markets and effectively secure this distinction between public and private information has a profound impact on the price discovery process, and motivates our study of regulatory enforcement and its impact on the information environment.

6.1 Introduction

In 1980, Grossman and Stiglitz defined a seminal paradox that shook the conventionally-held Efficient Market Hypothesis. Succintly, the paradox poses two interwoven questions:

- If markets are already informationally efficient, what reason would a rational agent have to pursue costly due diligence and acquire information?
- If no agent has cause to pursue material information, how do markets become informationally efficient in the first place?

They concluded that an informationally efficient market is impossible, and that gains must exist for investors to be actively pursuing material information. In our study of financial regulation, we concern ourselves with the methods by which that information is acquired and deployed - and more precisely, potential loopholes introduced by the regulatory framework governing that information exchange.

US legislation regarding the disclosure of material, private information makes an important distinction between information shared *with*, or *without* personal gain.

- If an insider shares material, private information *with* personal financial gain or outside of a corporate purpose (e.g. sharing with friends or family), then they and the recipient of that information are engaging in *insider trading* - a criminal offence governed by the Securities and Exchange Act (1934).

- If, on the other hand, an insider shares material, private information *without* personal financial gain and within a corporate context (e.g. in a Q&A at a private meeting with shareholders), then that information disclosure constitutes *informed trading* - which falls under the arguably much more lenient purview of Regulation Fair Disclosure, commonly referred to as Reg FD (2000).

It is the latter of these two instances that we examine in greater detail throughout this chapter, with a focus on the content of Reg FD, the wrongdoing it aims to police, and the market's response to enforcement actions taken by the Securities and Exchange Commission (SEC).

6.2 Background

In 2012, the chairwoman of the SEC argued that market participants had "short memories" and that the SEC as a result had to take regular enforcement actions to remind them of their legal obligations and keep markets clean (Wyatt, 2012). This claim has intuitive appeal: if the capital markets regulator does not enforce its rules

with some regularity, potential wrongdoers may interpret the resulting inactivity as a reduced risk of apprehension and increase their production of wrongdoing (Becker, 1968). Surprisingly, the literature on this topic lacks conclusive evidence of how market participants actually react to enforcement actions - or inaction - by the SEC.

In this chapter, we examine the unique properties of Reg FD - the SEC regulation that aims to prohibit insiders' disclosure of information to selected investors in favour of broad public disclosure methods - to assess how insiders adapt their behaviour after SEC enforcement actions. We construct correlation metrics for gauging the propensity towards informed trading ahead of earnings disclosures. We find that SEC enforcement has an immediate deterrent effect on insider leakage and that this effect is particularly notable when the SEC enforces Reg FD after a long period of inactivity. We also produce a novel framework for modelling the memory of enforcement and estimate that SEC escalations cause insiders to change their leakage behaviour for approximately 24 months.

In examining the impact of SEC enforcement on information leakage, this research contributes to at least three strands of literature.

6.2.1 Measuring The Deterrent Effect of Enforcement

First, it builds on existing literature that studies the importance of public enforcement. La Porta et al (2006) and Jackson and Roe (2009) used cross-country data to study the importance of enforcement mechanisms for capital market development. The latter paper presented persuasive evidence on the importance of public enforcement, specified with resource-based metrics such as regulators' budgets and staff numbers. While Armour et al (2016) and Leuz and Wysocki (2016) have noted that there is scant empirical research on the effects of enforcement in financial markets, two recent studies have built on Jackson and Roe (2009) using SEC resource data. Lohse et al (2014) examined the relationship between changes in the SEC's budget

and its enforcement activity and argued that firms reacted to increases in the SEC budget by improving their compliance behaviour (measured as the amount of SEC injunctive actions). The findings are however subject to the concerns that the amount of enforcement may not be an appropriate proxy for the amount of wrongdoing, and that a general increase in the SEC's budget may not cause increased compliance in all areas that the SEC supervises.

Del Guercio et al (2017), the paper closest to this research, found that share price run-ups before news announcements were negatively related to SEC budgets and staffing levels, concluding that SEC efforts deterred insider trading and weakened price discovery. Our findings expand on this in two ways. First, we place a stronger emphasis on understanding the legal environment, enabling us to tailor our enquiry. Del Guercio et al treated all trading around earnings announcements as subject to the same legal framework, but - as we show in Section 6.3 - outsiders trading on leaked information are subject to a more lenient regulatory environment than insiders, and may often be able to lawfully trade even when insiders leak material information to them. The run-up patterns they observed are thus not necessarily due to illegal activity. Furthermore, we evaluate the impact of actual SEC enforcement actions (rather than proxies such as budgets) on insider behaviour and find, for the first time, that SEC enforcement has a significant and immediate deterrent effect on insider leakage. In particular, SEC escalations in sanctioning - by which we mean the application of harsher penalties than ever before - create the most deterrence.

6.2.2 The Deployment of Private Information

This study also adds to a fast-growing but under-theorised literature on how insiders deploy private information. Bengtzen (2017) describes the four different options by which insiders can cause private information to impact stock prices: broad public disclosure, selective disclosure (or "leakage") to outsiders who trade, personal (insider)

trading, and trading via the firm (such as buybacks). By shining a light on insider leakage, we hope to assist researchers aiming to synthesise the literature and develop a theory of how insiders choose between these options for information deployment.

Important earlier work in this field with a focus on *insider trading* include Cohen et al (2012), who found that opportunistic insiders reduced their trading activity when they perceived the SEC to be more active in insider trading enforcement, as well as Huddart et al (2007) and Hu et al (2017) which found that insiders avoid trading in high-jeopardy periods such as just before earnings announcements. It is interesting to juxtapose those studies with prior work with a focus on *insider leakage* such as Campbell et al (2009), Yan and Zhang (2009), Berkman and McKenzie (2012), and Hendershott et al (2015) which all found evidence that institutional investors are informed of earnings surprises ahead of time. We are not aware of empirical work on the complementarities between insider trading and insider leakage, but one may suspect based on this literature that insiders prefer leakage to trading when they consider themselves subject to more regulatory scrutiny.

Other notable studies of leakage include Cohen et al (2008), who identified that mutual fund managers with the same educational background as insiders outperformed when investing in those insiders' stock. Butler and Gurun (2012) found that such fund managers are more likely to support insiders in contentious votes on their pay packages. Notable studies of the sources of private information include Griffin et al (2012) who found that connected brokerage houses were not the source of informed trading prior to earnings announcements, while Solomon and Soltes (2015) found that private meetings with insiders helped certain investors make better trading decisions and Akbas et al (2016) found that firms with more connected board members experienced more informed trading.

6.2.3 A Dynamic Analysis of Regulation Fair Disclosure

Finally, this work contributes to the literature on Reg FD, where prior studies of its effect on the information environment, e.g. Bailey et al (2003) and Ahmed and Schneible (2007), assess the effects of Reg FD by comparing data before and after its introduction. By examining how the regulator actually makes use of its Reg FD powers over time, we offer a more dynamic approach for evaluating its effectiveness. This work thus complements Griffin et al (2011) - to our knowledge, the only prior paper on Reg FD enforcement - which studied violating firms' stock price reactions to enforcement announcements. As an exploration of potential loopholes in regulatory design, this research may also be of interest to policy makers and capital market regulators. The data set employed throughout this chapter covers all Reg FD enforcement actions at time of writing, due to the SEC not taking any action in this area since 2013. This is - by far - the longest stretch of enforcement inactivity since Reg FD was introduced, and our results indicate that the SEC would reduce insider leakage if it took action.

Section 6.3 outlines the regulatory framework of information leakage in the US and formulates the hypotheses of the chapter. In Section 6.4 we describe our data and methodology. Section 6.5 presents the results and Section 6.6 concludes.

6.3 Regulatory Setting and Hypotheses

6.3.1 The SEC's Regulation of Information Leakage

The SEC introduced Reg FD in 2000 to curb the common practice of corporate insiders strategically leaking valuable information to their favoured analysts and investors. Formally, Reg FD requires that "[w]henever an issuer, or any person acting on its behalf, discloses any material nonpublic information regarding that issuer or its securities to [certain persons such as securities market professionals and shareholders],

the issuer shall make public disclosure of that information" either simultaneously (if the leakage is intentional) or promptly (if unintentional).[1]

By requiring simultaneous disclosure of material information, the SEC aimed to level the playing field so that all investors had access to the same information at the same time. While Coffee (2016) has argued that Reg FD has curbed systematic information leakage, Bengtzen (2017) instead posits that the regulation has severe design flaws which allow insiders significant opportunity to leak information.

The first notable feature of Reg FD is that its introduction had the effect of clearly bifurcating the regulation of selective disclosure and insider trading. After Reg FD, insiders who disclose material nonpublic information for a corporate purpose (for example, to investors and analysts) are only subject to Reg FD, while insiders who disclose without a corporate purpose (for example, to friends and family, or in return for a personal benefit) remain under the insider trading framework. This means that we are able to focus on Reg FD as the only legal framework governing strategic information leakage - intentional selective disclosures to investors and analysts for corporate benefit.

The second important attribute of Reg FD is that it leaves the decision as to whether or not valuable information can be selectively disclosed to corporate insiders. A common misconception is that Reg FD prohibits insiders from selectively disclosing any *valuable* information,[2] but this is not the case - it only prohibits disclosure of *material* information, which is a higher threshold.[3] Insiders can thus selectively disclose some information that professional recipients may consider valuable without falling foul of Reg FD. This is well illustrated by the case of *SEC v. Siebel Systems*,

[1] 17 C.F.R. §§ 243.100-103 (2015).

[2] For example, it appears that Del Guercio et al (2017), in labelling all share price run-ups before news announcements as evidence of "illegal trading", fall victim to this misconception.

[3] Securities regulation provides that information is material if a reasonable investor would view the relevant piece of information as significantly changing the 'total mix' of information available about a company. TSC Industries, Inc. v. Northway, Inc., 426 U.S. 438, 449 (1970).

the only Reg FD event the SEC has opted to litigate.[4] In *Siebel*, the SEC argued that an insider had selectively released material information in private investor meetings which caused the share price to increase sharply as the investors aggressively purchased shares.[5] The court, however, did not consider the information material and dismissed the SEC's action.[6] Insiders will thus continuously have to make judgment calls about how much valuable information to disclose, and each such judgment call will be unique, depending on the question asked and the actual situation in each firm at any given time.

Thirdly, Reg FD only applies to public companies: it does not impose any legal obligations on the analysts and investors who receive selectively disclosed information. If such outsiders receive leaked information, they are free to trade on it as long as they did not provide a personal benefit to the insider (in which case they would be caught by the insider trading framework as described above).[7] This means that the amount of leakage we observe in the market at any time depends on the contemporaneous risk aversion of insiders only: since recipients do not face any legal risk, they should demand as much information as possible and can trade freely.

But perhaps the most interesting attribute of Reg FD is its distinction between *intentional* and accidental, *unintentional* disclosures. Whilst intentional disclosures must be immediately addressed with public disclosure of the same information, firms have a 24-hour window to remedy an unintentional private disclosure.[8] While the

[4]SEC v. Siebel Systems, Inc., 384 F.Supp.2d 694 (S.D.N.Y. 2005).

[5]SEC Complaint at §§ 46, 53, Siebel Systems, 384 F.Supp.2d at 694 (No. 04-CV-5130).

[6]For further examples of information that may be valuable but not material in a legal sense, see Bengtzen (2017), n. 22 on p. 47.

[7]Indeed, Fidelity traded on material information selectively disclosed by Schering-Plough in 2003 and, when later asked about the incident by a journalist, replied that "We complied with all rules and regulations in our meeting with Schering-Plough and in our conduct thereafter" (Norris, 2003). Both Schering-Plough's CEO and the company itself, however, were found to have breached Reg FD and had to pay penalties (SEC 2003).

[8]Technically, Reg FD requires "prompt" public disclosure, which it defines to mean as soon as reasonably practicable after a senior official of the issuer learns of the leakage, but not later than the latest to occur of (i) 24 hours, or (ii) the start of next day's stock exchange trading. While issuers technically should disclose sooner than 24 hours if that is practicable, that determination is left to the issuers themselves. In practice, the SEC calls it a "24-hour requirement" (SEC 2000, 51722) and

information *itself* must be released to the market, firms do not have to disclose that the information had been leaked. Consequently, if there is a leak within 24 hours of a scheduled press release, firms only have to issue that press release on schedule to comply with their Reg FD duties, meaning that market participants at large need never be made aware of information leaks. The SEC has also announced that it will apply a lower standard for behaviour in private meetings since it understands the difficulty for insiders to make legal assessments in real-time unrehearsed settings,[9] and that it will impose less severe sanctions on leaks that last for shorter periods of time (SEC 2000, 51726). As a result of the SEC's approach to pre-announcement leaks, we expect insiders to feel relatively unconstrained in investor discussions during the 24 hours before a scheduled press release.

