Forecasting Financial Markets Using Linear, Nonlinear & Model Combination Methods.

Zac Harland

Thesis Submitted for the Degree of Ph.D. in Finance

Cass Business School
Faculty of Finance

July 13, 2010

Contents

List of Figures

List of Tables

Abstract

In this thesis we investigate the question of asset price predictability. The two major themes that we focus on are firstly; whether machine learning and statistical modelling techniques, which impose less restrictive assumptions on asset price dynamics than do classical linear methods, can be used to forecast and trade financial markets to a degree greater than that which traditional asset pricing models would lead us to expect and secondly; to what extent model combination/ensemble strategies can add value in this pursuit. The approaches used include support vector regression (SVR), k-nearest neighbours (KNN), trading rules, linear regression (LR) and the random subspace ensemble method.

We investigate these two themes using inherently data-driven models across datasets of sufficient size to render statistically meaningful results in three self-contained contexts. The first piece of empirical work compares the relative forecasting performance of SVR, KNN and LR models when applied to predicting daily returns of 58 UK stocks in the FTSE 100 over 4000 days.

Bootstrap simulations are used to shed further statistical light on model performance.

Secondly, we investigate the extent to which model combinations can improve forecasting performance with the use of the random subspace ensemble method for constructing ensembles of linear regression models to predict the returns of a portfolio of FTSE 100 stocks. The primary ensemble consists of 62500 component models estimated by randomly sampling subsets of the feature set and the final result combined via a majority vote

Lastly, we conduct an in-depth study of the channel break-out trading rule over a portfolio of 37 futures markets. We borrow a page from the book of modern portfolio theory where it is the performance of individual markets in the context of a portfolio that is ultimately of interest rather than on an individual basis. This approach is rarely used in the literature but is able to shed more light on the question of trading rule efficacy. Bootstrap resampling is employed to derive robust performance statistics. Our results show the Sharpe Ratio of the portfolio to be three times greater than of individual markets as a result of diversification in addition to being greater than that of S&P500 benchmark.

We did not set out in an attempt to refute the weak form of Fama's (1970) classic taxonomy of information sets or, colloquially, "to beat the market"; nonetheless, some of our results suggest economically significant returns.

Acknowledgments

Thanks to my supervisor, Roy Batchelor, for his guidance, insights, criticisms, thoughtful suggestions and continual prodding during the course of this study. I was always given enough rope but never enough to actually hang myself.

I would like to express my sincere thanks to a number of the faculty at the University, especially Anna Maria Fuertes for her excellent teaching, helpful advice and for making one feel more than just another student. Additionally, thanks to Alec Crystal for providing support when it was needed.

A big thank-you goes to my family, especially my wonderful parents for always providing unconditional support, they deserve far more credit than I can ever give them.

Finally, this thesis is dedicated to Maria to whom I owe my most heartfelt gratitude for her love, support, encouragement and companionship - only you know what it took to make it this far.

Declaration

I grant powers of discretion to the University Librarian to allow this dissertation to be copied in whole or in parts without further reference to me. This permission covers only single copies made for study purposes, subject to normal conditions of acknowledgement.

Chapter 1

Introduction

1.1 Motivation, Scope & Background

In this thesis we investigate the question of asset price predictability. The two major themes that we focus on are

1) Whether machine learning and statistical modelling techniques, which impose less restrictive assumptions on asset price dynamics than do classical linear methods, can be used to forecast and trade financial markets to a degree greater than that which traditional asset pricing models would lead us to expect.

2) To what extent model combination strategies can add value in this pursuit.

1.1.1 The EMH & Behavioural Paradigms

Financial markets are notoriously difficult to predict [1](Timmermann 2008). This is not surprising given that incremental success in this endeavour renders the task progressively more challenging as detectable structure is systematically ironed out through exploitative trading and markets become more efficient. The well known Efficient Markets Hypothesis (EMH) (Fama 1970) embodies this process of ever increasing market efficiency at its limit, a point where it is deemed impossible to realize above average risk-adjusted returns. Given that the EMH implies that market prices are for the most part unpredictable it is imperative for any investigation into the predictability of asset prices to address those theoretical and practical limitations it imposes. However, whilst reference to the EMH is necessary it is important to note that the paradigm does not preclude predictability per se. In fact we would expect some deterministic structure due to such factors as transaction cost frictions and systematically time varying risk premia.

Prima facie, attempting to predict asset returns is by most accounts liable to be a futile exercise, especially if prices move randomly in accordance to an independent and identically distributed process (*i.i.d.*). Weak form market efficiency would suggest that prices traded in a market that is weak form efficient are not predictable using historical price information. This would imply that prices traded in such a market are serially uncorrelated. One

[1] At least in terms of the first moment.

method that has been adopted in the extant literature for testing weak form market efficiency has been an examination of asset prices for evidence of non-random behaviour. The Random walk hypothesis (RWH) posits that successive price changes in an efficient market are random. However, the RWH is somewhat restrictive in practice, as it implies that in the process:

$$P_t = P_{t-1} + \varepsilon_t \tag{1.1}$$

the error term ε is *i.i.d.*. If P_t represents the log of a market price, this implies that the market returns are *i.i.d.* It is possible to relax the assumption of *i.i.d.* returns in the context of weak form efficiency by replacing 1.1 with:

$$E[P_{t+1} \mid I_t] = P_t \tag{1.2}$$

where I_t is any information set which includes $P_{t-j}, j \geq 0$. 1.2 is a martingale process. It was Samuelson (1965) and Mandelbrot (1966) who formally recognised the importance of a martingale when describing efficient markets, though the seeds were sown by Bachelier (1900). For a more detailed exposition of the EMH see appendix A

In spite of their intellectual power, in recent decades the EMH and standard asset pricing models, underpinned by assumptions of rational expectations and homogeneity, are starting to give way to alternatives that incorporate a different view of the world, as the former fall short in explaining the inherently complex and volatile nature of financial markets. The classical models assume the existence of *rational economic agents* who always act in their own self-interest, conducting optimal cost/benefit analysis when arriving at decisions using the statistically correct probabilities (perhaps more aptly termed *hyper-rational agents*). These rational agents are assumed to have homogenous expectations. Many authors have questioned the foundation of these assumptions, most notably Simon (1955) who argued that individuals are unlikely to able to carry out the kind of optimization assumed in the classical models and rather engage in *satisficing*.

This recent shift (*a Kuhnian paradigmatic shift* in some authors' views) away from the rational expectations, efficient markets model towards a so-called behavioural/Heterogeneous Agents framework, which arguably accounts more satisfactorily for the stylized facts of financial markets, places a different emphasis on the issue of asset price predictability. If heterogeneity does not simply average out via *representative agents* who exhibit an average agent's behaviour (Kirman 1992) and individuals are boundedly rational satisficers (Herbert 1957) then markets, rather than fully reflecting all that is known and reacting to fundamental 'news', may in reality be in a constantly evolving state, at times displaying nonlinearity and/or chaos (as in the results of the two-agent models of De Grauwe & Grimaldi (2005)).

In a similar vein, Lo's Adaptive Market Hypothesis (AMH) (Lo 2004) recasts the EMH view of the world incorporating evolutionary models of human behavior positing that individuals make their decisions based on trial and error and that processes of learning, adaptation, competition and natural selection play a significant role in the formation of prices. Investors use trial and error to establish heuristics in the markets. Their skills improve as they climb a learning curve, and then, inevitably, the market evolves rendering some strategies obsolete. The survivors innovate and adapt, establishing new methods of making money.

Many systematic biases in human decision making processes have been documented by behavioral economists and psychologists and include: overconfidence (Fischoff & Slovic 1980), (Barber & Odean 2000), (Gervais & Odean 2001), overreaction (DeBondt & Thaler 1986), loss aversion (Kahneman & Tversky 1979), (Shefrin & Statman 1985), (Odean 1998), herding (Huberman & Regev 2001), psychological accounting (Tversky & Kahneman 1981), miscalibration of probabilities (Lichtenstein, Fischoff & Phillips 1982), hyperbolic discounting (Laibson 1997), and regret (Bell 1982*a*),(Bell 1982*b*), (Clarke, Krase & Statman 1994).

1.1.2 Nonlinearities

If prices evolve via complex interactions between heterogenous groups of investors in a manner similar to that advocated by the behaviouralists, clearly nonlinear modelling approaches are required and may result in greater accuracy than traditionally motivated classical linear approaches [2]. As Campbell, Lo & MacKinlay (1997) argue

> " *many aspects of economic behavior may not be linear. Experimental evidence and casual introspection suggest that investors attitudes towards risk and expected return are nonlinear. The terms of many financial contracts such as options and other derivative securities are nonlinear. And the strategic interactions among market participants, the process by which information is incorporated into security prices, and the dynamics of economy-wide fluctuations are all inherently nonlinear. Therefore, a natural frontier for financial econometrics is the modeling of nonlinear phenomena.* "

Questions regarding *why* we might expect nonlinearities in financial markets aside there is an extensive literature pertaining to the testing for nonlinearity and chaos in financial data. Antoniou & Vorlow (2005) investigated the compass rose patterns revealed in phase portraits (delay plots) of FTSE 100 stock returns and found a strong nonlinear and possibly deterministic signature in the data generating processes. Yang & Brorsen (1993) found evidence of nonlinearity in several futures markets, which was consistent with deterministic chaos in about half of the cases. Sewell, Stansell, Lee &

[2] Heterogeneous agent models do not necessarily imply markets are anymore predictable in terms of gaining excess returns beyond that expected under a rational expectations framework but rather, there may be more to predict.

Below (1996) examined weekly changes for the period 1980 to 1994 in six major stock indices (the US, Korea, Taiwan, Japan, Singapore and Hong Kong) and the World Index as well as the corresponding foreign exchange rates between the US and the other five countries. They concluded that their results did not conclusively prove the existence of chaos in these markets but that they were consistent with its existence in some cases. Examining 30 years of the Canadian/US dollar exchange rate data Kyrtsoua & Serletisb (2006) conclude, "There is a strong evidence that noisy chaotic structures are responsible for nonlinearity in the mean of the Canadian/US exchange rate series". In chapter 4, page 108 using the BDS test (Brock, Dechert, Scheinkman & LeBaron 1996) we find nonlinearity beyond GARCH effects cannot be rejected for 49 of 65 FTSE 100 index stocks.

The theoretical motivation provided by the behaviouralist agent based paradigm, in addition to the empirical motivation provided by literature on nonlinearity in markets, both underpin the first major theme of this thesis; the empirical investigation into the efficacy of linear to nonlinear modelling methods applied to forecasting financial markets.

1.1.3 Model ensembles and combination

The second major theme of this thesis is model combination. Following seminal papers on model combination in business and economic forecasting by

Barnard (1963), Bates & Granger (1969) and Roberts (1965), a considerable literature has developed over four decades showing accuracy improvement under certain conditions using combinations of forecasts - for an overview see Clemen (1989). Recently this has been consolidated by Granger & Jeon (2004) where it is termed "thick modeling" in contrast to "thin modeling", the use of just a single model. The benefits have not been lost on researchers in other fields such as machine learning, where classifier or model ensembles as they are known, contribute a large part of the literature (Hansen & Salamon 1990)(Breiman 1996). The general consensus is that in order for an ensemble to improve overall accuracy the component models should be both accurate and diverse (Batchelor & Dua 1995), and at a minimum they need only be slightly more accurate than random guessing as long as they exhibit sufficient diversity - see Dietterich (2002) for an additional overview in the ML literature. For a more detailed exposition of model ensembles and machine learning see Chapter 2.

Forecasters of financial markets essentially face a moving target as the data generating process for financial returns changes over time. Timmermann (2008) frames the difficulty faced by forecasters in practical terms, providing an intuitive explanation for the benefits of forecast model combination:

> "*We would expect competition between a multitude of forecasting methods to cause instability both in the parameter estimates associated with particular forecasting models and in their (relative)*

forecasting performances. Indeed, the performance of individual forecasting methods may follow a life cycle pattern. Before a particular forecasting approach is widely discovered and adopted, it may perform quite well. Then, given a suitably long historical track record indicating good performance, the forecasting method will become more widely adopted. Finally, as this learning and adoption process becomes more complete and the information in the forecasts gets incorporated into prices, the method will cease to predict future return movements."