Finally, two features specific to the SEC's enforcement practices are important to our research design. Firstly, Reg FD can only be enforced by the SEC - no private securities suits are possible.[10] The SEC also prefers formal enforcement actions aimed at generating market-wide deterrence to more informal enforcement methods such as those seen in the UK (Coffee, 2007; Armour, 2009). By studying all SEC enforcement actions of Reg FD, we can thus be sure to cover all events where the regulator has acted to directly deter strategic information leakage. Secondly, the SEC's practices in connection with Reg FD investigations and sanctioning is to only issue one announcement of enforcement action at the completion of its investigation. While some leakage events were subject to media speculation at the time they took place, the timing of SEC announcements is unrelated to the leakage events and often occur more than a year after the violations.[11] The SEC's enforcement announcements are

a prominent law firm has described the timing requirement as "within 24 hours" (Shearman and Sterling, 2005), thus indicating that disclosure within 24 hours is acceptable.

[9]In the words of the SEC (2000): "[A] materiality judgment that might be reckless in the context of a prepared written statement would not necessarily be reckless in the context of an impromptu answer to an unanticipated question."

[10]The regulation explicitly stipulates this in 17 C.F.R. §243.102 (2015).

[11]For violation and SEC announcement dates, see Table 6.10.

thus unexpected by market participants, which provides us with an opportunity to study how insiders react to them with precision.

6.3.2 Formulation of Hypotheses

Building on the seminal contribution of Becker (1968) on the economics of criminal activity, our research design assumes that an insider considering information leakage will weigh up its expected benefits and costs. The benefits that insiders gain from engaging in selective disclosure include buying support from selected analysts and investors (SEC 1999, 72592), which could accrue both to insiders and to their corporations (Coffee, 1997). The expected costs of engaging in leakage, on the other hand, can be conceptualised as a function of the probability that the SEC detects the behaviour and the severity of the sanction if detected.

As is well-established in criminology literature, we consider the amount of deterrence Reg FD creates in the mind of insiders to be susceptible to external influences that vary in strength over time (Apel and Nagin, 2011; Chalfin and McCrary, 2017). Deterrence is achieved when potential wrongdoers refrain from an activity due to the perceived risk of detection and sanctioning, which means that the regulator may be able to increase deterrence by signaling that it pays attention to certain behaviour (Geerken and Gove, 1975).

Examining the specific setting of Reg FD, it appears reasonable that insiders will perceive the strength of SEC deterrence to change over time, for two main reasons. First, the SEC's enforcement priorities are politically influenced due to its dependency on Congress for annual budget approvals and Congress has been described as "micromanaging" the SEC (Velikonja, 2015). As a result of this political influence, the SEC is known to change its priorities frequently. Secondly, the SEC has limited resources and its performance is measured in terms of how many cases it brings and the amount of fines it collects (Macey, 2010). The SEC prefers cases that do not require lengthy

or difficult investigations and to take action only after an issue becomes highly visible in the financial press (Macey and O'Hara, 2009). Since selective disclosure occurs in private conversations, it is costly to detect and difficult to substantiate allegations. The small size of penalties that the SEC can assess under Reg FD relative to the amounts at stake in other areas such as insider trading[12] may thus cause the SEC to direct its resources to areas that entail more publicity and larger penalties than Reg FD enforcement, unless insider leakage becomes a salient issue for Congress or in the press. We therefore hypothesise that the intensity of SEC enforcement of Reg FD will change over time, and than an enforcement action will serve as a signal to insiders that leakage is salient:

Hypothesis H1: Reg FD enforcement actions deter leakage.

Insiders will interpret SEC enforcement of Reg FD as a signal that the regulator is focusing on this regulatory framework and respond by reducing their production of information leakage. Building on this first hypothesis, we also want to investigate the claim that insiders have "short memories", which asserts that if the regulator does not enforce its rules with some regularity, potential wrongdoers will interpret the resulting inactivity as a reduced risk of apprehension and increase their production of wrongdoing. We thus conjecture that SEC enforcement actions that take place after a significant period of inactivity will have a particularly notable deterrent effect on leakage:

Hypothesis H2: Reg FD enforcement actions that occur after significant SEC inactivity will be characterised by increasing leakage before the events and decreasing leakage after the events.

The third hypothesis we wish to examine relates to the types of signals the SEC

[12]For example, the SEC obtained a $92.8 million fine from Raj Rajaratnam for insider trading (SEC 2011), an amount 28 times higher than the sum of all fines it has ever issued under Reg FD ($3.3 million) since it entered into force in 2000.

can send to corporate insiders. As noted above, the SEC has two levers at its disposal to raise the costs of information leakage and increase deterrence: it can increase its rate of enforcement to counteract the "short memories" problem and it can increase the sanctions it imposes. Since we study the entire lifespan of Reg FD, we have the opportunity to study how market participants adapt as the SEC's enforcement practices evolve. We hypothesise that insiders will take particular notice when the SEC steps up its enforcement to deploy a "bigger gun" in Reg FD enforcement:

Hypothesis H3: SEC escalations in sanctioning will cause insiders to immediately adjust upwards their expected costs of engaging in leakage, producing a notable reduction in leakage.

Finally, we examine the "short memories" claim and seek to quantify the persistence of enforcement actions in the minds of insiders.

6.4 Data and Methodology

After describing the data employed in our study, we outline the parametric and non-parametric methodologies employed to evaluate the market's behaviour around corporate earnings announcements. We provide the means of replicating these datasets in Appendix A.4.

6.4.1 Measuring Information Leakage

The main problem of studying strategic information leakage is that the details of leakage events are known only to the two parties involved – a leaking insider and an outside investor or analyst. Studies of the impact of regulation on insider behaviour therefore infer information flow from stock price data, Jaffe (1974) being an early example, and we follow a similar approach. We study quarterly earnings announcements - the typical example of pre-scheduled releases - of S&P500 stocks, and use 21

years of market data (Jan-1995 to Dec-2015) covering the 5 years before Reg FD was introduced and the 16 years thereafter. As is later shown in Table 6.10, the SEC has not enforced Reg FD since 2013, so our dataset covers all Reg FD actions. Quarterly earnings announcements are a suitable event for our purposes, since insiders themselves learn increasingly precise information about their results as they finalise them. Since this information is known only by a select group of corporate insiders prior to release, Occam's razor suggests that insider leakage should be the starting assumption if late-stage informed trading is observed in connection with earnings announcements.

We estimate the impact of enforcement on insider leakage by analysing changes in the correlation between companies' Abnormal Returns on the day before earnings announcement and the Earnings Surprise in that (subsequent) announcement. We measure Abnormal Returns by Carhart alpha and the Earnings Surprise as the difference between actual quarterly EPS and analysts' consensus forecast EPS before the announcement, scaled by the most recent closing price:

$$\text{Earnings Surprise(t)} = \frac{\text{EPS(t)} - \text{Forecast EPS(t)}}{\text{Share Price(t)}} \tag{6.1}$$

For each S&P500 firm n and earnings announcement k, we designate τ as the first trading day after the earnings release, hereafter abbreviated to ER. We monitor on a case-by-case basis whether the earnings release occurred before market open (τ is the announcement date) or after market close (τ is the first business day directly following the announcement date). We track the excess return $R_{n,k}$ above the 1-month risk-free rate at various mutually exclusive timescales. These include:

- *ER Day Return:* the excess return on the first trading day after ER, $R_{n,k}(\tau)$.
- *Prior Day Return:* the excess return on the day prior, $R_{n,k}(\tau - 1)$.

- *Following Week Return:* the excess return for the week following ER, excluding the ER date $\sum_{\tau=T+1}^{\tau=T+4} R_{n,k}(T)$.

- *Prior Week Return:* the excess return for the week preceding ER, excluding the prior day $\sum_{\tau=T-5}^{\tau=T-2} R_{n,k}(T)$.

In all instances we wish to pinpoint the Abnormal Return component attributable to firm-specific earnings announcement information rather than the established factors of market sensitivity (Sharpe, 1964; Lintner, 1965), size effect, growth effect (Fama and French, 1992) and cross-sectional momentum (Carhart, 1997). To this end, we equate Abnormal Return to the Carhart 4-factor model alpha $\alpha_{n,k}$ for all time frames, drawing our firm-specific parameters $\beta_{n,k}^{MKT}$, $\beta_{n,k}^{SMB}$, $\beta_{n,k}^{HML}$ and $\beta_{n,k}^{MOM}$ from 3-year multivariable rolling regressions against the 4 factor returns of the Carhart model: market return R_k^{MKT}, small minus big market capitalisation portfolio return R_k^{SMB}, high minus low price-to-book ratio portfolio return R_k^{HML} and high minus low momentum portfolio return R_k^{MOM}. For example, ER Day Abnormal Return $\alpha_{n,k}(\tau)$ for S&P500 firm n and earnings announcement k would be:

$$\begin{aligned}\alpha_{n,k}(\tau) = R_{n,k}(\tau) - \beta_{n,k}^{MKT} \times R_k^{MKT}(\tau) - \beta_{n,k}^{SMB} \times R_k^{SMB}(\tau) \\ -\beta_{n,k}^{HML} \times R_k^{HML}(\tau) - \beta_{n,k}^{MOM} \times R_k^{MOM}(\tau)\end{aligned} \tag{6.2}$$

Abnormal Returns at other timescales are derived by replacing τ in Equation (6.2) and summing across the appropriate date ranges defined above.

6.4.2 Correlation Analysis

We begin by establishing the characteristics of our correlation metric to ensure it is correctly specified. We measure the Pearson (Table 6.1) and Spearman (Table 6.2) correlation between Abnormal Returns at various timeframes and our target variable, Earnings Surprise. Using the date the first Reg FD proposals were published by the

Table 6.1: Linear Correlation between Abnormal Returns and Earnings Surprise on the S&P500 dataset, N=31,706 (6,111 in the Jan-95 to Dec-99 pre-Reg. FD era, 25,595 in the Jan-00 to Dec-15 post-Reg. FD era).

	Pre-Reg FD		Post-Reg FD	
Feature	Corr	p-value	Corr	p-value
ER Day Abnormal Return	**+0.0860**	**< 0.0001**	**+0.1485**	**< 0.0001**
Following Week Abnormal Return	+0.0208	0.1038	+0.0187	0.0028
Prior Day Abnormal Return	**+0.0662**	**< 0.0001**	**+0.0304**	**< 0.0001**
Prior Week Abnormal Return	**+0.0550**	**< 0.0001**	**+0.0249**	**< 0.0001**
ER Day Volume	+0.0075	0.5601	−0.0004	0.9441
Following Week Volume	−0.0147	0.2521	−0.0122	0.0515
Prior Day Volume	−0.0133	0.2985	+0.0173	0.0058
Prior Week Volume	+0.0007	0.9560	+0.0118	0.0592

SEC (Dec. 20, 1999) as our demarcation line, we group the results by domain for both the pre-Reg FD (Jan 1995 - Dec 1999, N=6,111) and post-Reg FD (Jan 2000 - Dec 2015, N = 25,595) eras. With outliers (often errors in recorded data) clipped via 97.5% winsorisation, linear and rank correlation broadly identify the same features as salient. We hereafter quote only the Spearman correlation for brevity and identify significance when p-values are below the 0.001 significance threshold.

Table 6.2: Rank Correlation between Abnormal Returns and Earning Surprise on the S&P500 dataset, N=31,706 (6,111 in the Jan-95 to Dec-99 pre-Reg. FD era, 25,595 in the Jan-00 to Dec-15 post-Reg. FD era).

	Pre-Reg FD		Post-Reg FD	
Feature	Corr	p-value	Corr	p-value
ER Day Abnormal Return	**+0.1229**	**< 0.0001**	**+0.2210**	**< 0.0001**
Following Week Abnormal Return	+0.0163	0.2026	**+0.0254**	**< 0.0001**
Prior Day Abnormal Return	**+0.0734**	**< 0.0001**	**+0.0463**	**< 0.0001**
Prior Week Abnormal Return	**+0.0509**	**< 0.0001**	**+0.0336**	**< 0.0001**
ER Day Volume	+0.0252	0.0487	+0.0140	0.0256
Following Week Volume	−0.0112	0.3805	−0.0193	0.0020
Prior Day Volume	+0.0107	0.4033	**+0.0376**	**< 0.0001**
Prior Week Volume	+0.0269	0.0357	**+0.0210**	**0.0007**

We observe a strong positive correlation between Abnormal Returns on a firm's earnings report date and the Earnings Surprise they just disclosed, a relationship which has strengthened since Reg FD's introduction (+12.29% in the pre-Reg FD era

vs. +22.10% in the post-Reg FD era). The stronger reaction in ER Day Abnormal Return since Reg FD's implementation (1.38bp per 1% of Earnings Surprise vs 1.19bp per 1% Earnings Surprise before the regulation came into effect) is consistent with Gomes et al (2007) and suggests that Reg FD does help to level the informational playing field among investors. We also find evidence of post-earnings-announcement drift (Ball and Brown, 1968), indicated by light, consistently positive correlations between surprise and returns on the week (+1.63% pre-Reg FD, 2.54% post-Reg FD) following ER.

More interesting for our purposes is the strong positive correlation between Earnings Surprise and Abnormal Returns at both anticipatory timeframes in our study: the prior day and week. These correlations decline nominally after Reg FD's announcement, but remain significant. In particular, the Prior Day Abnormal Return is more sharply aligned with upcoming Earnings Surprise than the Prior Week Abnormal Return, suggesting more leakage in the 24-hour window, consistent with our expectations based on the regulatory framework described in Section 6.2. The cross-correlation function of Earnings Surprise as a function of lagged daily Abnormal Returns (Figure 6.1) confirms this. The last day before ER exhibits a considerably higher correlation with Earnings Surprise (+5.1%) than any of the earlier days. Notably, the correlation is positive at almost every single lag.

Regressing Earnings Surprise on Abnormal Returns at various lags provides further evidence that this "pre-earnings announcement drift" is more significant than the post-earnings announcement drift. Table 6.3 provides the t-statistics for each factor in the multivariable regression of Earnings Surprise on Abnormal Returns across 20 lags. The results mirror the correlation findings of Tables 6.1 and 6.2, and are consistent with segments of the market possessing foreknowledge of the Earnings Surprise and deploying it most extensively in the Prior Day.[13]

[13]We also examined regression techniques with built-in feature selection. Applied to the features of Table 6.3, LASSO gradually prunes every feature except Prior Day as we increase regularisation.

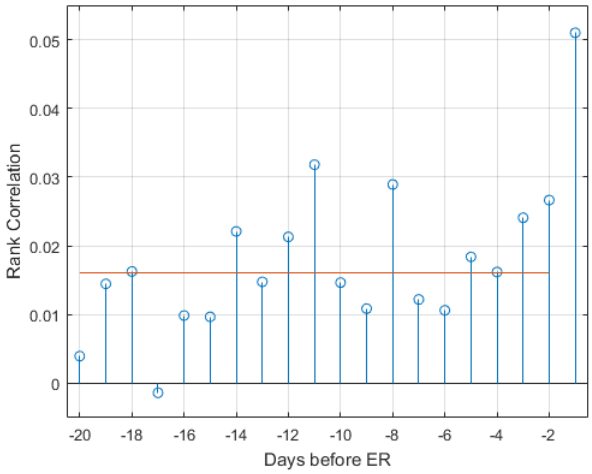

Figure 6.1: Rank Cross-Correlation Function of Earnings Surprise against Daily Abnormal Returns at various lags (mean correlation of +1.6% denoted by the transversal line).

Table 6.3: t-statistics and associated p-values of the regression of Earnings Surprise on Abnormal Returns at various lags.

	Pre-Reg FD		Post-Reg FD	
Lag	t-stat	p-value	t-stat	p-value
Prior Day	**5.5662**	**<0.0001**	**5.4483**	**<0.0001**
ER Day -2	4.7756	<0.0001	1.9854	0.0471
ER Day -3	3.1516	0.0016	1.7883	0.0737
ER Day -4	2.2508	0.0244	2.7127	0.0067
ER Day -5	2.9554	0.0031	4.6761	<0.0001
ER Day -6	0.3869	0.6989	2.8648	0.0042
ER Day -7	1.3340	0.1823	2.0423	0.0411
ER Day -8	3.6728	0.0002	4.5948	<0.0001
ER Day -9	0.3388	0.7348	1.5083	0.1315
ER Day -10	1.5781	0.1146	2.1716	0.0299
ER Day -11	1.9140	0.0557	4.3640	<0.0001
ER Day -12	3.1397	0.0017	2.6192	0.0088
ER Day -13	1.7579	0.0788	2.2931	0.0219
ER Day -14	2.0774	0.0378	3.2357	0.0012
ER Day -15	-0.3662	0.7142	2.7966	0.0052
ER Day -16	-1.2047	0.2284	1.0759	0.2820
ER Day -17	1.1909	0.2337	-0.3403	0.7336
ER Day -18	0.5767	0.5642	0.9596	0.3373
ER Day -19	1.7734	0.0762	2.3196	0.0204
ER Day -20	1.8173	0.0692	-0.1499	0.8809

In Table 6.4 we examine the relevance of Prior Day Abnormal Returns across quintiles of the S&P500 to assess the impact of market capitalisation, and trace its evolution over 4 equal periods of 5 years to compare samples of similar size (N≈1600 per 5-year period in each quintile). Whilst fairly stable across the index and through time, the correlation is more pronounced amongst smaller companies (bottom quintile of Table 6.4), confirming that our data is in line with prior Reg FD findings from Ahmed and Schneible (2007) and Gomes et al (2007).

Table 6.4: Rank Correlation between Prior Day Abnormal Return and Earnings Surprise tiered by S&P500 quintile, N≈1,600 per period for each quintile (Jan-95 to Dec-14).

	Pre-Reg FD		Post-Reg FD					
	1995-1999		2000-2004		2005-2009		2010-2014	
Quintile	Corr	*p*-value	Corr	*p*-value	Corr	*p*-value	Corr	*p*-value
1st	+0.0646	0.0160	+0.0126	0.5878	**+0.1237**	**< 0.0001**	**+0.0735**	**0.0009**
2nd	+0.0690	0.0143	+0.0583	0.0128	+0.0666	0.0023	+0.0267	0.2186
3rd	+0.0414	0.1458	−0.0083	0.7321	−0.0035	0.8748	+0.0045	0.8378
4th	+0.0698	0.0170	+0.0286	0.2276	+0.0378	0.0817	+0.0566	0.0094
5th	+0.1110	0.0003	+0.0699	0.0071	**+0.0945**	**< 0.0001**	**+0.0933**	**0.0002**

As a final check on the correlation metric's properties, we run similar, separate correlation analyses on 11 recent years of data for some of the largest European stocks - FTSE100 companies in the UK and CAC40 companies in France (Tables 6.5 and 6.6). While the most severe sanction for leakage in the US is an SEC penalty, the same activity in each of these European countries would be a criminal offence for which both a leaking insider and a trading outsider can go to prison (Clarke, 2013). None of these European markets provided evidence at even the 0.05 significance level that information about earnings is manifesting prematurely, further indicating that the permissive regulatory setting in the US may have explanatory value for the amount of observed leakage.

Table 6.5: Rank Correlation between Abnormal Returns and Earning Surprise on the FTSE100 dataset, N=1,307 (Jan-05 to Dec-15).

	Pearson		Spearman	
Feature	Corr	p-value	Corr	p-value
ER Day Return	+0.0468	0.0906	**+0.1209**	**< 0.0001**
Following Week Return	+0.0031	0.9109	+0.0387	0.1624
Prior Day Return	+0.0250	0.3664	+0.0247	0.3716
Prior Week Return	+0.0073	0.7913	+0.0033	0.9041

Table 6.6: Rank Correlation between Abnormal Returns and Earning Surprise on the CAC40 dataset, N=756 (Jan-05 to Dec-15).

	Pearson		Spearman	
Feature	Corr	p-value	Corr	p-value
ER Day Return	+0.0401	0.2710	+0.0563	0.1220
Following Week Return	−0.0462	0.2042	−0.0294	0.4200
Prior Day Return	+0.0824	0.0235	+0.0487	0.1811
Prior Week Return	+0.0064	0.8599	−0.0044	0.9038

6.4.3 Non-Parametric Function Learning

We examine more closely the relationship between Abnormal Returns and Earnings Surprise in the post Reg-FD era, using the ARD Gaussian Process framework described in Section 2.4. Having determined via cross-validation the kernel that generalises best,[14] we implement separate Gaussian Process regressions for each feature grouping (post-earnings features denoting the market's reaction and pre-earnings features denoting the market's anticipation) using the Squared Exponential kernel.

$$k(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\Big(- \frac{(\mathbf{x} - \mathbf{x}')^2}{2l^2} \Big) \tag{6.3}$$

A sidenote: while the Squared Exponential kernel performed marginally better than the alternatives, the performance range across all options was narrow. Function learning on this domain is not sensitive to kernel selection, and our findings would not be materially altered by an alternative choice of covariance function.

Table 6.7: Relevance Ratio and Spearman p-value of market reaction features.

	Relevance	
Feature	Ratio	p-value
ER Day Return	1.3×10^2	< 0.0001
Following Week Return	8.7×10^1	< 0.0001
Noise	1	0.9362

Table 6.7 provides an analysis of the relevance of market reaction features. ER Day market reaction and surprise move in near-lockstep: short-term price impact (x-axis in Figure 6.2) monotonically follows surprise, with the medium term (y-axis in Figure 6.2) reflecting the post-earnings announcement drift.

[14] Optimal kernel selection was achieved via 5-fold cross-validation on a wide range of commonly used kernels (Squared Exponential, Rational Quadratic, Matérn 1/2, Matérn 3/2, Matérn 5/2), with predictive performance of each kernel measured in Normalised Root Mean Squared Error (NRMSE), Median Absolute Deviation (MAD) and Spearman correlation between actual and forecast Earnings Surprise.

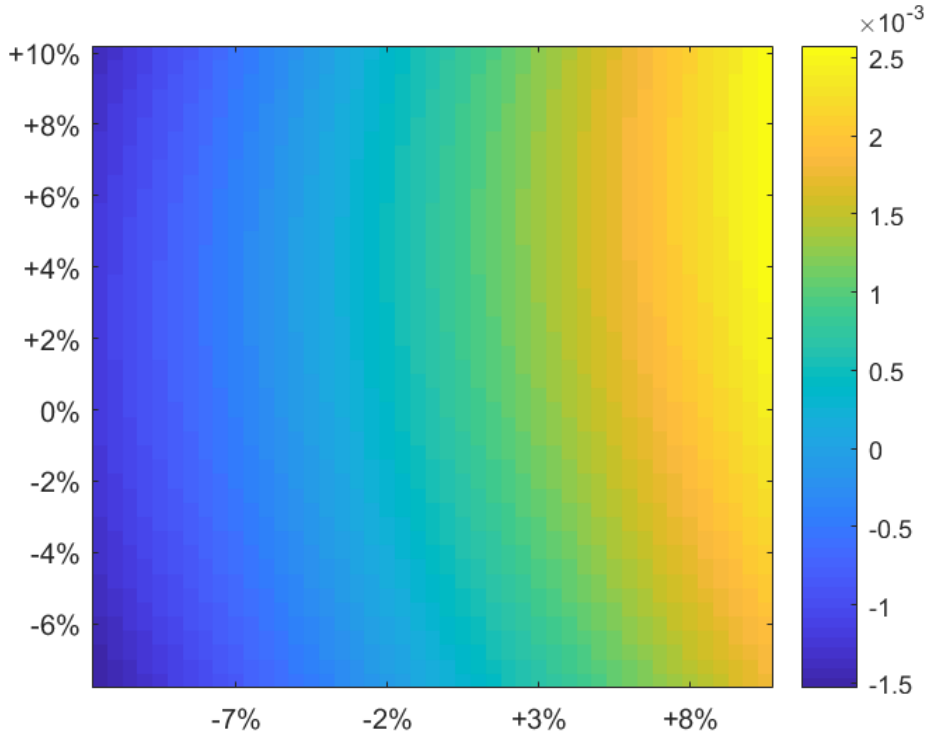

Figure 6.2: Earnings Surprise variation as a function of Earnings Day Return (x-axis) and Following Week Return (y-axis). The impact of an Earnings Surprise is most pronounced in the immediate aftermath of its disclosure, manifesting as variation predominantly along the x-axis.