A model combination approach may ameliorate this problem of a single model life cycle to some degree. We cover the subject of model combination in detail in chapter 2

1.1.4 The modelling approaches

The No Free Lunch Theorem tells us that no one modelling method dominates all others in all applications (Wolpert & Macready 1995). It may well be that certain methods dominate within a particular problem space where their strengths are aligned with characteristics of the underlying data that is, it is more aligned with bayesian priors than other algorithms. The modelling approaches used in this study of asset price predictability and model combination are chosen specifically for their flexibility and problem-fit and include support vector machines (SVM) - a somewhat recent addition to the set of nonlinear techniques available for classification and regression, k-nearest neighbours (KNN), the channel break-out trading rule (which can be considered a form of nonlinear model), linear regression (LR) and random subspace ensemble method.

Support Vector Machines

The Support Vector Machine (SVM) is a powerful machine learning method for classification and regression (see section 2.6 for further details) and is fast replacing neural networks (though it is actually a super set of neural networks) as the tool of choice for nonlinear prediction and pattern recognition tasks, primarily due to their ability to generalise well on new data and their solid theoretical foundation. SVMs are a member of a large class of learning algorithms known as *kernel methods*. Generally speaking kernel methods exploit information regarding the inner product between data instances and it is possible to re-write many algorithms such that they only need to consider these inner products. To do this kernels are introduced which are essentially distance measures in some feature space.

SVM Regression involves a nonlinear mapping of an n-dimensional input space into a high dimensional feature space. A linear regression is then performed in this feature space. In the field of finance SVMs have been used in interest rate curve estimation (Monteiro 2001), credit rating prediction (Huang, Chen, Hsu, Chen & Wu 2004), financial time-series prediction (Yang, Chan & King 2002) and modelling stock indices (Abraham, Philip & Saratchandran 2003).

K-Nearest Neighbours

The K-Nearest Neighbours method is form of Instance-based learning (IBL) (Aha, Kibler & Albert 1991) . These methods are often used in classification tasks and use specific instances (vectors of training examples or independent variables) to perform the classification, rather than using generalizations of the training data. IBL algorithms are also known as lazy learning algorithms, as they save some or all of the training examples and delay all effort towards inductive classification until requests for classifying yet unseen instances are received. They assume that similar instances have similar classifications: novel instances are classified according to the classifications of their most similar neighbours. In this case the KNN method performs so-called local learning resulting in a globally nonlinear function. See Chapter 2.8 for further details.

Model combination with the Random Subspace Method

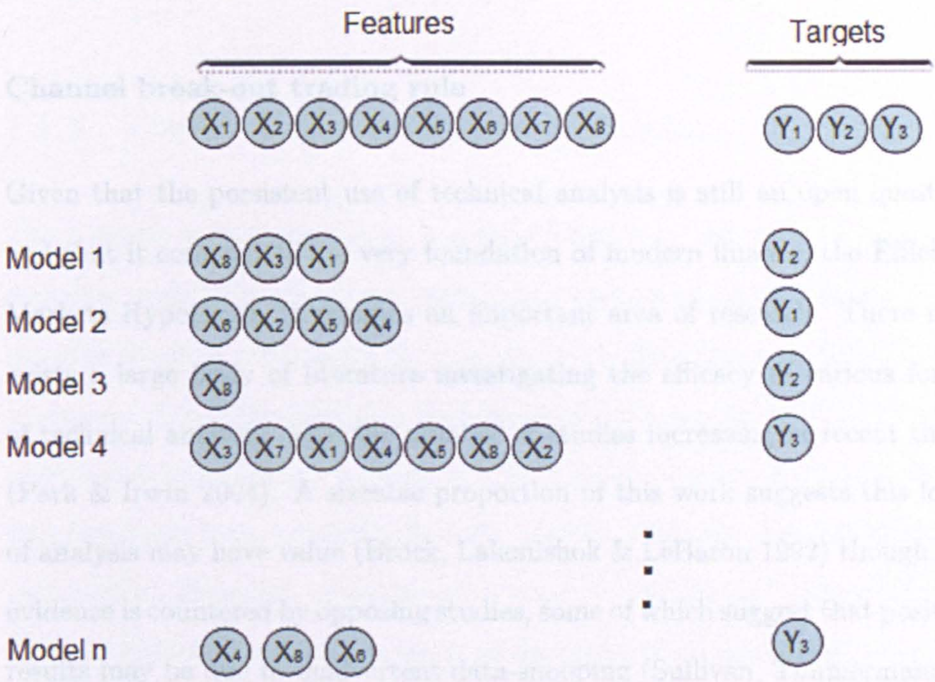
The random subspace method (RSM) of Ho (1998), rooted in the theory of stochastic discrimination (Kleinberg 2000), involves creating an ensemble of models whereby ensemble components use only a subset of the available training features/inputs.

Like bagging (Breiman 1996) and boosting (Freund & Schapire 1996); (Kearns 1988) the Random Subspace Method also works by modifying the training data but it does so in feature space. Given a training set of N feature set variables, a learning machine is repeatedly applied to randomly selected feature subsets of size $M < N$. Model outputs are then aggregated via majority voting to determine the final ensemble output. Although Boosting has proven to be a very effective combining algorithm it has a tendency to overfit the data in the presence of noise (Opitz & Maclin 1999) which makes it less appropriate in the context of financial market data where the signal to noise ratio is low.

The RSM is derived from the method of Stochastic Discrimination(SD) introduced by Kleinberg (1990),Kleinberg (2000) and shares some of its theoretical roots. SD is a method whereby weak component models are created and combined to produce accurate models. Kleinberg argues that it is not just another method of combining classifiers in the sense of attempting to create somewhat orthogonal learners for combination but rests on strict mathematical notions of enrichment, uniformity, and projectability. Its effectiveness is shown in Kleinberg (2000) where it outperforms bagging and boosting on a large number of benchmark problems.

Ho (2000) found that RSM exhibited better performance when the discriminatory power of the input space is distributed evenly over many features. Skurichina & Duin (2002) confirmed this result and further noted that RSM was more effective in this case than when the discriminatory power is condensed to a few features.

Figure 1.1: Random Subspace Sample Method



RSM ensembles have been shown to outperform single models in a variety of applications (Munro, Ler & Patrick n.d.),(Skurichina & Duin 2002), (Chawla & Bowyer 2005), (Zhao, Tang, Lin, Samson & Remsen 2005), (Bertoni, Folgieri & Valentini 2005). Rooney, Patterson, Tsymbal & Anand (2004) compared RSM, bagging and boosting using regression and nearest neighbours models on 15 data sets selected from the WEKA (Witten & Frank 1999) repository and found RSM to be most effective.

Figure 1.1 gives a visual explanation of the method as it is used in this thesis. In our case we not only sample from the input feature set but also from the output set (dependent variable set).

Channel break-out trading rule

Given that the persistent use of technical analysis is still an open question and that it contradicts the very foundation of modern finance, the Efficient Markets Hypothesis, it remains an important area of research. There now exists a large body of literature investigating the efficacy of various forms of technical analysis, with the number of studies increasing in recent times (Park & Irwin 2004). A sizeable proportion of this work suggests this form of analysis may have value (Brock, Lakonishok & LeBaron 1992) though the evidence is countered by opposing studies, some of which suggest that positive results may be due to inadvertent data-snooping (Sullivan, Timmermann & White 1999) thus, despite these efforts, the academic jury is still out.

Many formulations of technical trading rules can be considered nonlinear models. The channel breakout is one of these and belongs to class of trading rule considered by practitioners to be "trend following" (Pring 2002) ,(Schwager 1996) in that it is assumed that market prices are persistent to some degree ("markets trend") and by taking a position early enough and in the direction of a trend the trader can gain positive returns. There exist innumerable variations, all following a common theme that involves trading when the price "breaks out" from a predefined channel which can be based on price or volatility - in which case it is a volatility breakout system (Pring 2002).

1.1.5 Conceptual framework and data.

The support vector regression, k-nearest neighbour and random subspace sampling modelling approaches were mainly researched and developed in the machine learning community and are inherently data-driven, atheoretic and nonparametric. The underlying paradigms of econometrics tend to be parametric and hypothesised in nature, with a greater emphasis on theoretical soundness and interpretability than on inference and experimental effectiveness per se. Therefore the empirical studies are carried out within the conceptual framework of both the machine learning and econometrics disciplines as they largely overlap each other, each one contributing and complementing the work respectively. That is not to say that the research will not have theoretical soundness at its heart and interpretability as an objective but rather, there will be a greater emphasis on letting the data "speak for itself", largely independent of expectations and presuppositions required to form prior hypotheses. Given the designated context the nomenclature of machine learning will be used where appropriate.

Given that asset return predictability is still an open question and the low signal to noise ratio inherent to financial markets, this research leans towards the incorporation of comparatively large data sets combined with sparsely parameterized models, in addition to minimally pre-processed independent and dependent variables. It is generally the case that as the complexity of a model increases its capacity for being understood diminishes, so the fewer parameters and assumptions incorporated into the modelling process the closer we can get towards answering the underlying question of predictability or, informally, *less is more.*

The emphasis on large datasets is adhered to throughout the thesis to also mitigate the fact that this category of models is inherently data driven (data-rich, atheoretic). The research is carried out on data from both equity and futures markets. The FTSE 100 index (Financial Times Stock Exchange Index 100) is a share index of the 100 most highly capitalised UK companies listed on the London Stock Exchange. We use up to sixty five stocks from the FTSE in addition to thirty seven futures contracts over a number of years, in all comprising almost 500,000 data points. Last, but by no means least, we emphasize effective out-of-sample testing throughout.

1.1.6 The wider context.

In a wider context, underlying this research are two main imperatives. Predictability of asset returns is clearly of interest to investment practitioners, especially those involved in market timing and active asset allocation. From a broader economic perspective it has important implications for the efficient allocation of capital; ideally asset prices provide accurate signals for investment-consumption decisions. If these signals are inaccurate due to information not being fully reflected in prices then the resulting decisions may be inefficient.

The results of this study, apart from contributing to the literature, should also prove useful to traders and investors, whether they be retail or institutional, by providing further information on potentially profitable trading opportunities across a number of different asset classes. Additionally, our results provide a basis for comparison, against which other similar empirical studies may be evaluated.

1.2 Summary of Chapters

1.2.1 Chapter 2

Chapter 2 provides an in-depth exposition of machine learning & ensemble/model combination methods. We compare and contrast machine learning and statistics and present the two main categories of machine learning algorithms, regression and classification. This is followed by an explanation of the support vector regression method that we use in chapter 3 and how it performs nonlinear regression. Instance-based learning and the nearest neighbour method of forecasting is then covered. Finally we discuss model selection, the bias/variance trade-off, data-snooping and boostrap randomisation tests.

1.2.2 Chapter 3

In chapter 3 we compare the performance of three different forecasting methodologies in predicting individual daily returns of 58 UK stocks in the FTSE 100. The methods used are support vector regression, k-nearest neighbours and linear regression The study is conducted across a dataset of sufficient size to render statistically meaningful results and provides insight into the question of nonlinear vs linear predictability. Concurrently, we wish to ascertain to what extent UK stock returns are predictable, if at all.

The data for each stock consist 4100 days from 1986 - circa 240,000 data points in total. Applying each methodology in such a way as to produce meaningful comparative out-of-sample results by incorporating in-sample, validation and out-of-sample data sets we find that the nonlinear support vector regression models outperform linear and KNN models by a factor of 2 in terms of Sharpe Ratio. There is no significant difference between the linear and KNN results. Bootstrap tests are used to test the significance of the results.

1.2.3 Chapter 4

Chapter 4 investigates the application of linear ensemble models to forecast and trade 65 component stocks within the FTSE 100, using daily data over the years 1991-2006. Specifically, we are interested in what performance benefits if any are accrued by using ensemble models - models that combine the forecasts of a number of individual component models - over single model specifications. We do this by comparing the performance of a single linear AR model and four ensembles with differing specifications.