Table 6.8: Relevance Ratio and Spearman p-value of market anticipation features.

	Relevance	
Feature	Ratio	p-value
Prior Day Return	1.5×10^2	< 0.0001
Prior Week Return	1.3×10^2	< 0.0001
Noise	1	0.9362

Relevance for the market's anticipatory features is provided in Table 6.8. We provide a bivariate visualisation of ER Surprise as a function of Prior Day and Prior Week Returns (Figure 6.3). The mean function heatmap varies almost equally in function of the past day (x-axis) and week (y-axis) for much of the distribution,

consistent with the hypothesis that informed trading may be occurring in the day and week prior to earnings disclosures. Whilst the alignment manifesting over the preceding week could be due to the market pursuing efficient price discovery ahead of the announcement and being rewarded for it (Grossman and Stiglitz, 1980), it is hard to conceive of due diligence producing such a pattern on Prior Day Returns.

Furthermore, an asymmetry is apparent in the function mapping of Figure 6.3: at the lowest percentiles of Prior Week Return (-7% and lower), Earnings Surprise barely covaries with Prior Day Return, and the Earnings Surprise is uniformly negative. Prior Day Returns are most informative on stocks that have exhibited a run-up in the preceding week. Our evidence is consistent with prior day information leaks ahead of earnings beats but not of major misses, which would require conspicuously timed naked shorting to profit from. The ability to identify such asymmetries and non-parametrically model the interaction of variables is one of the benefits of GP function learning over a parametric or correlation-based approach.

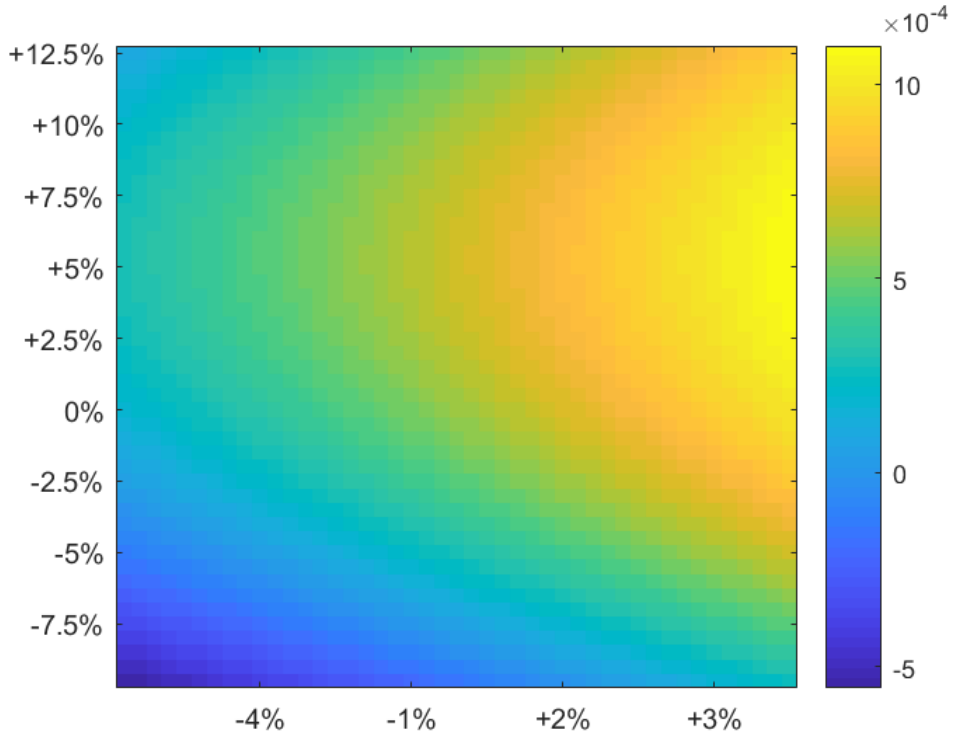

Figure 6.3: Earnings Surprise variation as a function of Prior Day Return (x-axis) and Prior Week Return (y-axis). As shown earlier in Tables 6.1 and 6.2, Earnings Surprise covaries positively with both Prior Day Return and Prior Week Return, but for deeply negative (-7% and below) Prior Week Returns, the Prior Day ceases to be informative.

6.4.4 SEC Enforcement Events

The SEC shapes the risk perception of would-be violators by showing a general interest in a particular activity (such as making statements about its stance on an activity) and by taking and publishing enforcement actions (Apel, 2013). To maximise the data available, we thus include both the SEC's statements surrounding the introduction of Reg FD (Items A, B and C in Table 6.9 below),[15] and its enforcement

[15]These include the date the SEC first issued its proposal for Reg FD, the date the SEC published the finalised regulation, and the date it entered into effect, since all three events may lead to increased publicity of the SEC's initiatives.

actions to enforce Reg FD (Items 1-13 in Table 6.10 below).[16]

Table 6.9: Reg FD Statements by the SEC.

ID	Announcement Date	Event
A	Dec 20, 1999	First Reg FD proposal
B	Aug 15, 2000	Final Reg FD published
C	Oct 23, 2000	Reg FD effective date

Table 6.10: Reg FD Enforcement Actions by the SEC.

ID	Announcement Date	Issuer	Issuer Penalty	Insider Penalty	Violation Date
1	Nov 25, 2002	Raytheon	$0	$0	Feb 2001
2	Nov 25, 2002	Secure Computing	$0	$0	Mar 2002
3	Nov 25, 2002	Siebel Systems I	$250,000	$0	Nov 2001
4	Sep 9, 2003	Schering-Plough	$1,000,000	$50,000	Sep-Oct 2002
5	Jun 29, 2004	Siebel Systems II	N/A	N/A	Apr 2003
6	Sep 16, 2004	Senetek	$0	$0	Jun-Sep 2003
7	Mar 24, 2005	Flowserve	$350,000	$50,000	Nov 2002
8	Sep 25, 2007	EDS	$0	$0	Sep 2002
9	Sep 24, 2009	ACL	$0	$25,000	Jun 2007
10	Mar 9, 2010	Presstek	$400,000	$50,000	Sep 2006
11	Oct 21, 2010	Office Depot	$1,000,000	$100,000	Jun 2007
12	Nov 22, 2011	Fifth Third Bancorp	$0	$0	May 2011
13	Sep 6, 2013	First Solar	$0	$50,000	Sep 2011

Since Reg FD entered into force in October 2000, the SEC has taken action in a total of 13 cases, one of which it chose to litigate in court and subsequently lost. The remaining 12 actions were all settlements where the issuer and/or the leaking insider agreed to cease-and-desist orders and, in seven of those cases, to pay penalties.

6.5 Results

6.5.1 The Deterrence Effect of SEC Enforcement

As noted in Section 6.4, we adopt the Spearman correlation between Prior Day Abnormal Return and Earnings Surprise as our gauge of information leakage. To test our H1, we evaluate leakage over two time frames: the 3 months directly preceding each SEC enforcement action, and the 3 months directly following it. If SEC enforcement actions affect the information environment, our leakage metric should be

[16]For *Siebel Systems II*, we use the date of the SEC's announcement that it was pressing charges and not the later date when the court gave its verdict. This is because we are focused on actions by the SEC aimed at creating deterrence.

Table 6.11: Spearman ρ before and after each SEC enforcement action related to Reg FD. Highlighted are the events responsible for the sharpest declines in inferred leakage.

ID	Announcement Date	Event / Company Investigated	Spearman ρ before	Spearman ρ after	Impact
A	Dec 20, 1999	Proposed Reg FD	-0.65%	-3.05%	-2.40%
B	Aug 15, 2000	Published Reg FD	6.03%	1.01%	-5.02%
C	Oct 23, 2000	Reg FD into effect	5.73%	-5.42%	-11.15%
1/2/3	Nov 25, 2002	Raytheon/Secure/Siebel	7.62%	3.27%	-4.35%
4	Sep 9, 2003	Schering-Plough	6.54%	-1.91%	-8.45%
5	Jun 29, 2004	Siebel Systems 2	-1.02%	1.29%	+2.31%
6	Sep 16, 2004	Senetek	5.79%	10.72%	+4.93%
7	Mar 4, 2005	Flowserve	7.04%	0.92%	-6.12%
8	Sep 5, 2007	EDS	8.27%	6.34%	-1.93%
9	Sep 24, 2009	ACL	13.29%	0.57%	-12.72%
10	Mar 3, 2010	Presstek	2.56%	9.49%	+6.93%
11	Oct 21, 2010	Office Depot	9.10%	0.47%	-8.63%
12	Nov 22, 2011	Fifth Third Bancorp	-1.21%	1.31%	+2.52%
13	Sep 6, 2013	First Solar	11.97%	3.36%	-8.61%
	Mean		5.79%	2.03%	-3.76%

different in the two periods: effective enforcement should reduce insiders' propensity to leak private information, lessening the alignment between Prior Day Abnormal Return and Earnings Surprise. The results, provided in Table 6.11, evidence a considerable but uneven drop in correlation: in 10 out of 14 distinct dates, leakage drops after SEC action. Mean leakage drops from 5.79% Spearman ρ in the three months prior to an SEC announcement, to 2.03% in the three months after. We provide the histograms and associated boxplot of the correlation distributions directly preceding and following enforcement actions in Figure 6.4. The absence of overlap in the boxplot notches provides evidence at the 95% confidence threshold that the medians of the two distributions differ (Chambers et al 1983), and supports the hypothesis that Reg FD enforcement produces a statistically significant drop in leakage.

We also measure how leakage behaviour varies cross-sectionally (Table 6.12). Tiering firms on the basis of Carhart factors (top 30% versus bottom 30% of each metric), we find that leakage is more pronounced amongst high beta, highly capitalised, low-momentum value stocks. The immediate impact of SEC enforcement on leakage is sharpest amongst highly capitalised, momentum-driven stocks, and is relatively in-

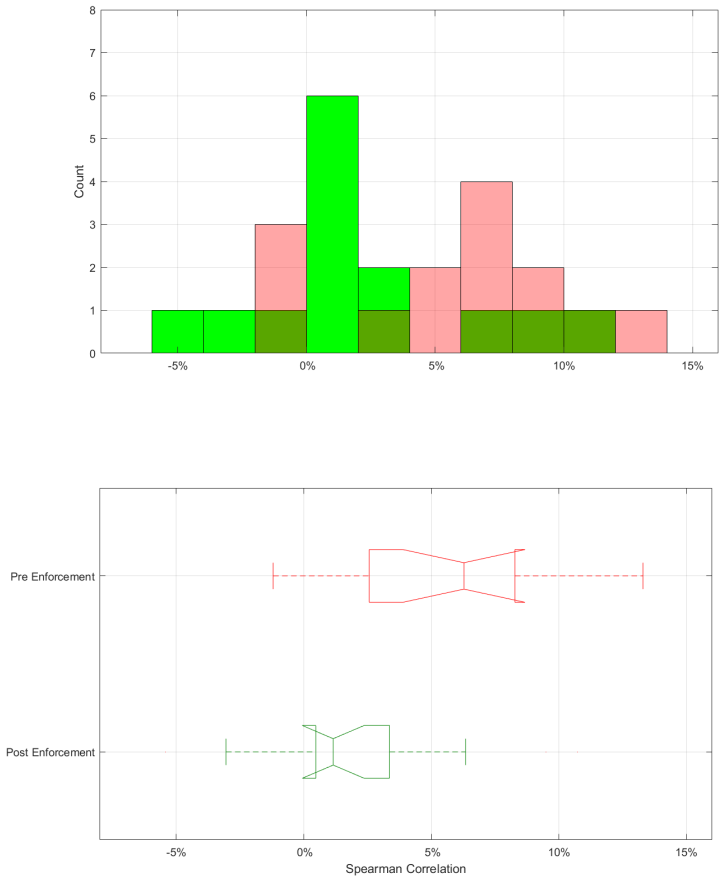

Figure 6.4: Histogram (Upper Panel) and corresponding Tukey boxplot (Lower Panel) for the distribution of correlation in the 3 months directly preceding and directly following an enforcement action. Whiskers cover 1.5 times the interquartile range. The absence of overlap in the boxplot notches provides evidence at the 95% confidence threshold that the medians of the 2 distributions differ.