The primary ensemble consists of 62500 component models built using the random subspace method in which randomly sampled subsets of the feature set are used to estimate each model with the final result combined via a majority vote. The performance is compared to a number of benchmarks and it is found that this ensemble methodology improves the overall results both in terms of consistency across time periods and economic significance. It is also found that model selection or thinning improves performance further.

1.2.4 Chapter 5

In Chapter 5 we examine the channel break-out technical trading rule over a portfolio of thirty seven futures markets from 1982 to 2005. We look not only at a large portfolio of markets, which few studies attempt, but also in a manner which closely resembles how practitioners implement these rules by including a trade management strategy (TMS) using stop-loss and profit-limit orders and accounting for transaction costs. Bootstrap tests are used to test the significance of the results and it is found that both the vanilla trading rule and the rule with the TMS added consistently realise significant net returns over the whole data sample.

Chapter 2

Machine Learning & Ensemble Methods.

The term Machine Learning (ML) is said to have been coined by Samuel (1959) in a study of checkers, where, to paraphrase, he described ML as, the field of study that gives computers the ability to learn without being explicitly programmed. Mitchell (1997) defines ML thus, "a computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience." The common theme throughout ML is the adaptation from experience with a concomitant improvement in generalisation performance on similar tasks.

2.1 Statistics and machine learning

Whilst is it often said that what is termed "machine learning" is the practice of statistics by computer scientists, it is sometimes useful to differentiate between the two. Typically the difference between inferential statistics and machine learning is the manner in which they deal with prior knowledge in that with the former, a hypothesis is posited which is then tested using various statistical machinery to either confirm or deny it - so in this sense we can say it is "confirmatory". Moreover, often a specific prior assumption regarding the distribution of the data is required as the efficacy of many statistical tools rely on certain assumptions of "well-behaved" distributions. ML can be considered less confirmatory and more an exercise of data exploration, in that often little or no assumptions are made regarding the data, the aim being to derive this information from the data itself. In this sense statistics can be seen to be having more self-imposed limits and in certain situations can be less flexible. The trade-off is that often statistical methods deliver benefits in terms of greater theoretical soundness and interpretability.

The reality is that both fields largely overlap each other; statistical methods are frequently explicitly incorporated within machine learning algorithms, such as regularisation in support vector machines and neural networks. Statistical model evaluation techniques such as bootstrapping (Enfron & Tibshirani, 1993) and cross-validation are standard techniques used in ML. ML frequently offers a more pragmatic perspective in the face of ever increasing data, often tackling the computational issues when scaling up algorithms to large data sets, resulting in more practical algorithms for real world applications and improved experimental effectiveness.

2.2 Classes of machine learning algorithms

In the field of ML the main classes of learning algorithms consist of supervised, unsupervised, reinforcement and, more recently, multiple-instance learning (Dietterich et. al. 1997)(sometimes considered a subset of supervised learning). This thesis deals predominantly with supervised learning in which the overall task involves the selection of a function, or hypothesis, from a given hypothesis space that approximates a desired response contained within a set of labelled training examples. Supervised learning algorithms can be subdivided into two main categories, regression and classification.

2.2.1 Regression

In this case the labels take on real values. Given a set of training data:

$$S = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_l, y_l)\} \tag{2.1}$$

where $\boldsymbol{x}_i \in \boldsymbol{X} \subset \Re^N$ and the labels $y_i \in Y \subset \Re$

our objective is to find a function, or hypothesis, which minimizes some risk functional:

$$R(f_s) = \int c(f_s(\boldsymbol{x}), y) d, P(\boldsymbol{x}, y) \tag{2.2}$$

where c is the cost of making prediction $f_s(\boldsymbol{x})$ when the actual value is y.

The prediction function f_s is learned from the training data 2.1 using some learning algorithm.

In practice the true distribution $P(\boldsymbol{x}, y)$ is unknown so we end up minimizing the empirical risk:

$$R_l(f_s) = \frac{1}{l} \sum_{i=1}^{l} c(f_s(\boldsymbol{x}), y) \tag{2.3}$$

(for example the mean square error.) The problem is that minimizing 2.3 is not necessarily equivalent to minimizing 2.2, and pursuing the former can lead to the problem of over-fitting It is important to take some measures to avoid this, the form of which will depend on the algorithm in question. This problem of over-fitting is encapsulated by the so-called bias-variance dilemma that is covered in section 2.11.

2.2.2 Classification

Where the variable of interest is categorical and each output is assigned a label which represents the class information of the example. Classification involves the process of learning to separate objects into different classes based on a set of features. Examples include face, voice or handwriting recognition, medical diagnosis, forecasting the direction or state of market prices etc.

During training the class labels are used as the desired outputs of the classifier which is trained to minimize some measure of error.

More formally, we wish to classify objects:

$$\boldsymbol{x}_i \in \boldsymbol{X} \subset \Re_N \tag{2.4}$$

As belonging to one of two classes: $y_i \in \{-1, +1\}$ On the basis of some labelled training data:

$$S = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_l, y_l)\} \tag{2.5}$$

(2.5 is the same as 2.1 but is repeated here for clarity.)

That is, we wish to learn a mapping:

$f_s : \boldsymbol{X} \rightarrow \boldsymbol{Y}$

$f_s : \boldsymbol{x} \mapsto y$

that generalizes to unseen examples.

In the context of financial time-series forecasting regression is generally used although the resulting forecasts can be converted into simple directional forecasts that is, whether a series will rise or fall in the next time period.

2.3 Model combination and ensembles in machine learning

Model combination is analogous to diversification in investment management. Portfolios of assets are created in such a way as to diversify away as much risk as possible, usually within a mean/variance framework. In this sense creating ensembles is partly an attempt to diversify away model uncertainty or risk.

Note in the case of models attempting to forecast financial markets, model accuracy, in terms of the percentage of times the model correctly forecasts the sign of the return, is only part of the story. It is also the magnitude of the return that must be accounted for. It is quite conceivable for a model to exhibit less than 50% directional accuracy but to still make money if the return per trade/forecast is enough to offset this. This can be measured by the following:

$$\frac{R^W}{R^L}\frac{A^W}{A^L}$$

where R^W is the average return of winning trades, R^L the average absolute return of losing trades, A^W , the percentage of winning trades and A^L , the percentage of losing trades. A value greater one indicates a positive expectation for the model.

2.4 Ensembles

In the context of supervised learning an *ensemble* is a collection of individual "learning machines" (trained regression or classification models - be they support vector machines, linear regression models, decision trees etc.) whose predictions are combined - typically via a weighted voting scheme - to create a final prediction model with the ultimate objective of improving overall accuracy.

There is a rich and diverse literature of research conducted across multiple disciplines concerning optimal combination of regression and classification models. Recently much attention has been paid to this area in the machine learning community due to the increasing availability of cheap and fast computational resources. This research has been both theoretical; (Hanson & Salomon 1990), (Krogh & Vedelsby 1995), (Kittler, Hatef, Duin & Matas 1998) and empirical;(Opitz & Maclin 1999), (Dietterich 2000*b*). Predominantly the research addresses classification though much of it generalises to regression models.

There exist a multitude of different methods for creating the individual models for the ensemble and of amalgamating the resulting prediction outputs. Note the term *ensemble* is used in this thesis though other terms, such as committee machines (Tresp 2001) , aggregation (Brors, Kohlmann, Schoch, Schnittger, Haferlach & Eils 2000), multiple classifier systems (Ho 1992), classifier fusion (Varshney 1996) can be found in the literature, reflective of an emerging discipline.

The properties of a good ensemble are that the individual models are accurate and they make errors on different parts of the feature space. Although much of the empirical work utilises datasets from the UCI WEKA machine learning repository (Asuncion & Newman 2007) in order to mitigate the problem of "one algorithm, one dataset", it has not yet been established theoretically exactly which method of ensemble creation should be used. Empirical results so far show this to be context dependent. Moreover, this is complicated by the fact that although the UCI machine learning repository goes some way in standardising research in this area by allowing researchers to investigate methods using similar datasets, often only a subset of the available datasets is used, without proper clarification of motivation.

2.4.1 Why ensembles work.

Dietterich (2000a) provides a number of fundamental reasons why it is possible to build ensembles that perform better than a single classifier. The first is statistical. Learning machines search a given space of H hypotheses with the objective of finding the best hypothesis that fits the data. Given that many real world applications suffer from insufficient data is it quite conceivable for the algorithm to find many different potential hypotheses all of which fit the data equally well but which may or may not be close to the true function f. In this case an average of all the individual H's may result in a H that is closer to the true function f than any individual model.

The second reason is computational. Many learning machines perform a local search in the H-space leaving them open to the problem of getting caught in local minima. Even if large amounts of data are available it can still be difficult to be certain that any given H is not simply a local optimum as opposed the a global one. This can be ameliorated by the ensemble averaging process which can provide a closer approximation to the true function. Neural networks are known to be prone to this problem and generally the analyst must perform many training runs using different starting weights in attempting to find a global optimum. Support vectors machines should suffer less from this issue as they involve a quadratic optimisation problem and therefore have a single global optimum.

The third reason is representational. Although many learning algorithms can be shown to be universal function approximators with proven asymptotic representation theorems, this is with the assumption of unlimited data. In the context of limited data the algorithms will perform only a limited search of possible Hypotheses. It may be the case that the true function cannot be represented by any of the hypotheses in H. Again, ensemble averaging can serve to expand the space of representable functions resulting in better approximation.

2.5 Methods for constructing ensembles

Valentini & Masulli (2002) provide a detailed taxonomy of ensemble methods under the headings of "generative" and "non-generative" methods. Non-generative ensemble methods operate on a given set of previously designed models, attempting to find some optimal combination. They are not involved in the actual creation of new models. Generative methods create ensembles by attempting to increase the accuracy and diversity of the base learner [1].

[1] A *base learner* is any classification or regression model

2.5.1 Non-Generative Methods

Probably the most common type of non-generative method consists of simply majority voting a set of component models' outputs. This method can be refined by including a weighting scheme in which those models achieving greater accuracy are given larger weights with the intention of increasing the overall accuracy. Other approaches include fuzzy aggregation methods (Cho & Kim 1995),(Keller & Yang 2000), (Vanccka & snd S Alya 2009), and Dempster-Schafer combination rules if the outputs are possibilistic (Rogova 1994). A hierarchical method called stacked generalisation (Duin & Tax 2000) creates a number of second level models that take as inputs the outputs of the base models and then tries to learn an optimal combination.

2.5.2 Generative Methods.

These methods manipulate and modify the structure and characteristics of the training data to create different hypotheses and are best suited to so-called unstable algorithms where small changes in the training data can produce large differences in the output predictions, an example being neural networks and decision trees, as opposed to nearest neighbour methods which are considered relatively stable.

Bagging

With the Bagging (Bootstrap Aggregating) (Breiman 1996) algorithm, a learning machine is repeatedly applied to bootstrap replicate training sets of size m (Grossman & Williams 2000), where each bootstrap replicate is drawn by randomly sampling with replacement from the original training set (typically $m = n$, the size of the training set). The individual base learners are then trained on these replicates which contain on average 63.2% of the original training data hence it is sometimes called the .632 bootstrap. Their outputs are then aggregated via majority voting to determine the final ensemble output. That is, the final model evaluates out-of-sample data by querying each of the base classifiers on the sample and outputting their majority opinion. The main arguments for its effectiveness are (1) running several trials on uniform samples of a population results in more significant (less variant) statistical results, and (2) deferring to the majority decision can rid the classifier of noise-induced errors that occur only in a handful of the base classifiers. Typically the base classifiers are homogenous with only the parameters differing.

Boosting

The goal of boosting, due to Schapire (1990) and developed by (Freund & Schapire 1996), is to increase the accuracy of a given algorithm on a given set of training instances, via successive creation of composite classifiers based on a filtration of the training set.

Boosting is similar to Bagging in that it also manipulates the training set however it does so in a slightly more intelligent manner. With Adaboost, the most popular variant of boosting, a model is applied to the training data and its accuracy then measured. The probability distribution over the samples is then re-weighted based on a measure of how hard each sample was to classify for the current model. In each iteration a new model is invoked to minimize the weighted error over the training set. The effect of the change in weights is to place more emphasis on samples that were miss-classified and less on those correctly classified. The final ensemble is then constructed with a weighted vote of each base learner, on the basis of its performance on its weighted training set.