Table 6.12: Mean inferred leakage and deterrence effect (impact, as defined in Table 6.11) as a function of Carhart factors.

Carhart Factor	Inferred Leakage	Deterrence Effect
Baseline	4.86%	7.85%
High β	5.40%	12.90%
Low β	2.81%	12.88%
Big	6.08%	19.62%
Small	4.06%	10.91%
High P/B	3.17%	12.32%
Low P/B	5.25%	12.59%
High Momentum	2.40%	15.30%
Low Momentum	5.01%	7.88%

sensitive to market β and price-to-book ratios.

6.5.2 Measuring Deterrence in Undisturbed Markets

Our second hypothesis for investigation is based on the SEC chairwoman's claim that insiders have "short memories". If the regulator does not enforce its rules with some regularity, potential wrongdoers will interpret the resulting inactivity as a reduced risk of apprehension and increase their production of wrongdoing. We thus conjecture that SEC enforcement actions that take place after a significant period of inactivity will have a particularly notable deterrent effect on leakage.

We measure the time period preceding each SEC action in Table 6.13, and select for further study those events preceded by SEC inactivity above the mean SEC inactivity (422 days).[17] This offers an arguably distinctive cluster of enforcement actions to study, since it divides our sample at the point of the largest difference in days between any two events, to include the First Solar announcement (preceded by 654

[17]We measure the period of SEC inactivity preceding the proposal of Reg FD as starting at the date of the Supreme Court's decision in US v. O'Hagan (June 25, 1997), since this was the most recent preceding event relating to questions of leakage. As events 1, 2 and 3 were announced on the same day, they are counted as a single Reg FD event for the purposes of calculating the mean inactivity.

Table 6.13: Time elapsed since last event for each SEC enforcement action under Reg FD. We highlight in gray the events following an inactive period above the mean inactivity of 422 days.

ID	Announcement Date	Event / Company Investigated	Preceding SEC Inactivity
A	Dec 20, 1999	Proposed Reg FD	903
B	Aug 15, 2000	Published Reg FD	239
C	Oct 23, 2000	Reg FD into effect	69
1/2/3	Nov 25, 2002	Raytheon/Secure/Siebel Systems	763
4	Sep 9, 2003	Schering-Plough	288
5	June 29, 2004	Siebel Systems 2	294
6	Sep 16, 2004	Senetek	79
7	Mar 4, 2005	Flowserve	169
8	Sep 5, 2007	EDS	915
9	Sep 24, 2009	ACL	750
10	Mar 3, 2010	Presstek	160
11	Oct 21, 2010	Office Depot	232
12	Nov 22, 2011	Fifth Third Bancorp	397
13	Sep 6, 2013	First Solar	654

days of SEC inactivity) and exclude Fifth Third Bancorp (397 days).

We find evidence that enforcement undertaken after long periods of regulatory inactivity lead to the most reliable drops in insider leakage. In Figure 6.5 we follow the time series of Spearman correlation between Prior Day Abnormal Return and Earnings Surprise, batched by year. Circles in Figure 6.5 denote SEC actions following lengthy periods of inactivity (events A, 1/2/3, 8, 9 and 13 per Table 6.13), and each constitutes a local peak in estimated leakage. Long periods of SEC inactivity systematically produce local maxima in leakage: insiders' memories (and thus the deterrence effect) of SEC enforcement actions appear to fade over time. The SEC chairwoman's remarks quoted at the beginning of this chapter appear correctly founded. These findings should however be interpreted as evidencing rational calculations by insiders as to the SEC's enforcement activity rather than insider memory loss.

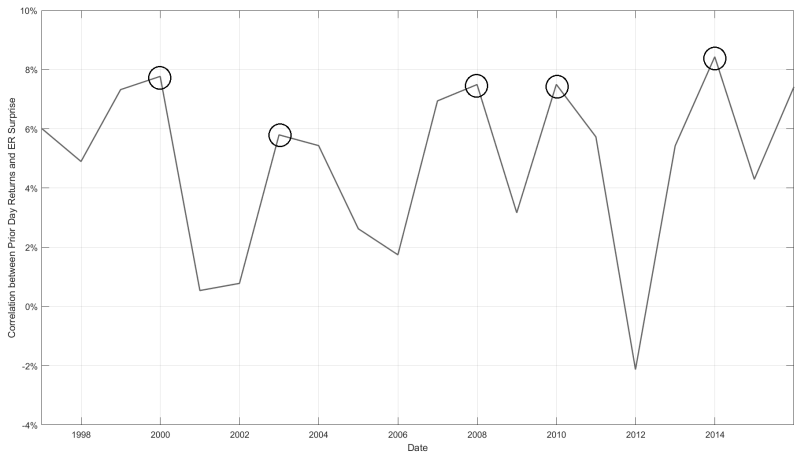

Figure 6.5: Rank correlation between Prior Day Abnormal Returns and Earnings Surprise on the S&P500 as a function of time, aggregated by year.

6.5.3 The Effect of SEC Escalations

Our third hypothesis is that SEC escalations in sanctioning will cause insiders to immediately adjust upwards their expected costs of engaging in leakage, producing a notable reduction in leakage. To test it, we review the SEC's enforcement actions in Tables 6.9 and 6.10 to determine when the SEC escalated its sanctioning powers. We categorise the following SEC actions as escalations:

- The proposal of Reg FD in December 1999, since it signalled the SEC's intentions to actively intervene to reduce information leakage in the market (event A).
- The Schering-Plough settlement in September 2003, which was the first time the SEC fined a corporate insider under Reg FD (event 4).
- The Office Depot action in October 2010, the first (and still only) time that the SEC stipulated that the insiders responsible for the leak were not permitted to seek reimbursement from their employer for the fines they had to pay (event 11).

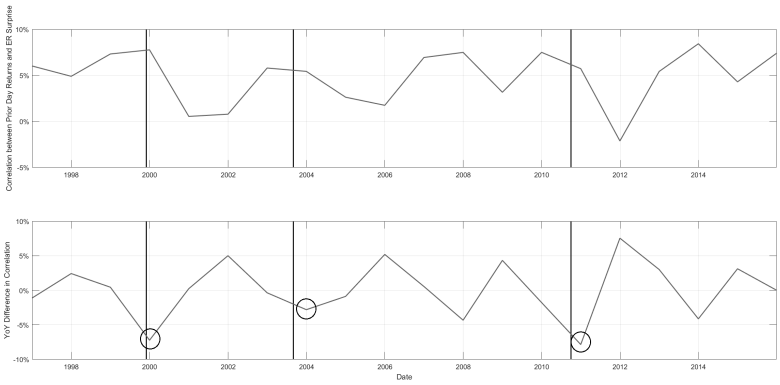

Figure 6.6: Spearman correlation between Prior Day Abnormal Returns and Earnings Surprise on the S&P500 as a function of time, aggregated by year (Upper Panel), and annual difference in correlation (Lower Panel). SEC escalation events (as defined in Section 6.5.3) denoted by transversal lines.

These three cases all indicated clear SEC intent (through action) to increase the expected cost of insider leakage, which would be expected to increase deterrence.

Notably, the correlation measure that we use as an estimate of information leakage only dips below 2% three times in the entire time series, each time in the aftermath of an SEC escalation. Escalations are also associated with steep annual declines in leakage (circled in Figure 6.6). The SEC escalation in Office Depot, which required insiders to pay their fines out of personal (not corporate) funds, is associated with the only instance where our leakage metric turned negative. Similarly to Cohen et al (2012), who find evidence that insiders who trade opportunistically in their own stock reduce their activity after the SEC releases news about insider trading prosecutions, this finding may represent the selective disclosure analogue: insiders reducing their leakage activity after SEC escalations.

6.5.4 The Memory of Enforcement

To evaluate the persistence of SEC enforcement, we compare the unconditional distribution of the Spearman correlation time series with its conditional counterpart, conditioned on observing an SEC escalation in the preceding N quarters. If the SEC

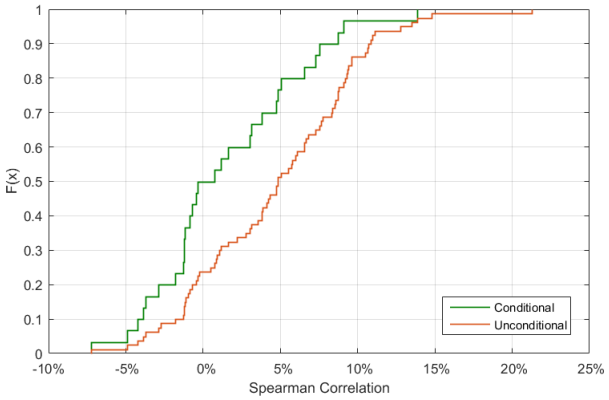

Figure 6.7: Empirical cumulative distribution functions of unconditional correlation and correlation conditioned on SEC escalation in the preceding N=10 quarters.

escalations are effective, conditioning on them should significantly alter the distribution. Denoting by $\{\rho_N{}_{t=1}^{n_1}\}$ the subset of quarterly Spearman correlations conditioned on SEC escalations in the last N quarters and $\{\rho_{t=1}^{n_2}\}$ the full set of quarterly Spearman correlations over the 21-year window, we can compute their empirical cumulative distribution functions $F_N(z)$ and $F(z)$. As a reminder, the two-sample Kolmogorov-Smirnov (K-S) test (Massey, 1951) evaluates the null hypothesis that the distributions generating both samples have identical cumulative distribution functions, by evaluating the K-S statistic:

$$\gamma_N = \left(\frac{n_1 n_2}{n_1 + n_2}\right)^{1/2} \sup_{-\infty<z<\infty} |F_N(z) - F(z)| \tag{6.4}$$

The limiting distribution of γ_N provides percentile thresholds above which we reject the null hypothesis, and may therefore infer that the information environment has been altered for N quarters. As an example of this approach, we provide the empirical cumulative distribution functions of both unconditional correlation and correlation conditioned on SEC escalations within the past $N = 10$ quarters in Figure 6.7. The supremum of the vertical gap between the two distributions is substantial, providing evidence of a statistically significant change in market behaviour for 10 quarters after SEC escalations.

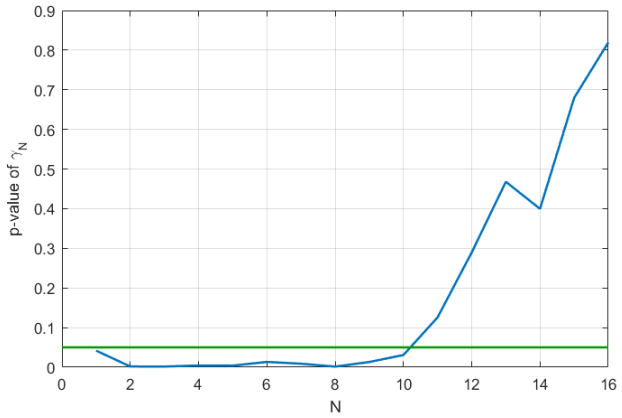

Figure 6.8: Significance of the Kolmogorov-Smirnov test statistic γ_N as a function of the number of quarters N. The horizontal line traces the 5% significance threshold.

In Figure 6.8 we provide the p-values associated with γ_N for $1 \leq N \leq 16$. Our first observation is that SEC escalation produces an instantaneous and measurable deterrent effect: the conditional and unconditional distributions differ significantly for all $N \leq 10$, suggesting that the market's behaviour vis-a-vis earnings announcements is altered for up to 30 months but no further. This provides an upper bound on the market's memory of SEC escalation: past the 30-month mark, the distributions become statistically indistinguishable.

By measuring the p-value associated with each γ_N, we can identify the number of quarters N over which the information environment is most heavily affected. The distributional gap, as measured by the significance of the K-S test statistic, is maximised for $N = 8$. By our best estimate, the deterrent effect of SEC escalations under Reg. FD lasts on average 24 months.