The method can suffer problems especially if base model/learning method is not that weak (i.e., the error gets small very quickly), or the data contains a fair amount of noise, which can then lead to over-fitting as the model concentrates on what are likely to be noisy training instances.

Cross-validated committees

Another training set sampling method consists of constructing training sets by leaving out disjoint subsets of the training data as in cross-validated committees (Parmanto, Munro & Doyle 1996). For example, the training set can be randomly divided into n disjoint subsets. The base learner is trained n times, each time leaving out one of the subsets from the training set to use as a validation set. The final model is then creating using an appropriate voting scheme.

Stochastic Discrimination

The Stochastic Discrimination (Kleinberg 1990) method operates by forming weak base learners which are then combined via the Central Limit Theorem to create a strong classifier. The method is designed to be resistant to over-training (Kleinberg 2000) and has been shown to work well in practice (Chen & Cheng 2000). Kleinberg argues that it is not just another method of combining classifiers in the sense of attempting to create somewhat orthogonal learners for combination but rests on strict mathematical notions of enrichment, uniformity, and projectability. Its effectiveness is shown in Kleinberg (2000) where it outperforms bagging and boosting on a large number of benchmark problems. It is perhaps less popular as it requires a number of strict assumptions which may not hold in practice and the relevant papers are mathematically dense, restricting accessibility to practitioners.

Random subspace method (RSM)

With the random subspace method (Ho 1998) ensemble components use only a subset of the available training features/inputs. A subset of features is randomly selected and assigned to each component learner. This way, one obtains a random subspace of the original feature space with models being constructed inside this reduced subspace. The final decision rule is based on a weighted majority voting on the basis of each individual model's accuracy. The RSM is derived from the method of Stochastic Discrimination(SD) introduced by Kleinberg (1990),Kleinberg (2000) and shares some of its theoretical roots.

RSM ensembles have been shown to outperform single models in a variety of applications (Munro et al. n.d.),(Skurichina & Duin 2002), (Chawla & Bowyer 2005), (Zhao et al. 2005), (Bertoni et al. 2005). Rooney et al. (2004) compared RSM, bagging and boosting using regression and nearest neighbours models on 15 data sets selected from the WEKA (Witten & Frank 1999) repository and found RSM to be most effective.

The Input Decimation Approach.

Input decimation (Tumer & Oza 1999) attempts to reduce the correlation among the errors of the base classifiers, decoupling the base classifiers by training them with different subsets of the inputs features. It differs from

RSM as for each class the correlation between each feature and the output of the class is explicitly computed, and the base classifier is trained only on the most correlated subset of features. Other methods for combining different feature sets using genetic algorithms are proposed by Kuncheva & Jain (2000).

Output Coding decomposition methods

Output Coding (OC) methods decompose a multi-class classification problem into a set of two-class sub-problems and then recompose the outputs to form the classification (Mayoraz & Moreira 1997). With this method the various classifiers are not solving the same problem but rather they are each solving a binary classification problem. Each class is encoded as a bit string, or "codeword", with a different two class base learner (dichotomizer) trained to learn each codeword bit. When dichotomies are used to classify new points, a suitable measure of similarity between the codeword computed by the ensemble and the codeword classes is used to predict the class. Different decomposition schemes have been proposed in the literature.

Manipulating the output targets

This involves using different y values for the output(target) data. For example, typically financial models will utilise log returns as the output/dependent variable data but of course there are an infinite amount of possible outputs to choose from. One example is the use of binary targets whereby if the price goes up the output is a one and if the price goes down it is a minus one. This idea can be taken further with the returns being divided into a number of deciles. Moreover, the first difference price transformation(log returns) essentially acts as a specific form of high pass filter. There is nothing stopping the analyst from using a different filter, for example it may be that using daily log returns for a financial forecasting model results in a model that changes forecasts too frequency, producing large transaction costs. It might be possible to use a smoothed target in order to reduce the trading frequency in the hope of producing a more efficacious model.

2.6 Support Vector Machines

2.6.1 Introduction

The most widely used technique in econometric analysis is that of Classical Statistical Inference in the form of *linear regression*. This methodology is based on a number of fundamental assumptions:

1. The data one is working with can be modelled by a set of linear in parameter functions.
2. In most real-life applications, the stochastic component of data is the normal probability distribution law, i.e., the underlying joint probability distribution is Gaussian.
3. Due to the second assumption, the induction paradigm for parameter estimation is the maximum likelihood method that is reduced to minimization of the residual sum of the squared errors (RSS) cost function, known as *least squares*.

When using parametric approaches apriori assumptions are made with respect to the structure of the underlying data generating process. Assuming the correct functional form, a model is then built by estimating the unknown parameters from the data. The model choice decision is frequently based on issues of tractability and interpretability rather than on whether the correct functional form has been chosen from those available. In many real life problems it is often the case that these assumptions are violated resulting in model miss-specification.

Given a vector of inputs $X = (X_1, X_2, \ldots, X_n)$ we wish to predict the value Y given via the model

$$\hat{Y} = \hat{\beta}_0 + \sum_{j=1}^{n} X_j \hat{\beta}_j \tag{2.6}$$

The term $\hat{\beta}_0$ is the *intercept* or *bias.* If the bias term $\hat{\beta}_0$ is included in the vector of coefficients $\hat{\beta}_j$ and a constant variable 1 is included in X_j, in matrix notation we have

$$\hat{Y} = X^T \hat{\beta} \tag{2.7}$$

The most widely used method of fitting this model to data is the aforementioned least squares method. In this case the coefficients β are chosen such the residual sum of the squared errors (RSS) are minimized

$$RSS(\beta) = \sum_{i=1}^{N} (y_i - x_i^T \beta)^2 \tag{2.8}$$

With non-parametric approaches the unknown density and distribution functions are replaced by their non-parametric density estimators. In cases with sufficient data non-parametric estimators often reveal features of the data that are undetectable under parametric techniques.

In chapter 3 of this thesis we focus on comparing the nonlinear non-parametric approach of support vector machines with traditional linear methods in the context of financial forecasting. The main focus is on Support Vector Machine Regression of which a detailed exposition follows.

2.6.2 Support Vector Machines

The Support Vector Machine (SVM) is a powerful machine learning method for classification and regression and is fast replacing neural networks (though it is actually a super set of neural networks) as the tool of choice for nonlinear prediction and pattern recognition tasks, primarily due to their ability to generalise well on new data and their solid theoretical foundation. SVMs are a member of a large class of learning algorithms known as *kernel methods*. Generally speaking kernel methods exploit information regarding the inner product between data instances and it is possible to re-write many algorithms such that they only need to consider these inner products. To do this kernels are introduced which are essentially distance measures in some feature space.

SVM Regression involves a nonlinear mapping of an n-dimensional input space into a high dimensional feature space. A linear regression is then performed in this feature space. SVMs use the structural risk minimization

(SRM) (Vapnik 2000) induction principle which differentiates the method from many other conventional learning algorithms based on empirical risk minimization (ERM) alone, for example standard neural networks. This is equivalent to minimizing an upper bound in probability on the test set error as opposed to minimizing the training set error, which should result in better generalisation. Importantly for practitioners, recently published research has shown successful application of the SVM methodology in a wide variety of fields (Barabino, Pallavicini, Petrolini, Pontil & Verri 1999), (Joachims 1997), (Mukherjee, Tamayo, Mesirov, Slonim, Verri & Poggio 1999) ,(Ince 2000).

In the field of finance SVMs have been used in interest rate curve estimation (Monteiro 2001), credit rating prediction (Huang et al. 2004), financial time-series prediction (Yang et al. 2002) and modelling stock indices (Abraham et al. 2003).

The method has a number of advantages over other techniques; the parameters that need to be fitted are relatively low in number and, unlike other methods such as neural networks, they do not suffer from local minima. The two main features of SVMs are their theoretical motivation from statistical learning theory and the use of kernel substitution to transform a linear method into a general nonlinear method, with little added complexity.

2.7 How SVMs work.

2.7.1 SVM Classification

The task of classification involves the process of learning to separate objects into different classes based on a set of features. Examples include face, voice or handwriting recognition, medical diagnosis, forecasting the direction of market prices etc.

More formally, we wish to classify objects:

$x_i \in X \subset \Re^N$

As belonging to one of two classes:

$y_i \in \{-1, +1\}$

On the basis of some labelled *training data*:

$S = \{(x_1, y_1), \ldots, (x_l, y_l)\}$

That is, we wish to learn a mapping:

$f_s : X \rightarrow Y$

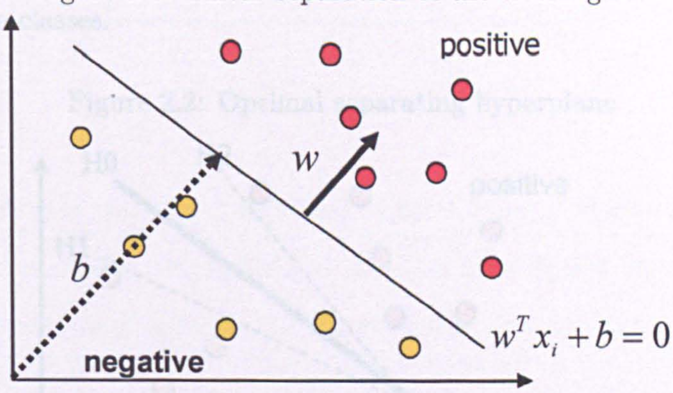
$f_s : x \mapsto y$

that *generalizes* to unseen examples.

Linear separation of the training set.

We wish to find a decision rule that separates or explains the training set to a certain level of accuracy. In the example shown in figure 2.1 the task is to separate the red dots from the yellow dots. In order to do this we want to find a hyperplane $w^T x_i + b = 0$, corresponding to the decision function $f_{w,b}(x) = sgn(w^T x + b)$ that separates red positives $+$ from yellow negatives $-$.

Figure 2.1: Linear separation of the training set

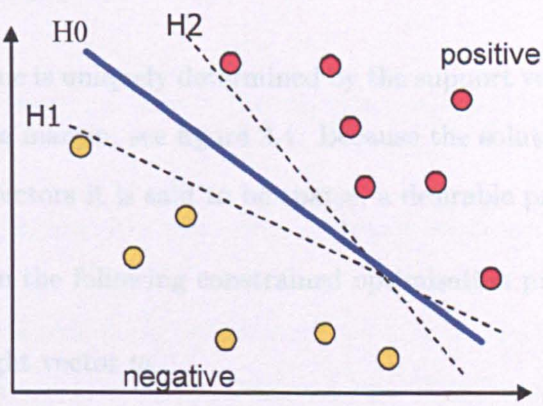


The separating hyperplane is defined by the weight vector w , which is normal to the plane, and the offset b , the distance from the origin. Both w and b are learnt from the training data.

The optimal separating hyperplane.

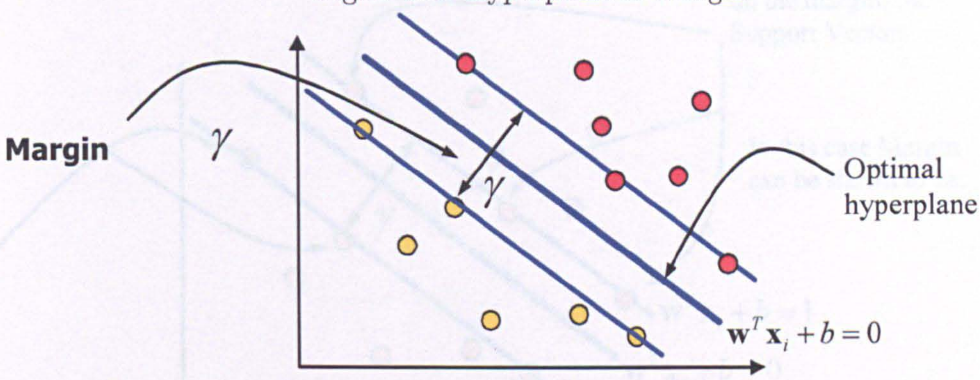
It is easy to find a hyperplane that minimizes the empirical error and there are of course infinitely many hyperplanes that will separate the data but what we're interested in is that hyperplane which results in the best generalisation performance on unseen examples - we are interested in the hyperplane that minimizes the *expected error*. In the figure 2.2 below hyperplanes H0, H1 and H2 all separate the two classes with zero empirical risk or error but which will generalize best? Intuitively H0 is best as it provides the largest separation between the classes.

Figure 2.2: Optimal separating hyperplane



The support vector machine algorithm constructs a hyperplane where the margin of separation is maximised - this is defined as the optimal separating hyperplane and in this example is a decision surface that maximises the margin between red and yellow examples see figure 2.3. More specifically it maximises the margin between the convex hulls of the data (to picture the convex hulls imagine rubber bands placed around the two classes of data).

Figure 2.3: Hyperplane & Margin



The hyperplane is uniquely determined by the support vectors - those vectors that lie on the margin, see figure 2.4. Because the solution only depends on the support vectors it is said to be sparse, a desirable property.

This results in the following constrained optimisation problem:

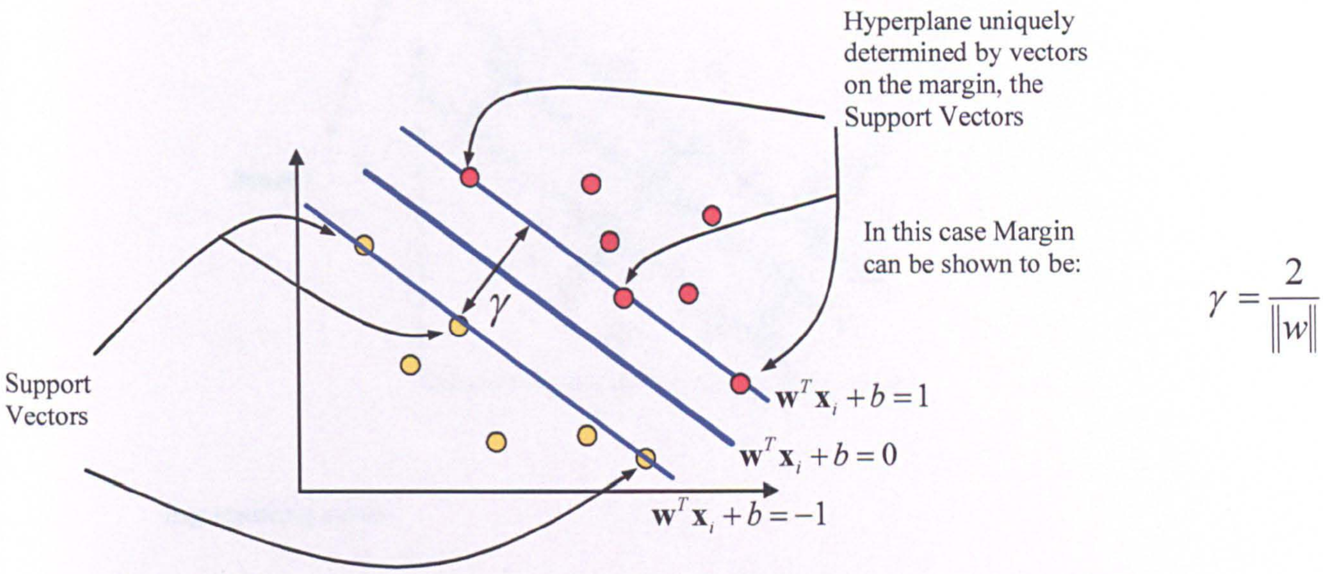
Find the weight vector w

that minimizes the objective function:

$$\phi(w) = \frac{1}{2}\|w\|^2 \tag{2.9}$$

subject to $y_i(w^T x_i + b) \geq 1, i = 1, \ldots, l$.

Figure 2.4: Support Vectors & Margin



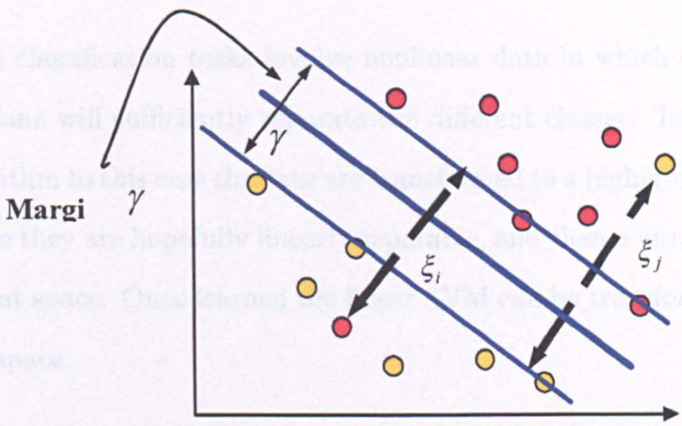
- because the scale is arbitrary the numerical value of the margin is set to 1.

Linearly nonseparable patterns: Introduction of slack variables.

The majority of real world datasets are unlikely to be linearly separable so a hyperplane that exactly separates the data will not exist. The concept of a *soft margin* is introduced by using *slack variables* that allow some errors to exist, see figure 2.5.

Therefore when data are not linearly separable the objective is to maximize the margin with respect to correctly classified examples, subject to minimiz-

Figure 2.5: Introduction of slack variables.



ing training error.

Minimize:

$$\phi\left(w,\xi\right)=\frac{1}{2}\|w\|^{2}+C\sum_{i=1}^{i}\xi_{i} \tag{2.10}$$

subject to $y_i(w^T x_i + b) \geq 1 - \xi_i, i = 1, \ldots, l, \xi_i \geq 0$

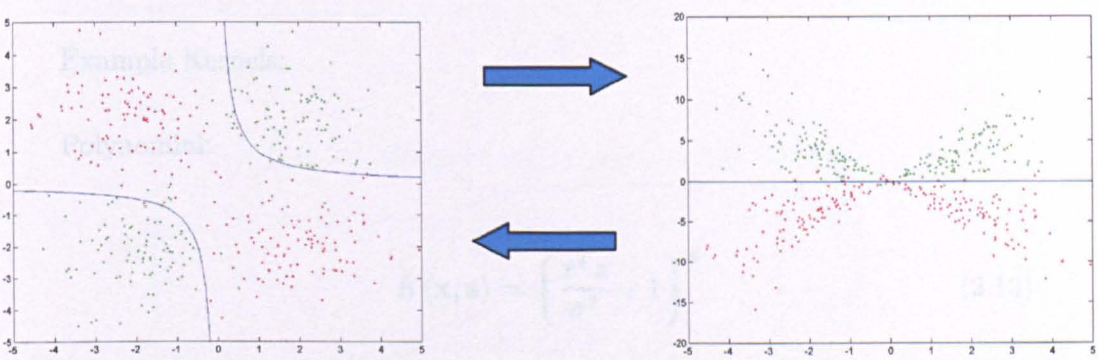
$C > 0$ (chosen by user) determines the trade-off between margin maximization and training error (empirical risk) minimization.

ξ_i are slack variables.

2.7.2 Nonlinear data.

Some classification tasks involve nonlinear data in which case no linear hyperplane will sufficiently separate the different classes. To extend the SVM algorithm to this case the data are transformed to a higher-dimensional space, where they are hopefully linearly separable, and then a linear SVM is learned in that space. Once learned the linear SVM can be transformed back to original space.

Figure 2.6: Transformation into higher dimensional space



The "Kernel trick", nonlinearity via kernel substitution.

The transformation into a higher dimensional space is carried out using kernel functions. Kernel substitution is possible as feature vectors only occur as inner products in the decision function and training algorithm.

Denote the mapping $X \rightarrow H, x \mapsto \phi(x)$

then the decision function becomes:

$$f(x) = sgn(\phi(x)^T w^* + b^*) \tag{2.11}$$

$$= sgn\left(\sum_{i=1}^{l} y_i \alpha_i^* \phi(x)^T \phi(x_i) + b^*\right) \tag{2.12}$$

The mapping is never explicitly carried out; a kernel function computes the inner product in H:

$$K(\mathbf{x}, \mathbf{z}) \equiv \phi(x)^T \phi(z)$$

Example Kernels:

Polynomial:

$$K(\mathbf{x}, \mathbf{z}) = \left(\frac{x^T z}{\sigma^2} + 1\right)^d \tag{2.13}$$

RBF:

$$K(\mathbf{x}, \mathbf{z}) = exp\left(-\frac{\|x - z\|^2}{2\sigma^2}\right) \tag{2.14}$$

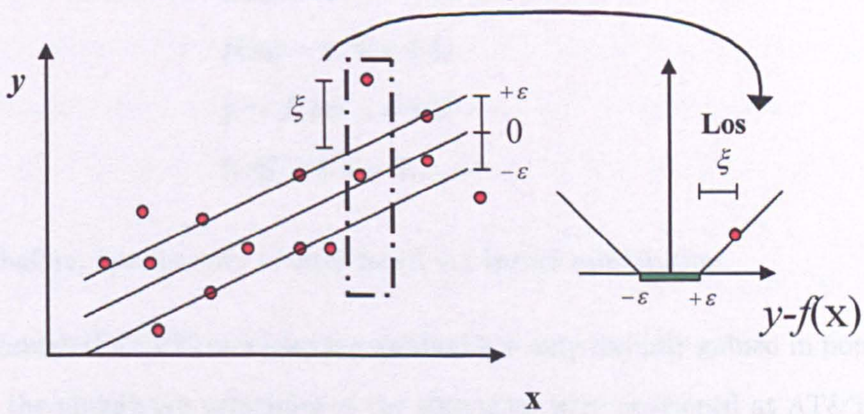
2.7.3 Support Vector Regression

The SVM algorithm is extended to regression by introducing Vapnik's so-called ϵ - insensitive loss:

$$L^{\epsilon}(x, y, f) = max(0, \mid y - f(x) - \epsilon \mid) \tag{2.15}$$

This counts as training errors only those points which fall outside an ϵ -band of the fitted solution. In other words we do not take into consideration errors unless they are greater than ϵ - this allows the concept of margin to be carried over to the regression case see figure 2.7.

Figure 2.7: Support vector regression & the tube



A tube with radius ϵ is fitted to the data. The trade-off between model complexity and points lying outside the tube is controlled by minimizing:

$$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l} \mid y_i - f(x_i) \mid_\epsilon \tag{2.16}$$

(in this case C controls penalty magnitude for points lying outside the ϵ tube)

In this case we need two types of slack variable,

$$f(x_i) - y_i > \epsilon \text{ and } y_i - f(x_i) > \epsilon$$

Minimize:

$$\phi(w, \xi^{(*)}) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}(\xi_i, \xi_i^*) \tag{2.17}$$

subject to:

$f(x_i) - y_i \leq \epsilon + \xi_i$

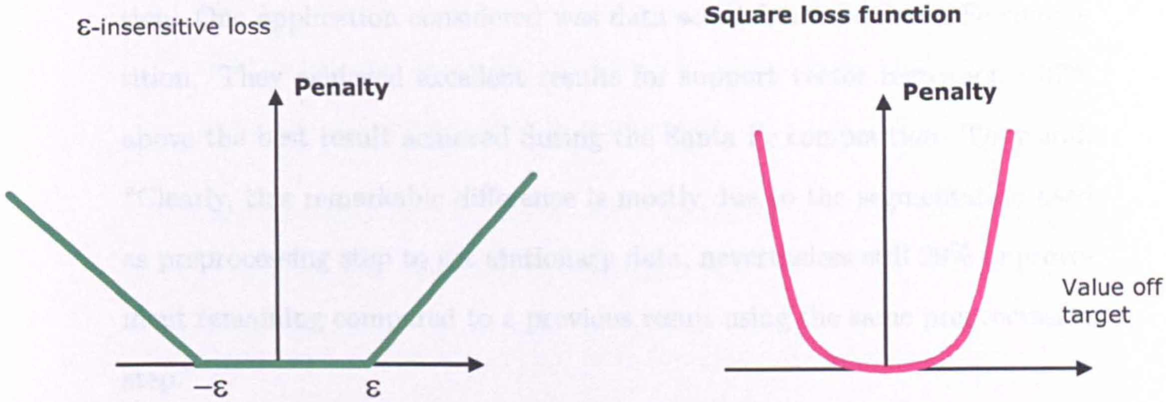
$y_i - f(x_i) \leq \epsilon + \xi_i^*$

$\xi_i, \xi_i^* \geq 0 \; i = 1, \ldots, l$

As before, nonlinearity is introduced via kernel substitution.