We provide the histograms of the conditional and unconditional correlation distributions and associated boxplot (Figure 6.9) for $N = 8$. The conditional distribution presents periods of muted aberrant activity, in contrast to the more elevated unconditional baseline. The distributions are indeed discernibly different: the median correlation in periods following an SEC escalation is slightly negative (-0.8%), a significant divergence from the 4.8% median observed in the unconditional baseline. The

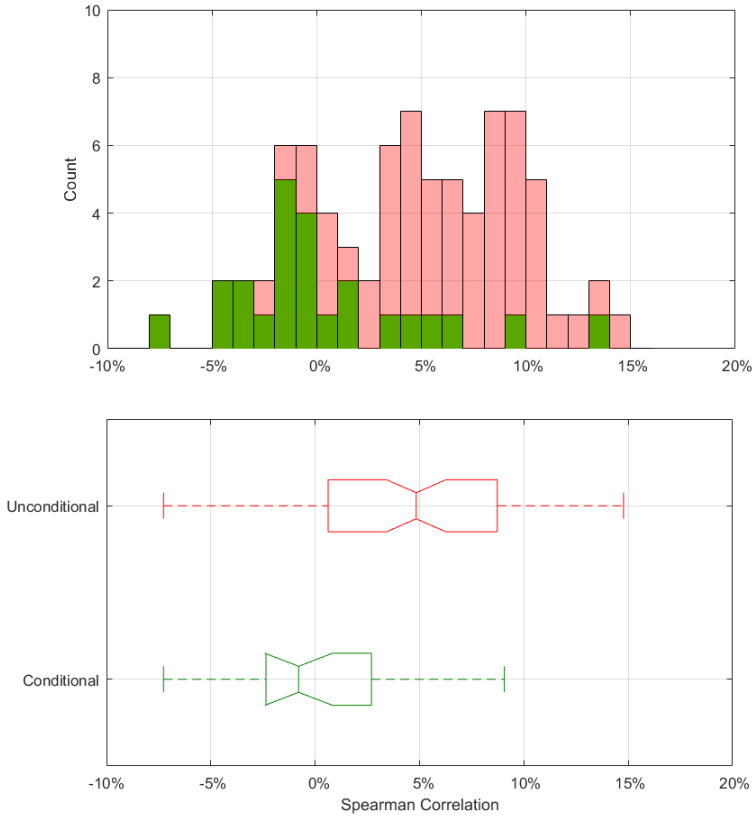

Figure 6.9: Histogram (Upper Panel) and corresponding Tukey boxplot (Lower Panel) for the distribution of correlation, conditioned on an SEC escalation in the past 8 quarters (in green) vs. the unconditional distribution (in red). Whiskers cover 1.5 times the interquartile range.

absence of overlap in boxplot notches supports the belief that the medians of the two distributions are statistically different. This evidence lends further credibility to the former SEC chairwoman's claim regarding "short memories": public enforcement is required on a regular basis to deter misbehaviour.

6.6 Summary

SEC enforcement under Regulation Fair Disclosure has an immediate deterrent effect on insider leakage, with enforcement actions undertaken after long periods of SEC inactivity being more significant. SEC escalations - instances in which the SEC deploys stronger sanctions for Reg FD violations than it ever had previously - produced the most credible deterrence of insider leakage, altering the information environment for approximately 24 months. These findings offer a dynamic evaluation of the effectiveness of Reg FD, and underscore the importance of recurring interventions by capital market regulators wishing to keep markets clean.

Chapter 7

Conclusion

The research in this thesis deliberately spans a wide range of topics in finance, to showcase the benefits of an algorithmic, rigorously quantitative approach.

7.1 Price Discovery

Faced with an overwhelming abundance of potentially informative data, liquidity takers must devise methods of selecting or generating salient features before making trading decisions.

7.1.1 Feature Selection

Financial markets present more data sources than can reasonably be studied. When deciding whether to invest in a particular company, the trader must consider the price history of comparable assets, the fundamentals of the business, newsflow surrounding its activities as well as professional opinions. To be successful, the buy-side must devise methods for sifting through financial complexity and pruning features. Our work in Chapter 3 finds that, for the purposes of feature selection, Gaussian Process regression with Automatic Relevance Determination kernels can identify a more informative subset of features than conventional approaches to shrinkage such as LASSO.

7.1.2 Feature Extraction

Data fusion for stock market prediction is extremely challenging, and much of its difficulty lies in engineering the appropriate predictive features. Despite its ubiquity, much of modern technical analysis still relies on features of sometimes arcane provenance. The convolutional layers in the ConvNets of Chapter 4 obviate the need to hand-design or hand-pick features, functioning as bespoke feature extractors. By training models on recent data, a convolutional replacement to technical analysis has the potential to adapt to changes in market behaviour and replace the dogmatic, demonstrably inaccurate chartist approach.

7.2 Liquidity Provision

Today's market makers operate in a heavily-competed space, where rational clients obtain bid-offer prices from a wide range of sources via quote aggregators and only trade at the most advantageous price. Adverse selection is a natural consequence of this: a market maker's quote will only get traded if they showed a better bid or offer price than all their competitors, and this may well be the result of erroneous pricing.

Some predatory liquidity takers may design their business around such types of opportunism, looking for mistakes to arbitrage. By modelling the trading activity of individual counterparties as Gaussian Processes, Chapter 5 offered a tentative framework for evaluating a client's propensity towards adverse selection. More generally, it offers a comprehensive picture of the conditions under which a market maker's counterparties act: this may present opportunities beyond the tiering of their worth, such as the identification of predictive signals in their quote requests.

7.3 Financial Regulation

Machine learning is already widely adopted amongst buy-side and sell-side systematic hedge funds, and we expect that many of the algorithms discussed in this thesis - and

in our humble estimation, far more sophisticated variants thereof - are likely trading on financial markets today.

By contrast, quantitative methods are not as commonplace in policy discussions, and empirical research in the academia of law and finance may benefit from the adoption of more advanced methods. In Chapter 6, we deployed several widely accepted statistical methodologies (t-tests, Kolmogorov-Smirnov tests) alongside function learning via Gaussian Processes, to identify a stock market anomaly around US corporate earnings announcements. This anomaly's instantaneous response to high-profile Reg FD investigations and its absence from European jurisdictions leads us to hypothesise that we're witnessing a footprint of informed trading in US financial markets.

7.4 Further Work

Several strands of research in this thesis bear the potential for extension.

7.4.1 The Price Space Representation

The feature ranking in Chapter 3 showcased the informativeness of options markets in predicting near-term market movements. We can conceive of logical reasons why this might be the case:

- A rational agent possessing material information - whether through due diligence or less conventional means such as those of Chapter 6 - should deploy their edge in leveraged instruments to maximise their gains. Options markets are the most liquid means of achieving that leverage. As such, accumulations of call or put option positions with near-term expiries (which we termed *Directionality* in Chapter 3) deserve closer inspection.

- Large outstanding positions in near-dated straddles will provoke mutually reinforcing behaviours from both the liquidity takers and market makers of those

options. Gamma hedging by the buy side will constrain any deviation from the straddle's strike, pinning prices locally (*Viscosity* in the terminology of Chapter 3). This favours the sell side, whose rational self-interest may even include buying or selling the underlying asset outright to keep its price near the strike, to minimise option payouts.

These considerations encourage a different mindset around price prediction. Rather than seeking patterns in return space, our work encourages a closer study of the importance of price levels, and the patterns that may emerge around specific price loci. Options markets aside, psychological factors are likely to strengthen our case: financial media - and by extension, speculative frenzies - tend to focus on the price level of assets (WTI crude oil at $30/barrel or the S&P500 at 2000 in 2016, for example), not their returns.

7.4.2 Neural Network Architectures for Price Prediction

Even shallow convolutional neural networks, comprised of a single convolutional layer, were able to produce considerably more powerful classifiers than a wide range of alternative algorithms, as shown in Chapter 4. The simplicity of our architecture merely shows the potential for ultra-parametric models to supplant technical analysis - we do not for a moment expect shallow convolution to be optimal. In the context of computer vision, accuracy improved dramatically through the expansion of model depth.

- AlexNet (2012): 5 convolutional layers, 3 dense layers - 16.4% error rate on ILSVRC-2012.
- VGG19 (2014): 16 convolutional layers, 3 dense layers - 7.3% error rate on ILSVRC-2014.

- ResNet (2015): 152 convolutional layers, 1 dense layer - 3.6% error rate on ILSVRC-2015.

The gains in financial prediction will likely never match those of a physical setting like image recognition - after all, successful trading strategies arbitrage away the financial regularities being exploited. Nevertheless, we believe that an architecture tailored for financial prediction bears tremendous potential for identifying those patterns and, through rolling training windows, adapting to changes in market regime - unlike the dogmatic approach of chartism.

Our findings in this domain were spurred by a concern over the obsolescence of technical analysis. As such, we restricted our input space to open, close, high and low price data, at the frequency it is most commonly applied in mainstream media (daily). Active research in this field has begun examining whether convolution may also be applied at higher frequencies, using the full depth of the order book. It is plausible that order book imbalances would exhibit visual regularities that presage a 'break' in one direction or the other, and that the instinctively visual interpretation of ConvNets could help the high-frequency trader make short-term predictions.

7.4.3 Tracking the Flow of Material Information

Earnings announcements provided a rich dataset for scrutinising the behaviour of financial markets around a material, binary event: the transition of information from private to public. Yet some of the most important corporate events fall outside of quarterly disclosures: technological breakthroughs, production delays or leadership changes are all impactful instances that are certainly known to insiders well in advance. We hope to see the framework of Chapter 6 applied to a wider range of corporate disclosures. We see no reason to presume that insiders would view leakage of earnings information any differently from the selective disclosure of other material facts.

On a side note - while our perspective on this topic may appear deeply cynical, it is sincerely intended as a realist's view of financial markets. Much as the Grossman and Stiglitz paradox that prefaced Chapter 6, we need only presume the rational self-interest of insiders to suppose that information leakage would occur. Ours is but one of many studies to suggest that the information environment is not a level playing field.

7.4.4 Counterparty Knowledge

Dealing with that information asymmetry is an intrinsic part of market making. Liquidity providers live or die by their profit margin, and as such have long accepted that their counterparties are not all equally informed. Chapter 5 provided a means of identifying adverse selectors, whose OTC deal flow could then be tactically disincentivised through less competitive pricing.

That being said, the lion's share of this thesis has focused on the search for predictive signals. Even our foray into financial regulation produced, arguably, a predictive tool. Upon identifying a pre-earnings announcement drift, a savvy liquidity taker could opt to follow suit, betting on the earnings surprise that they've inferred from stock price movements in their final days before earnings disclosures.

In much the same way, the liquidity provider's visibility on the trading patterns of multiple counterparties must be construed as a vantage point. Predictable counterparties may help the market maker both manage their risk better, and anticipate market movements over short horizons. Liquidity provision in US equities is dominated by a small number of high-frequency players,[1] many of whom already employ cutting-edge machine learning to generate their statistical edge. Market making datasets are immensely valuable for that reason, and though public, academic research may have

[1]Citadel Securities and Virtu Financial each serviced approximately 20% of the entire US equity market in 2016, per research by the Tabb Group reported in the Financial Times, April 25th 2017.

limited access to them, we are confident in the belief that private sector researchers keenly investigate the very same considerations we have.

Appendix A

Data Provenance

We include in this appendix the provenance of all datasets used throughout the thesis, in the interest of transparency and reproducibility by the interested reader. Most of our data was sourced from Reuters Datastream 5.1, with access kindly provided by the Oxford-Man Institute of Quantitative Finance. We provide below the Datastream codes to retrieve data chapter-by-chapter.

A.1 Datasets for Chapter 3

Chapter 3 employed a variety of datatypes - technicals, sentiment, options metrics and broker recommendations. Datasets sourced from Reuters Datastream are provided in Table A.1.

The technical indicators used in Chapter 3 (lagged returns, 50d EMA, MACD and Signal Line) were all produced from arithmetic computations on the Close Price of the S$P500 stock market index.

As the chapter aimed to rank the relevance of various data domains, we did not produce our own sentiment analysis and instead relied on sentiment scores from Twitter and Stocktwits compiled by PsychSignals, a leading provider of social media data for financial markets. While publicly available at the time of writing the associated research paper (Q3 2015), this dataset has since become private - once a free provider of data, PsychSignals now charges for access.

Table A.1: Datastream codes for the Data Fusion research.

Data Description	Datastream 5.1 Mnemonic
Daily Adjusted Close Price of the Index	PI
S&P500 Call Open Interest	SPX[mm][yy][k]C(OI)
S&P500 Put Open Interest	SPX[mm][yy][k]C(OI)

The options market metrics were derived through the application of Equations (3.3) and (3.5) to full matrices of daily Call and Put data tracking Open Interest by strike and by expiry. As such, each strike and expiry must be collected as a separate column. The mnemonic in Datastream for retrieving e.g. Call Open Interest data uses the formulation: SPX[mm][yy][k]C(OI), where mm provides the expiry month, yy the expiry year and k is the strike. For Call options expiring in December 2014 at a strike of 1800, the mnemonic would thus be SPX12141800C(OI).