Although the SVM as a learning method has only recently gained in popularity, the underlying principles of the algorithm were developed at AT&T Bell Laboratories by Vapnik and co-workers (Boser, Guyon & Vapnik 1992),(Guyon, Boser & Vapnik 1993),(Cortes & Vapnik 1995), (Scholkopf, Burges & V.Vapnik 1995) and (Vapnik, Golowich & Smola 1997) and are based on ideas derived from statistical learning theory. The roots of this approach, the SVM meth-

Figure 2.8: Loss Functions



ods of constructing the optimal separating hyperplane for pattern recognition goes back to 1964 (Vapnik & Chervonenkis 1964). In 1992 the SVM technique was generalized for nonlinear separating surfaces (Boser et al. 1992). In 1993 it was extended for constructing decision rules in non separable cases (Cortes & Vapnik 1995). In 1995 the SVM method for estimating real-valued function was obtained (Vapnik 1995), and lastly, in 1996 the SVM method was adopted for solving linear operator equations (Vapnik et al. 1997). This recent increase in popularity is due to advances in methods and theory which include the extension to regression from the original classification formulation. For a thorough treatment see (Vapnik 1998),(Vapnik 1995), the tutorials (Burges 1998), (Smola & Scholkopf 1998) and the introduction (Cristianini & Shawe-Taylor 2000).

Muller, Smola, Ratsch, Schokopf, Kohlmorgen & Vapnik (1999) used SVMs with radial basis function networks within a time series prediction application. One application considered was data set D from the Santa Fe competition. They achieved excellent results for support vector regression - 37% above the best result achieved during the Santa Fe competition. They add, "Clearly, this remarkable difference is mostly due to the segmentation used as preprocessing step to get stationary data, nevertheless still 29% improvement remaining compared to a previous result using the same preprocessing step."

2.7.4 Summary

In summary, the main advantages of the support vector machine methodology are:

1. It is explicitly based on a theoretical model of learning (Vapnik 1995).
2. There are theoretical guarantees with regards to their performance
3. Only a few parameters need to be chosen. We only need to choose the width of the RBF or the degree of the polynomial kernel, then the error rate C and epsilon.
4. It is not affected by local minima due quadratic optimization with global minimum.

5. It is somewhat more interpretable than neural networks in that the method provides a list of the support vectors, those points which are most important. In addition certain kernel functions such as the polynomial kernel can be directly analysed in feature space.

6. Its formulation means that it is less affected by curse of dimensionality so can be used with high dimensional data.

7. Sparseness of solution in that only the support vectors are required.

The main disadvantage of SVMs in applications is training time compared to methods such Linear Regression.

2.8 Instance-based learning and the Nearest Neighbour method of forecasting.

Instance-based learning (IBL) methods (Aha et al. 1991) are generally referred to in the context of classification where one is trying to assign a class label to an unknown object. The methods use specific instances (vectors of training examples or independent variables) to perform the classification, rather than using generalizations of the training data. IBL algorithms are also known as lazy learning algorithms, as they save some or all of the training examples and delay all effort towards inductive classification until requests for classifying yet unseen instances are received. They assume that similar instances have similar classifications: novel instances are classified according to the classifications of their most similar neighbours.

IBL algorithms ultimately derive from the original nearest neighbour method that is attributed to Fix & Hodges (1952) and Cover & Hart (1968). It is a simple yet powerful non-parametric forecasting and classification algorithm that, as its name suggests, classifies or generates a forecast for a new training instance to the class or forecast of its nearest neighbour in some measurement space using, most commonly, euclidean metrics. The approach is based on the idea of locating local variation in the time-series rather than attempting to model the global properties of the data.

Nearest Neighbour Rule:

1) Given training data or sample set $S = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_l, y_l)\}$

2) Define the nearest neighbour

$$N(x, S) \equiv \arg\ min_{\boldsymbol{x}_i \in S} \|\boldsymbol{x} - \boldsymbol{x}_i\| \tag{2.18}$$

3) Label test point with the label of $N(\boldsymbol{x}, S)$

A slight modification results in the so-called k-nearest neighbour technique which classifies a new instance to the class most heavily represented among its k nearest neighbours, rather than just a single neighbour.

It is also possible to use the algorithm for function approximation (read regression) as we do in chapter 3 of this thesis. Instead of assigning the most frequent class label among the k-nearest neighbours most similar to the pattern to be classified, an average of the function values (dependent variables) of the k training instances is calculated and serves as the prediction for x.

A variant of this approach calculates a weighted average of the k nearest neighbour's function values (Dudani 1975). Given a specific instance x, the weight of a neighbour increases by its respective proximity to x. An optimal value of k can be determined automatically from the training set by using leave-one-out cross-validation (Weiss & Kulikowski 1991).

To use the method one has to choose k, and a distance metric, which is usually the euclidean distance measure. In the context of forecasting financial time-series another parameter is the "lookback" period, L. That is, how far back in the past should the algorithm search for nearest neighbours. It might make sense to limit how far back in time nearest neighbours are searched for as recency may be relevant.

The following describes how the method is used in chapter 3 to forecast financial times-series:

1. Decide on values for k, L and use the euclidean distance measure.
2. Choose an initial starting point in the dataset from which to make the first prediction.
3. Calculate the euclidean distance between the most recent instance to be forecast and all other instances from $T - L$ (or the beginning of the data if the starting point is less than L) to $T - 1$.
4. Find the k nearest neighbours and sum their respective outputs/dependent variables weighted by their distance to the recent instance.
5. If the result from step 4 is greater than zero then the forecast is long else it is short.

2.9 Model Selection, Data Snooping and Randomisation Tests.

Given some data, there will always be an infinite number of viable models or hypotheses [2] that fit the data equally well and without making further assumptions there is no reason to prefer one model or hypothesis over another. Therefore, we are forced to make assumptions that provide us with an *inductive bias*. Inductive bias refers to the assumptions a given learning method has and uses in making its inductions. For example, when using the nearest neighbour method to classify an unseen example based on some measure of distance, one finds the nearest example and assigns its class label to the new point. This method has a bias: it assumes that similar examples (usually within a euclidian space) will have similar outcomes. Mitchell (1980) explains the importance of bias for generalisation.

> *If consistency with the training instances is taken as the sole determiner of appropriate generalizations, then a program can never make the inductive leap necessary to classify instances beyond those it has observed. Only if the program has other sources of information, or biases for choosing one generalization over the other, can it non-arbitrarily classify instances beyond those in the*

[2]The term *hypotheses* is used here in the vernacular of machine learning as opposed to statistics.

training set.

Mitchell (1980)

2.10 Model Selection.

Model selection is the task of choosing a model from a set of potential models with the best inductive bias, which in practice means selecting parameters in an attempt to create a model of optimal complexity given (finite) training data. Ultimately what is of interest in an applied forecasting exercise is the performance accuracy of the models[3] when presented with new unseen/out-of-sample data assumed to be drawn from the same distribution. If the model explains the training data perfectly but does not generalise it will be of little use. In other words, we would like to choose one that minimizes the expected out-of-sample error.

The need to carry out model selection is ubiquitous across many research disciplines and is an inherent part of scientific enquiry in general. Naturally many proposed solutions have arisen within a large body of literature dedicated to the subject. These methods can be thought of as being divided into empirical and theoretical model selection methods. Theoretical methods include Akaike Information Criterion (Akaike 1973), Structural

[3]Here the term *model* is used somewhat loosely in that it can refer to a forecasting model, method or trading rule which hasn't necessarily been estimated on data.

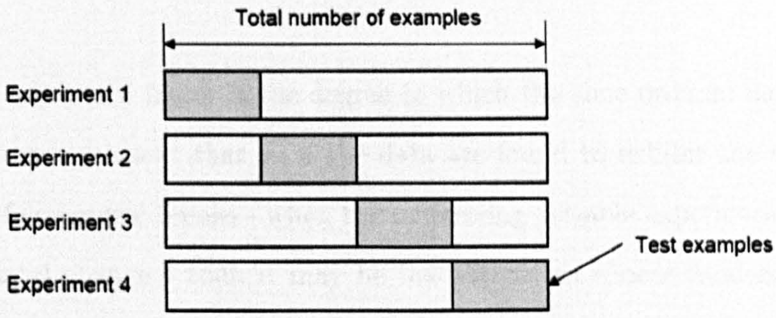
Risk Minimization (Vapnik & Chervonenkis 1974), Bayesian Information Criterion (Schwarz 1978). Empirical methods include adjusted R-squared (Wherry 1931), Bootstrap (Efron 1979) and Cross-validation (Stone 1978). Many of the methods use some form of penalty term in an attempt to balance the trade-off between model complexity and closeness of fit. The choice of method will depend on the context.

In this thesis we primarily use cross-validation, a method that is prevalent in machine learning. The term cross-validation or, more generally, *hold-out testing* entails a wide range of techniques but most commonly the procedure divides the data in two parts: the training set, on which the model is estimated, and the hold-out set, on which its performance is measured. More generally, the *k-fold cross-validation* [4] procedure divides the data into k equally sized folds. It then produces a model by training on $k-1$ folds and testing on the remaining fold. This is repeated for each fold, and the observed errors are averaged to form the k-fold estimate. Figure 2.9 shows an example using k=4. *Leave-one-out* cross-validation is k-fold cross-validation taken to its logical extreme, with k equal to N, the number of data points in the set. It should be noted that $k - fold$ cross validation is generally more effective on small data sets than the simple $1 - fold$ hold out method (Goutte 1997).

The empirical work contained within this thesis uses the following method-

[4] The term "k-fold" comes from the machine learning literature. See for example Bengio & Grandvalet (2004).

Figure 2.9: K-fold Cross Validation (K=4)



ology:

1. The data are divided into three separate sets, an in-sample, validation and out-of-sample set.

2. The models (learning algorithms, trading rules) are then estimated on the in-sample set.

3. Models are then tested on the validation data.

4. The model (or subset of models) exhibiting the optimum performance based on some measure is chosen and finally tested on the out-of-sample data.

The in-sample comprises the earliest available data and the out-of-sample the most recent. K-fold cross-validation would be preferable but requires greater time and computational resources. This computational overhead is especially costly in the context of the nonparametric nonlinear models evaluated in this thesis.

Another important factor is the degree to which the time ordered nature of asset prices is relevant that is, if the data are found to exhibit the characteristic of structural breaks - when the underlying variable experiences some fundamental change - then it may be inadvisable to choose models based on their performance on earlier and possibly redundant data as they will be biased.

2.11 Bias-Variance Tradeoff

As touched on before when estimating models it is rarely the objective to arrive at a function that represents the data exactly but rather, it is to build a model of the underlying data generating process in the hope that it will generalize well on new data.

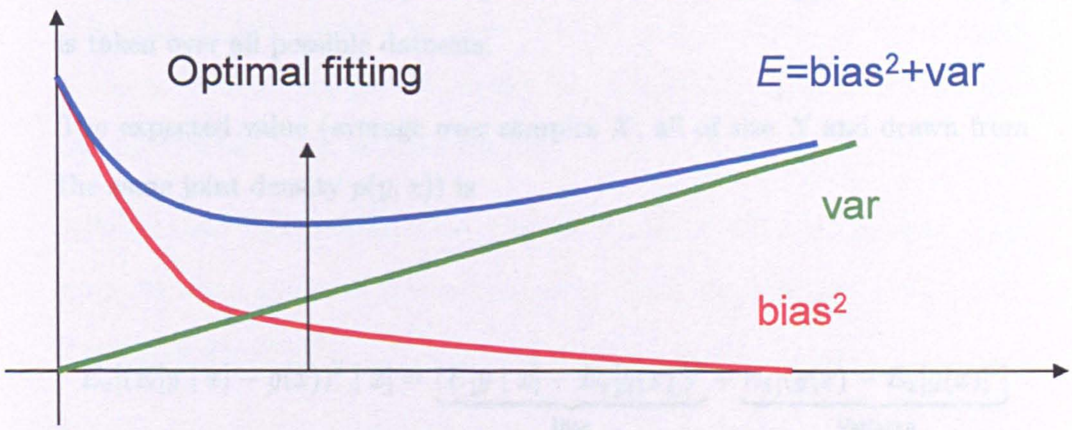
Intuitively, if the function fits the data exactly it is less likely to result in good out-of-sample performance as it will have fitted much of the noise contained within the sample. Conversely if the fit isn't close enough this will also generally lead to sub-optimal results as the function will not be sufficiently expressive.

It is instructive to think of this problem in the context of the so-called bias/variance trade off. The terms bias and variance come from the widely known decomposition of the generalization error of a model into the sum of the bias squared plus the variance (Geman, Bienenstock & Doursat 1992).

Bias. The extent to which the average (over all data sets) of the estimator differs from the desired function.

Variance. The extent to which the estimator fluctuates around its expected value as the sample varies. The best generalization is obtained with the optimal compromise between the conflicting requirements of small bias and small variance.

Figure 2.10: Bias Variance trade off.



More formally, a sample $X = x^t, y^t$, is drawn from unknown joint probability density $p(x, y)$. Using this sample, the estimate $g(.)$ is constructed. The expected square error (over the joint density) at x can be written as

$$E[(y-g(x))^2 \mid x] = \underbrace{E[(y-E[y \mid x])^2 \mid x]}_{\text{Noise}} + \underbrace{(E[y \mid x]-g(x))^2}_{\text{Squared error}} \tag{2.19}$$

The first term on the right is the variance of y given x; it does not depend on $g(.)$ or X. It is the variance of noise added, σ^2. This is the part of error that can never be removed, no matter what estimator is used. The second term quantifies how much $g(x)$ deviates from the regression function, $E[y|x]$. This does depend on the estimator and the training set. It may be the case that for one sample, $g(x)$ may be a very good fit; and for some other sample, it may make a bad fit. To quantify how well an estimator $g(.)$ is, the average is taken over all possible datasets.

The expected value (average over samples X, all of size N and drawn from the same joint density $p(y,x)$) is

$$E_x[(E[y \mid x]-g(x))^2 \mid x] = \underbrace{(E[y \mid x]-E_x[g(x)])^2}_{\text{Bias}} + \underbrace{E_x[(g(x)-E_x[g(x)]^2]}_{\text{Variance}} \tag{2.20}$$

As we mentioned previously, bias measures how much $g(x)$ is wrong disregarding the effort of varying samples, and variance measures how much $g(x)$ fluctuates around the expected value, $E[g(x)]$, as the sample varies.

For example: To estimate the bias and the variance, a number of datasets $X_i = x_i^t, y_i^t, i = 1, ..., M$, are generated from some known $f(.)$ with added noise. Each dataset is used to form an estimator $g_i(.)$, as $f(.)$ is unknown as are the parameters of the added noise. Then $E[g(x)]$ is estimated by the average over $g_i(.)$:

$\bar{g}(x) = \frac{1}{M}\sum_{i=1}^{M} g_i(x)$

Estimated bias and variance are:

$Bias^2(g) = \frac{1}{N}\sum_T[\bar{g}(x^t) - f(x^t)]^2$

$Variance(g) = \frac{1}{NM}\sum_t \sum_t [g_i(x^t) - \bar{g}_i(x^t)]^2$

Note that the expressions for bias and variance are functions of the input vector x. To best understand, consider the training set to be fixed and finite. Ideally it would be as large as possible.

Now consider a hypothesis space, that is the set of potential functions that map the input space to the output space. Less formally, just consider the number of parameters.

1. Too few parameters: high bias/low variance

2. Too many parameters: low bias/high variance

For example, consider a training set of n points.

1. One could fit a straight line – too few parameters (high bias/low variance).

2. The training set can always be fit exactly with a polynomial of degree $n - 1$ too many parameters (low bias/high variance).

Figure 2.11: Fitting a Polynomial - the bias variance trade off.

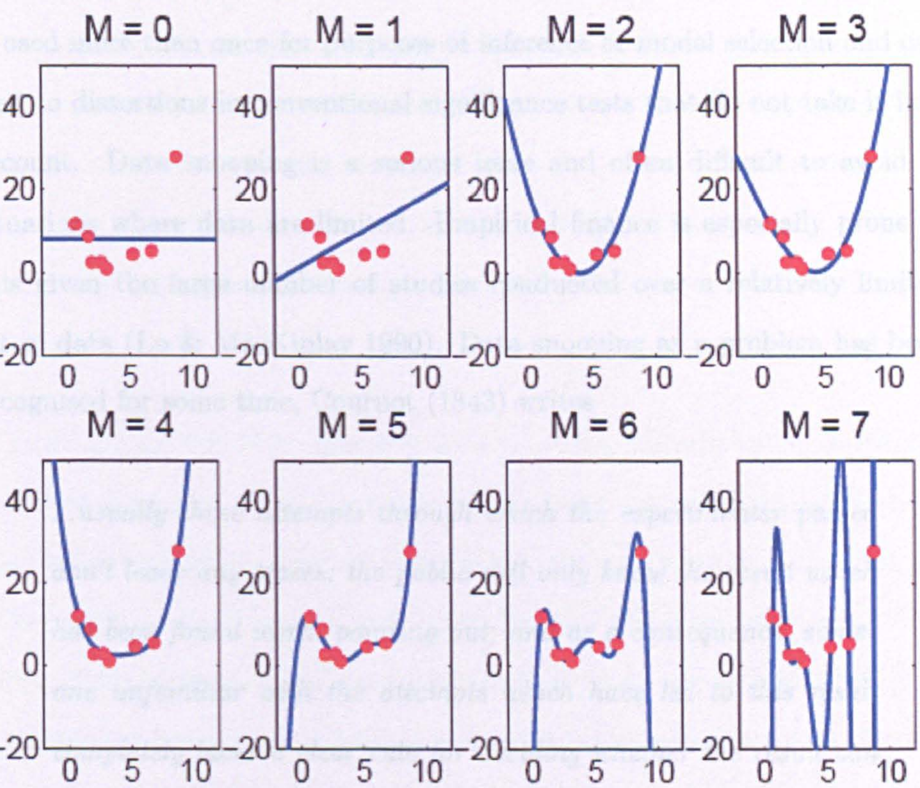


Figure 2.11 depicts an example where the order M of a polynomial is increased from 0 to 7. Initially polynomial of order zero can be seen to under-

fit the eight data points. The fit becomes more reasonable as the order is increased but finally it noticeably overfits the data.

2.12 Data snooping.

Data snooping is related to model selection and results when a set of data is used more than once for purposes of inference or model selection and can lead to distortions in conventional significance tests that do not take it into account. Data snooping is a serious issue and often difficult to avoid in situations where data are limited. Empirical finance is especially prone to this given the large number of studies conducted over a relatively limited set of data (Lo & MacKinlay 1990). Data snooping as a problem has been recognised for some time, Cournot (1843) writes

> *...usually these attempts through which the experimenter passed don't leave any traces; the public will only know the result which has been found worth pointing out; and as a consequence, someone unfamiliar with the attempts which have led to this result completely lacks a clear rule for deciding whether the result can or can not be attributed to chance.*

It can take a variety of forms, some obvious and others more subtle. An obvious example in the context of searching for profitable trading rules is where a researcher experiments with many different rules and parameters in-sample and chooses that rule which results in the best performance. The problem is that by testing many rules on the same data inevitably some will perform well even if just by chance, rather than due to any modelling of inherent structure in the data. As Sullivan et al. (1999) point out:

> *As time progresses, the rules that happened to perform well historically receive more attention and are considered serious contenders by the investment community, while unsuccessful trading rules are more likely to be forgotten. After a long sample period, only a small set of trading rules may be left for consideration, and these rules historical track record will be cited as evidence of their merits.*

A more subtle version is collective data snooping whereby many individual researchers conduct single in-sample ex-ante tests, yet unbeknownst to them, their individual efforts combine and result in the equivalent of data snooping. This occurs as those results that reject the null are more likely to be published while those that don't are filed away, never to see the light of day, resulting in the file-drawer problem/publication bias (Rosenthal 1979),(Denton 1985),(Ioannidis 2005). What is especially problematic is that researchers in this case are not aware of the consequences of their individual actions. There is an additional consequence in that later researchers exacerbate the situation by building on the original published, but ultimately snooped, results.

2.12.1 Accounting for data snooping

When conducting a single hypothesis test typically a maximum probability α, set at some conventional level (usually 0.05), is specified of rejecting the null when it is in fact true - this is the probability of committing a Type I error. When testing multiple hypotheses where each is assigned a Type I error probability, the probability that at least some Type I errors are committed increases. When it is recognised that a particular question of interest will involve testing more than one hypothesis at a time (multiple hypothesis testing or multiple inference as it is sometimes known) a number of methods have been developed, for a review see Shaffer (1995). The most commonly used method is to control the familywise error rate (FWE), which is the probability of making at least one false rejection.

White (2000) provides a formal framework that specifically quantifies the effects of data snooping and tests the null hypothesis that the best model unearthed during a specification search has no predictive ability over a benchmark model. It is appropriately named the "Bootstrap Reality Check" (BRC). In order to account for data snooping in the area of technical trading rules Sullivan et al. (1999) use the BRC to test whether trading rules found by Brock et al. (1992) still show significance when controlling for data snooping (they do).

The best rule is searched for by applying a performance statistic to all trading rules considered, then a p-value is obtained from comparing the performance of the best rule to approximations of the asymptotic distribution of the performance statistic. The BRC accounts for the increasing number of alternative models being tested by increasing the critical value as additional rules are added to the comparison. It is able to do this as the best performance statistic is a maximum, and the bootstrap procedure uses all rules being compared to compute bootstrap maxima, thus obtaining a non-parametric empirical distribution for the maximum (best) performance statistic under the null.

The hypotheses are:

H_0 : No method is better than the benchmark.

H_1: At least one method is better than the benchmark.

Hansen (2005) notes that the BRC is overly conservative and may reduce rejection probabilities under the null by not properly accounting for situations where trading rules that exhibit significantly inferior performance to the benchmark are included in the set of rules being tested.

2.12.2 Randomization tests

In this thesis we used bootstrap randomization tests to test for the significance of our results. All the studies are carefully constructed to allow robust out-of-sample testing. We do not explicitly account for the form of data-snooping above as it isn't necessary when estimating forecasting models whose model selection procedure involves selection ex-post on in-sample/validation data and testing only those models ex-ante on out-of-sample data.

The first empirical study we conduct is a study on forecasting FTSE100 stocks using three different types of models. Once we have completed the whole modelling procedure we wish to know how well the models perform out-of-sample. It's not enough to simply observe that the models' predictions resulted in an appreciable rate of return, we need also to know how they performed against models that have no "skill". One way of doing this is to compare the performance of each model to what we would expect if the models had no predictive ability.

A method of doing this is to use randomisation tests, initially introduced by R.A.Fisher (1935), which give one the probability of the observed model's performance assuming the null hypothesis of no skill. Randomisation tests are a useful alternative to more traditional parametric tests for analysing empirical research data. They have the advantage of not making any distributional assumptions about the data such as normality, as is often inappropriate for financial data, and remain as powerful as parametric tests such as the T-test.

In testing a single hypothesis, the probability of a Type I error, i.e., rejecting the null hypothesis when it is true, is usually controlled at some designated level α. The choice of α should be governed by considerations of the costs of rejecting a true hypothesis as compared with those of accepting a false one. Because of the difficulty in quantifying these costs and the subjectivity involved, α is usually set at some conventional level, often 0.05.