Finally, the broker recommendation data was sourced through web scraping of *Yahoo Finance* pages. A separate search was conducted for each constituent ticker of the S&P500 using the url:
https://finance.yahoo.com/quote/[*ticker*]/analysis?p=[*ticker*].
Replacing *ticker* with e.g. AAPL provides a list of all Upgrade, Downgrade and Hold actions on Apple Inc.

A.2 Datasets for Chapter 4

Chapter 4 was built on Open, Close, Daily High and Daily Low prices for stocks in the S&P500. All of these datatypes are available in Reuters Datastream, with their mnemonics provided in Table A.2.

A.3 Datasets for Chapter 5

Unlike all the other research threads in this thesis, Chapter 5 involved the proprietary trading data of a leading market maker. Our fully-anonymised dataset of foreign

Table A.2: Datastream codes for the ConvNet research.

Data Description	Datastream 5.1 Mnemonic
Daily Open Price	PO
Daily Close Price	P
Daily High Price	PH
Daily Low Price	PL

exchange trading activity was provided by BNP Paribas. We paired their trade logs with EURUSD tick data provided publicly by Pepperstone, an online foreign exchange broker, to produce our intraday return and volatility metrics.

A.4 Datasets for Chapter 6

Chapter 6 explored regulatory effectiveness, employing a wide range of firm-specific and macro data summarised in Table A.3. For each firm we computed Carhart alphas using 3-year rolling regressions of the Carhart factors of market beta, market capitalisation, price-to-book ratio and momentum. These models measure sensitivity to a stock's excess return over the risk-free rate, typically measured through the 1M deposit rate in the appropriate jurisdiction (in our studies, US, UK and France). Finally, to derive *Earnings Surprise*, we also retrieve the reported Earnings per Share, Consensus Earnings per Share and Quarterly Earnings Announcement Date for each firm. As noted in Section 6.4.1, some US firms report their earnings before the market opens (BMO), and some report after market closure (AMC). This needed to be tracked on an individual basis: the first trading day on which the earnings information is public is either the announcement date itself (in the case of BMO) or the trading day directly following the announcement (in the case of AMC).

Table A.3: Datastream codes for the financial regulation research.

Data Description	Datastream 5.1 Mnemonic
Daily Adjusted Close Price of the Index	PI
US Dollar 1M Deposit Rate	ECUSD1M(IR)
UK Sterling 1M Deposit Rate	ECUKP1M(IR)
Euro 1M Deposit Rate	ECEUR1M(IR)
Daily Close Price	P
Market Capitalisation	MV
Price-to-Book Ratio	PTBV
Reported Earnings per Share	RIEPS
I/B/E/S Consensus Earnings per Share	I1MN
Quarterly Earnings Announcement Date	W05901,W05902,W05903,W05904

Bibliography

[1] Agrawal, A. and Tandon, K. (1994). Anomalies or Illusions? Evidence from Stock Markets in eighteen Countries. *Journal of International Money and Finance*, 13(1), 1994.

[2] Akbas, F., Felix, M. and Wintoki, M. B. (2016). Director Networks and Informed Traders. *Journal of Accounting and Economics*, 62, pp. 1-23.

[3] Allen, F. and Karjailainen, R. (1999). Using Genetic Algorithms to find Technical Trading Rules. *Journal of Financial Economics*, 51, pp. 245-271.

[4] Anthony, J. H. (1998). The Interrelation of Stock and Options Market Trading-Volume Data. *The Journal of Finance*, 43(4), 1998.

[5] Antipova, G., Berrani, S. A. and Dugelay, J. L. (2016). Minimalistic CNN-based Ensemble Model for Gender Prediction from Face Images. *Pattern Recognition Letters*, 70, pp. 59-65.

[6] Apel, R. and Nagin, D. S. (2011). General Deterrence. *Oxford Handbook of Crime and Criminal Justice.*

[7] Apel, R. (2013). Sanctions, Perceptions and Crime: Implications for Criminal Deterrence. *Journal of Quantitative Criminology*, 29, pp. 67-101.

[8] Armour, J., Bengtzen, M. and Enriques, L. (2017). Investor Choice in Global Securities Markets. *The New Special Study of the Securities Markets.*

[9] Avellaneda, M. and Stoikov, S. (2008). High-frequency Trading in a Limit Order Book. *Quantitative Finance*, 8, pp. 217-224.

[10] Ball, R. and Brown, P. (1968). An Rmpirical Evaluation of Accounting Income Numbers. *Journal of Accounting Research*, 6(2), pp. 159-178.

[11] Ball, R. and Kothkari, S. P. (1991). Security Returns around Earnings Announcements. *The Accounting Review*, 66(4), pp. 718-738.

[12] Ballings, M., Van den Poel, D., Hespeels, N. and Gryp, R. (2015). Evaluating Multiple Classifiers for Stock Price Direction Prediction. *Expert Systems with Applications*, 42(20), pp. 7046-7056.

[13] Beaver, W. H, McNichols, M. F. and Wang, Z. Z. (2017). The Information Content of Earnings Announcements: New Insights from Intertemporal and Cross-Sectional Behavior. *Review of Accounting Studies*, 23(1) pp. 1-41.

[14] Becker, G. S. (1974). Crime and Punishment: an Economic Approach. *Journal of Political Economy*, 76(2), pp. 169-217.

[15] Bengtzen, M. (2017). Private Investor Meetings in Public Firms: The Case for Increasing Transparency. *Fordhman Journal of Corporate & Financial Law*, Vol. XXII (forthcoming).

[16] Bernard, V. L. and Thomas, J. K. (1990). Evidence that Stock Prices do not fully reflect the Implications of Current Earnings for Future Earnings. *Journal of Accounting and Economics*, 13, pp. 305-340.

[17] Bhattacharya, U. and Daouk, H. (2002). The World Price of Insider Trading. *Journal of Finance*, 57(1), pp. 75-108.

[18] Biller, B. and Corlu, C. G. (2012). Copula-based multivariate Input Modeling. *Surveys in Operations Research and Management Science* 17(2), 2012.

[19] Blume, L., Easley, D. and O'Hara, M. (1994). Market Statistics and Technical Analysis: The Role of Volume. *Journal of Finance*, 49(1), pp. 153-181.

[20] Buffa, A. M. and Nicodano, G. (2008). Should Insider Trading Be Prohibited when Share Repurchases Are Allowed? *Review of Finance* 12(4), p.735.

[21] Butler, A. W. and Gurun, U. G. (2012). Educational Networks, Mutual Fund Voting Patterns, and CEO Compensation. *Review of Financial Studies*, 25(8), pp. 2533-2562.

[22] Campbell, C., Li, Y. and Tipping, M. (2002). Bayesian Automatic Relevance Determination algorithms for classifying gene expression data. *Oxford University Press.*

[23] Carhart, M. M. (1997). On Persistence in Mutual Fund Performance. *Journal of Finance*, 52, pp. 57-82.

[24] Chalfin, A. and McCrary, J. (2017). Criminal Deterrence: A Review of the Literature. *Journal of Economic Literature*, 55(1), pp. 5-48.

[25] Chambers, J. M., Cleveland, W. S., Kleiner, B. and Tukey, P. A. (1983). Comparing Data Distributions. *Graphical Methods for Data Analysis*, pp. 60-63.

[26] Chapados, N. and Bengio, Y. (2007). Forecasting and Trading Commodity Contract Spreads with Gaussian Processes. *13th International Conference on Computing in Economics and Finance, 2007.*

[27] Christophe, S. E., Ferri, M. G. and Angel, J. J. (2004). Short-selling prior to Earnings Announcements. *Journal of Finance*, 59, pp. 1845-1876.

[28] Chung, J, Gulcehre, C., Cho, K. and Bengio Y. (2014). Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *NIPS 2014 Deep Learning and Representation Learning Workshop.*

[29] Clarke, S. (2013). Insider Dealing: Law and Practice. *Oxford University Press*, p. 30.

[30] Coffee Jr, J. C. (1997). Is Selective Disclosure Now Lawful? *218 New York Law Journal, 5.*

[31] Coffee Jr, J. C., and Goldschmid, H. (2016). The Scholar as Realistic Reformer. *116 Columbia Law Review, 1.*

[32] Cohen, L., Frazzini, A. and Malloy, C. (2008). The Small World of Investing: Board Connections and Mutual Fund Returns. *Journal of Political Economy*, 116.

[33] Cohen, L., Malloy, C. and Pomorski, L. (2012). Decoding Inside Information. *Journal of Finance*, 67, pp. 1009-1043.

[34] Cornell, B. and Landsman, W. R. (1989). Security Price Response to Quarterly Earnings Announcements and Analysts' Forecast Revisions. *The Accounting Review*, 64(4), pp. 680-692.

[35] Cybenko, G. (1989). Approximations by superpositions of sigmoidal functions. *Mathematics of Control, Signals and Systems*, 2(4), pp. 303-314.

[36] Dixon, M., Klabjan, D. and Bang, J.H. (2017). Classification-based Financial Markets Prediction using Deep Neural Networks. *Algorithmic Finance*, 6(3-4), pp. 67-77.

[37] Dixon, M. (2018). Sequence Classification of the Limit Order Book using Recurrent Neural Networks. *Journal of Computational Science*, 24, pp. 277-286.

[38] Dolgopolov, S. (2004). Insider Trading and the Bid-Ask Spread: A Critical Evaluation of Adverse Selection in Market Making. *Capital University Law Review 33*, pp. 83-180.

[39] Edwards, R. D. and Magee, J. (1948). Technical Analysis of Stock Trends. *Vision Books*.

[40] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least Angle Regression. *Annals of Statistics*, 32(2), pp. 407-499.

[41] Falk, H. and Levy, H. (1989). Market Reaction to Quarterly Earnings' Announcements: A Stochastic Dominance Based Test of Market Efficiency. *Management Science*, 35(4), pp. 425-446.

[42] Fama, E. F. and French, K. (1992). The Cross-Section of Expected Stock Returns. *Journal of Finance*, 47, pp. 427-466.

[43] Farrell, M.T. and Correa, A. (2007). Gaussian Process Regression Models for Predicting Stock Trends. *MIT University Technical Report*.

[44] Fischer, T. and Krauss, C. (2017). Deep Learning with Long Short-Term Memory Networks for Financial Market Predictions. *European Journal of Operational Research*, 270(2), pp. 654-669.

[45] Freis, J. H. (1996). An Outsider's Look into the Regulation of Insider Trading in Germany: A Guide to Securities, Banking, and Market Reform in Finanzplatz Deutschland. *19 B.C. Int'l & Comp. L. Rev., 1.*

[46] Fried, Jesse M. (2014). Insider Trading via the Corporation. *162 University of Pennsylvania Law Review, 801.*

[47] Garman, M. B. (1976). Market Microstructure. *Journal of Financial Economics*, 3(3), pp. 257-275.

[48] Geerken, M. R. and Grove, W. R. (1975). Deterrence: Some Theoretical Considerations. *Law & Society Review*, 9(3), pp. 497-513.

[49] Ghoshal, S. and Roberts, S. (2016). Extracting Predictive Information from heterogeneous Data Streams using Gaussian Processes. *Algorithmic Finance*, 5(1-2), pp. 21-30.

[50] Ghoshal, S. and Roberts, S. (2017). Reading the Tea Leaves: A Neural Network Perspective on Technical Trading. *KDD 2017, Mining and Learning from Time Series.*

[51] Ghoshal, S. and Roberts, S. (2018). Thresholded ConvNet Ensembles: Neural Networks for Technical Forecasting. *KDD 2018, Data Science in Fintech.*

[52] Glosten, L. R. and Harris, L. E. (1988). Estimating the Components of the Bid/Ask Spread. *Journal of Financial Economics*, 21(1), pp. 123-142.

[53] Glosten, L. R. and Milgrom, P. R. (1985). Bid, Ask and Transaction Prices in a Specialist Market with heterogeneously Informed Investors. *Journal of Financial Economics*, 14(1), pp. 71-100.

[54] Gomes, A., G. Gorton and L. Mudureira. (2007). SEC Regulation Fair Disclosure, Information, and the Cost of Capital. *Journal of Corporate Finance*, 13, pp. 300-334.

[55] Griffin, P. A., Lont, D. H. and Segal, B. (2011). Enforcement and Disclosure under Regulation Fair Disclosure: An Empirical Analysis. *Accounting and Finance*, 51, pp. 947-983.

[56] Griffin, J. M., Shu, T. and Topaloglu, S. (2012). Do Institutions Trade on Information from Investment Bank Connections? *Review of Financial Studies*, 25.

[57] Grossman, S. J. and Stiglitz, J. E. (1980). On the Impossibility of Informationally Efficient Markets. *The American Economic Review*, 70(3), pp. 393-408.

[58] Guéant, O., Lehalle, C. and Fernandez-Tapia, J. (2013). Dealing with the Inventory Risk: a Solution to the Market Making Problem. *Mathematics and Financial Economics*, 7(4), pp. 477-507.

[59] Del Guercio, D., Odders-White, E. R. and Ready, M. J. (2017). The Deterrence Effect of SEC Enforcement Intensity on Illegal Insider Trading: Evidence from Run-up before News Events. *Journal of Law and Economics*, 60(2), pp. 269-307.

[60] Haeberle, K. and Henderson, M. T. (2017). Making a Market for Corporate Disclosure. *Yale Journal on Regulation, 35.*

[61] Hastie, T., Tibshirani, R., Friedman, J. H. (2009). The Elements of Statistical Learning: Data Mining, Inference and Prediction. 2nd ed. New York: Springer.

[62] Hendershott, T., Livdan, D., Schürhoff, N. (2015). Are Institutions Informed about News? *Journal of Financial Economics*, 117.

[63] Hinton, G. E., Deng, L., Yu, D., Dahl, G. E., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., Kingsbury, B. (2012). Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. *IEEE Signal Processing Magazine*, 29(6), pp. 82-97.

[64] Ho, T. and Stoll, H. (1981). Optimal Dealer Pricing under Transactions and Return Uncertainty. *Journal of Financial Economics*, 9(1), pp. 47-73.

[65] Hochreiter, S. and Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8), pp. 1735-1780.

[66] Hu, G. X., Pan, J. and Wang, J. (2017). Early Peek Advantage? Efficient Price Discovery with Tiered Information Disclosure. *Journal of Financial Economics, forthcoming.*

[67] Hu, X., Wang X. and Xin, B. (2017). Insider Trading: Does Being a Neighbor of the Securities and Exchange Commission Matter? *Managerial and Decision Economics*, 38, pp. 144-165.

[68] Huddart, S., Ke, B. and Shi, C. (2007). Jeopardy, non-public Information, and Insider Trading around SEC 10-K and 10-Q filings. *Journal of Accounting and Economics*, 43, pp. 3-36.

[69] Irving, R. B. (1990). French Insider Trading Law: A Survey. *U. Miami Inter-Am L. Rev.*, 41.

[70] Jackson, H. E. and Roe, M. J. (2009). Public and Private Enforcement of Securities Laws: Resource-Based Evidence. *Journal of Financial Economics*, 93, pp. 207-238.

[71] Jaffe, J. F. (1974). Effect of Regulation Changes on Insider Trading. *Bell Journal of Economics*, 6, pp. 93-121.

[72] Jegadeesh, N. (1991). Seasonality in Stock Price Mean Reversion: Evidence from the US and the UK. *Journal of Finance*, 46(4), pp. 1427-1444.

[73] Jozefowicz, R., Zaremba, W and Sutskever, I. (2015). An Empirical Exploration of Recurrent Network Architectures. *Journal of Machine Learning Research*, 37, pp. 2342-2350.

[74] Kim, K. (2003). Financial Time Series Forecasting using Support Vector Machines. *Neurocomputing*, 55(1), pp. 307-319.

[75] Kingma, D. and Ba, J. (2015). Adam: A Method for Stochastic Optimization. *ICLR 2015*.

[76] Koch, A. S., Lefanowicz, C. E. and Robinson, J. R. (2013). Regulation FD: A Review and Synthesis of the Academic Literature. *Accounting Horizons: September 2013*, 27(3), pp. 619-646.

[77] Krauss, C., Do, X. A. and Huck, N. (2017). Deep neural networks, Gradient-boosted Trees, Random Forests: Statistical Arbitrage on the S&P500. *European Journal of Operational Research*, 259(2), pp. 689-702.

[78] Krizhevsky, A, Sutskever, I and Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems 2012*, 1106-1114.

[79] Kurov, A., Sancetta, A., Strasser, G. and Wolfe, M. H. (2016). Price Drift before US Macroeconomic News: Private Information about Public Announcements? *ECB Working Paper Series, May 2016.*

[80] Lakonishok, J. and Smidt, S. (1989). Are Seasonal Anomalies Real? A Ninety-Year Perspective. *Review of Financial Studies*, 1(4), 1989.

[81] La Porta, R., Lopez-de-Silanes, F. and Shleifer, A. (2006). What works in Securities Laws? *Journal of Finance*, 61(1), pp. 1-32.

[82] Latané, H. A. and Jones, C. P. (1977). Standardized Unexpected Earnings - A Progress Report. *Journal of Finance*, 32, pp. 1457-1465.

[83] Leuz, C. and Wysocki, P. D. (2016). The Economics of Disclosure and Financial Reporting Regulation: Evidence and Suggestions for Future Research. *Journal of Accounting Research*, 54, p. 525.

[84] Levenberg, A., Pulman, S., Moilanen, K., Simpson, E. and Roberts, S. (2014). Predicting Economic indicators from Web Text using Sentiment Composition. *Proceedings of ICICA 2014.*

[85] Levenberg, A., Simpson, E., Roberts, S. and Gottlob, G. (2013). Economic Prediction using heterogeneous Data Streams from the World Wide Web. *ECML Workshop on Scalable Methods in Decision Making.*

[86] Lintner, J. (1965). The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets. *Review of Economics and Statistics*, 47(1), pp. 13-37.

[87] Lo, A.W., Mamaysky, H and Wang, J. (2000). Foundations of Technical Analysis: Computational Algorithms, Statistical Inference and Empirical Implementation. *The Journal of Finance*, 55(4), 2000.

[88] Lohse, T., Pascalau, R. and Thomann, C. (2014). Public Enforcement of Securities Market Rules: Resource-Based Evidence from the Securities and Exchange Commission. *Journal of Economic Behavior and Organization*, 106, pp. 197-212.

[89] Lu, C. J., Lee, T.-S. and Chiu, C.-C. (2009). Financial Time Series Forecasting using Independent Component Analysis and Support Vector Regression. *Decision Support Systems*, 47(2).

[90] Macey, J. R. and O'Hara, M. (2009). Regulation and Scholarship: Constant Companions or Occasional Bedfellows? *Yale Journal on Regulation*, 26, p. 89.

[91] Macey, J. R. (2010). The Distorting Incentives Facing the U.S. Securities and Exchange Commission. *Harvard Journal of Law & Public Policy*, 33, p. 639.

[92] Manne, H. G. (1966). Insider Trading and the Stock Market. *The Free Press, New York.*

[93] Mason, S. J., Graham, N. E. (2002). Areas beneath the Relative Operating Characteristics (ROC) and Relative Operating Levels (ROL) Curves: Statistical

Significance and Interpretation. *Quarterly Journal of the Royal Meteorological Society*, 128, pp. 2145–2166.

[94] McLean, R. D. and Pontiff, J. (2016). Does Academic Research Destroy Stock Return Predictability? *Journal of Finance*, 71(1), pp. 5–32.

[95] Micchelli, C. A., Xu, Y. and Zhang, H. (2006). Universal Kernels. *Journal of Machine Learning Research*, 7, pp. 2651–2667.

[96] Moskowitz, T. J, Ooi, Y. H., Pedersen, L. H. (2012). Time series Momentum. *Journal of Financial Economics*, 104, pp. 228-250.

[97] Neely, C., Weller, P. and Dittmar, R. (1997). Is Technical Analysis in the Foreign Exchange Market Profitable? A Genetic Programming Approach. *Journal of Financial and Quantitative Analysis*, 32(4), pp. 405-426.

[98] Nikfarjam, A., Emadzadeh, E. and Muthaiyah, S. (2010). Text Mining Approaches for Stock Market Prediction. *2nd International Conference on Computer and Automation Engineering (ICCAE).*

[99] Norris, F. (2003). SEC Penalizes Schering-Plough over a Fair Disclosure Violation. *New York Times*, Sept. 10, 2003.

[100] Oh, C. K., Beck, J. L. and Yamada, M. (2008). Bayesian Learning Using Automatic Relevance Determination Prior with an Application to Earthquake Early Warning. *Journal of Engineering Mechanics*, 134(12), 2008.

[101] O'Hara, M. (2015). High Frequency Market Microstructure. *Journal of Financial Economics*, 116(2), pp. 257-270.

[102] Ou, P. and Wang, H. (2009). Modeling and Forecasting Stock Market Volatility by Gaussian Processes based on GARCH, EGARCH and GJR Models. *Proceedings of the World Congress on Engineering, 2009.*

[103] Psorakis, I., Roberts, S., Rezek, I. and Sheldon, B. (2012). Inferring Social Network Structure in Ecological Systems from spatio-temporal Data Streams. *Journal of the Royal Society Interface*.

[104] Psorakis, I., Roberts, S., Ebden, M. and Sheldon, B. (2011). Overlapping Community Detection using Bayesian non-Negative Matrix Factorization. *Physical Review*, 83(6).

[105] Qi, Y., Minka, T. P., Picard, R. W. and Ghahramani, Z. (2004). Predictive Automatic Relevance Determination by Expectation Propagation. *ICML 2004 Proceedings of the Twenty-First International Conference on Machine Learning*.

[106] Rasmussen, C. and Williams, C. (2006). Gaussian Processes for Machine Learning. *MIT Press*.

[107] Romaszko, L. (2015). Signal Correlation Prediction Using Convolutional Neural Networks. *Journal of Machine Learning Research: Workshop and Conference Proceedings*, 46, pp. 45-56.

[108] Sirignano, J. and Cont, R. (2018). Universal Features of Price Formation in Financial Markets: Perspectives from Deep Learning. *Available at SSRN: https://ssrn.com/abstract=3141294 or http://dx.doi.org/10.2139/ssrn.3141294*.

[109] Smith, M., Reece, S., Roberts, S., Psorakis, I. and Rezek, I. (2014). Maritime Abnormality Detection using Gaussian Processes. *Knowledge and Information Systems*, 38(3), pp. 717-741.

[110] Solomon, D. and Solters, E. (2015). What are We Meeting For? The Consequences of Private Meetings with Investors. *Journal of Law and Economics*, 58, pp. 325-355.

[111] Spooner, T., Fearnley, J., Savani, R. and Koukorinis, A. (2018). Market Making via Reinforcement Learning. *AAMAS 2018 Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 434-442.

[112] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15, pp. 1929-1958.

[113] Stickel, S. E. (1995). The Anatomy of the Performance of Buy and Sell Recommendations. *Financial Analysts Journal*, 51(5).

[114] Taylor, M. P. and Allen, H. (1992). The Use of Technical Analysis in the Foreign Exchange Market. *Journal of International Money and Finance*, 11(3), 1992.

[115] Tsantekidis, A., Passalis, N., Tefas, A., Kanniainen, J., Gabbouj, M. and Iosidis, A. (2017). Forecasting Stock Prices from the Limit Order Book using Convolutional Neural Networks. *19th IEEE Conference on Business Informatics (CBI)*, 1, pp. 7-12, 2017.

[116] Van Ness, B. F., Van Ness, R. A. and Warr, R. S. (2001). How well do Adverse Selection Components measure Adverse Selection? *Financial Management*, 30(3), pp. 77-98.

[117] Velikonja, U. (2015). Politics in Securities Enforcement. *Georgia Law Review*, 50, pp. 17-41.

[118] Womack, K. L. (1996). Do Brokerage Analysts' Recommendations have Investment Value? *The Journal of Finance*, 51(1).

[119] Wyatt, E. (2012). Responding to Critics, SEC Defends 'No Wrongdoing' Settlements. *New York Times*, Feb 23, 2012.

[120] Yan, X. S. and Zhang, Z. (2009). Institutional Investors and Equity Returns: Are Short-term Institutions Better Informed? *Review of Financial Studies*, 22, pp. 893-924.

[121] Zhang, Z., Zohren, S. and Roberts, S. (2018). DeepLOB: Deep Convolutional Neural Networks for Limit Order Books. *Arxiv preprint available at https://arxiv.org/pdf/1808.03668.pdf.*