When using the more common T-test one assumes that the data arose by drawing samples from two normally distributed populations, with the question being whether the two populations differ in their mean, that is, how likely is it that the observed difference between the samples would be realised if there is no difference between the population statistics. The randomisation test, on the other hand, involves creating a large number of randomised data set replicates that could have arisen under the null (the null in this case being no predictive ability) and then computing some statistic and examining its distribution. The empirical distribution of this statistic is then used to estimate alpha, the probability of rejecting the null hypothesis when in fact the null is true.

We can create an empirical distribution under the null by scrambling the order of a model's trades N times and then comparing its realised performance to this distribution. The statistic used to measure performance in this case is termed the modified sharpe ratio (MSR), which is calculated by simply dividing the percentage return realised in the out-of-sample data by the standard deviation of the daily returns.

We randomise the order of actual trades (a single trade being a time period during which the model's daily forecasts do not change direction) 10,000 times and calculate the MSR for each replicate. Then we calculate how many of the random replicates MSRs' were below that of the particular model being tested.

Chapter 3

Forecasting FTSE 100 stocks using support vector regression, linear regression and k-nearest neighbour methods.

3.1 Introduction

In this chapter we attempt to predict the daily returns of 58 UK stocks in the FTSE 100 using three different methodologies, namely; support vector regression, k-nearest neighbours and linear regression. The objective is to compare the results of the three modelling techniques, linear and nonlinear, across a dataset of sufficient size to render statistically meaningful results. Concurrently, we wish to ascertain to what extent UK stock returns are predictable, if at all.

In pursuing this task the main priority was to keep the initial model as simple as possible, using simple independent variables and dispensing with any attempts at feature selection. The idea was to create a benchmark model with which to compare later models that will include a more detailed model building procedure, including the construction of ensembles.

3.2 Model Design and Methodology.

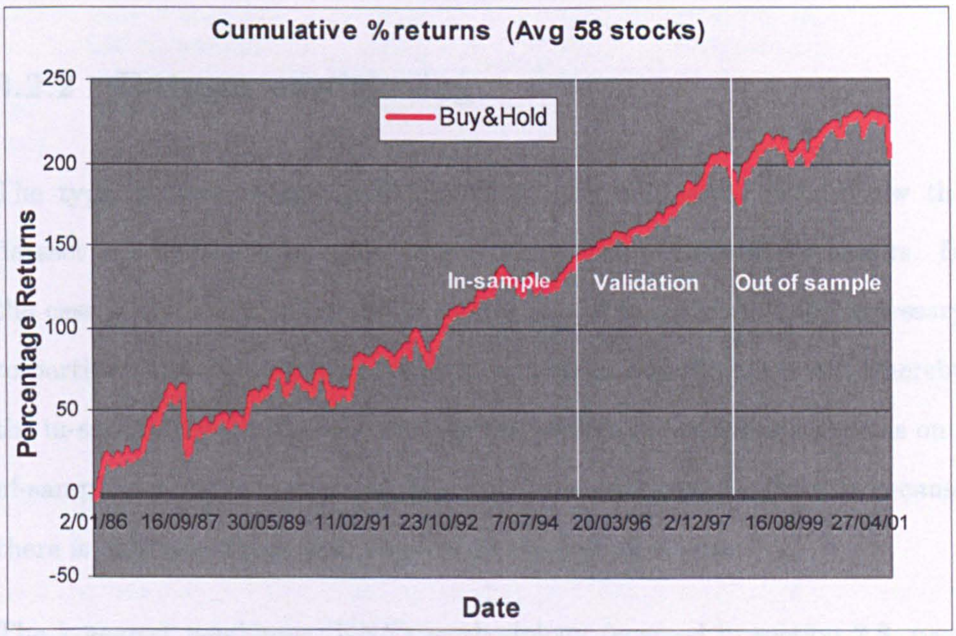
3.2.1 The Data

The FTSE 100 index (Financial Times Stock Exchange Index 100) is a share index of the 100 most highly capitalised UK companies listed on the London Stock Exchange that began on Jan 3rd 1984.

There are different approaches when it comes to deciding how much data to use when designing financial forecasting models. One view is that markets are always changing and therefore one does not want to use data too far back in history, as there is a danger that much of it will be redundant. The other approach is to use as much data as is available, reasoning that the only way to have confidence in the model's final results is if it has acceptable performance over as long a data history as possible. We subscribe to the latter approach.

As a sizeable data history was desirable only those stocks that had price histories going as far back as 1986 - approximately 4100 days - were chosen, resulting in a dataset of 58 FTSE 100 stocks of varying market capitalisations.

Figure 3.1: Cumulative percentage returns



The stocks were contained within the FTSE100 index as of August 2003 and the data consist of only the closing prices from 1986 to 2001. Data from 2001 to the present date are excluded as it is to be used for further out-of-sample testing in later models. This should go some way in ameliorating the data-snooping issue when analysing the results later in this thesis. Had we included and analysed data post 2001 our decisions may have been biased in later chapters regarding model choice/parameter values. As can be seen from figure 3.1 the data period includes the 87 crash in the in-sample period along with the events of 9/11 at the very end of the out-of-sample period.

The stocks that were including in this study are shown in Table 3.1 and were obtained from Datastream.

3.2.2 Dataset partitioning.

The type of forecasting algorithm being used will often dictate how the dataset is partitioned in order to get meaningful out-of-sample results. In the case of a linear regression forecasting model it is usually only necessary to partition the data into an in-sample set and an out-of-sample set, whereby the in-sample data is used to estimate the regression coefficients and the out-of-sample data is used to measure subsequent performance. This is because there is only one linear least squares fit through the data.

The k-nearest neighbour (KNN) methodology (covered in section 2.8, page 59) requires one to choose k, the number of neighbours, an initial starting point S from which to begin making predictions (in this case S was chosen to be 200 days), and a maximum "lookback" parameter L which determines

Table 3.1: Table of FTSE 100 stocks

Symbol	Name	MktCap		Symbol	Name	MktCap
ABF.L	ASSOC.BR.FOODS	4.101		KGF.L	KINGFISHER	6.335
ALLD.L	ALLIED DOMECQ	4.208		LGEN.L	LEGAL & GENERAL	6.47
AHM.L	AMERSHAM	3.601		LAND.L	LAND SECS GROUP	3.951
AVZ.L	AMVESCAP	3.943		MKS.L	MARKS & SP.	6.913
AV.L	AVIVA	11.319		MRW.L	MORRISON SUPMKT	3.259
BA.L	BAE SYSTEMS	5.225		PFG.L	PROVIDENT FINCL	1.624
BARC.L	BARCLAYS	30.126		PRU.L	PRUDENTIAL	8.761
BATS.L	BRIT AM TOBACCO	13.626		PSON.L	PEARSON	4.954
BOC.L	BOC GROUP	4.297		RB.L	RECKITT BENCKSR	8.399
BOOT.L	BOOTS GROUP	5.297		RBS.L	ROYAL BANK SCOT	45.715
BP.L	BP	94.816		REL.L	REED ELSEVIER	6.135
BT.L	BT GROUP	15.954		RTO.L	RENTOKIL INITIAL	3.843
BNZL.L	BUNZL PLC	2.107		REX.L	REXAM PLC	1.877
CBRY.L	CADBURY SCHWEPPE	7.857		RSA.L	ROYAL & SUN ALL	1.994
CW.L	CABLE & WIRELESS	2.836		RTR.L	REUTERS GROUP	3.595
DGE.L	DIAGEO	21.02		SN.L	SMITH&NEPHEW	3.775
DXNS.L	DIXONS GROUP	2.682		SFW.L	SAFEWAY	2.995
DMGOa.L	DAILY MAIL TST A	2.127		SBRY.L	SAINSBURY (J)	5.365
EMA.L	EMAP	2.191		SDR.L	SCHRODERS	1.602
EXL.L	EXEL	2.011		SCTN.L	SCOT & NEWCASTLE	3.422
FRCL.L	FOREIGN & COLONIAL	1.739		SHEL.L	SHELL TRNPT(REG)	38.332
GKN.L	GKN	1.923		SMIN.L	SMITHS GROUP	3.939
GSK.L	GLAXOSMITHKLINE	72.194		STAN.L	STNDRD CHART BK	9.903
GAA.L	GRANADA	2.837		TOMK.L	TOMKINS	2.033
GUS.L	GUS	6.916		TSCO.L	TESCO	15.191
HG.L	HILTON GROUP	3.068		ULVR.L	UNILEVER	14.979
HNS.L	HANSON	3.03		WTB.L	WHITBREAD	2.25
ICI.L	ICI	2.311		WOS.L	WOLSELEY	4.201
JMAT.L	JOHNSON MATTHEY	2.144		WPP.L	WPP GROUP	6.659

how far back in the price history neighbours will be searched for. Smaller values of L would imply recent data are more relevant. In the case where $L > S$ the lookback window will start at S and increase with t and will stay fixed once $L = t$. To keep the modelling procedure simple we created thirteen KNN models with different values for k and L. No attempt was made to optimize the parameters. This means that all forecasts made post S are effectively out-of-sample and we do not need to be concerned about partitioning the data set.

In the case of support vector regression there are a number of parameter values that need to be chosen and it is not known apriori what these values should be - this is explained in more detail below. In this case is it necessary to divide the data into three separate partitions, the in-sample, validation and out-of-sample sets. The algorithm is run using different starting values on the in-sample data. The validation set is then used to "validate" the models' parameters and those models exhibiting good performance, based on some metric (in this case the Sharpe Ratio), are then used to forecast the final out-of-sample data.

In order to make a fair comparison between the methods it is necessary to estimate the models using the same or at least similar data partitioning. The linear models are estimated with the in-sample data only. It might seem prudent to use both the in-sample and validation period combined as the in-sample set as these models only require two data partitions. However, this would not result in a fair comparison with the SVM models as they have only been estimated on the designated in-sample data.

Bearing the above in mind the data are divided into in-sample, validation and out-of-sample sets as shown in table 3.2. The in-sample training period covers 2500 days from 2nd January 1986 to 2nd August 1995, the validation set 800 days from 3rd August 1995 to 26th August 1998 and the out-of-sample set from 27th August 1998 to 25th September 2001, 804 days (see Table 3.2).

Table 3.2: Data set partitioning - Forecasting FTSE stocks.

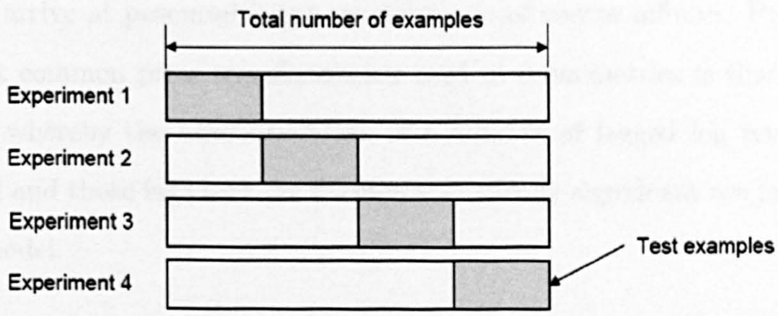
Set	Dates	Length
In-sample set	02 Jan 1986 to 02 Aug 1995	2500 Days
Validation set	03 Aug 1995 to 26 Aug 1998	800 Days
Out-of-sample set	27 Aug 1998 to 25 Sep 2001	804 Days

The SVM models were trained on the in-sample data using different starting parameter values. Subsequent performance was then checked on the validation set, and those models exhibiting the best performance were selected - this procedure is explained in more detail in section 3.3.1 below. The linear regression models were estimated using the same in-sample data set but without the need to check performance on the validation set. The knn models start making predictions from 200 days into the in-sample period. We dispensed with any knn model selection procedure and simply created thirteen models using different parameter values with the final prediction formed by a majority vote - see Table 3.11 for the values used.

One issue of concern can be seen in figure 3.1, which depicts the average of all 58 stocks as if they were simply bought and held. The validation period of 800 days consists of a very strong upward trend in the market and can have the effect of biasing the final models to only those that perform well during strong bull markets (when returns>0 are more prevalent than returns<0).

One way to ameliorate this would be to use $k - fold$ cross validation where the in-sample and validation data are combined and then split into k separate $folds$, k usually being 5-10. The models are then trained on $k-1$ sets, each time leaving one out for the validation process. This is repeated until all sets have at one point been used as a validation set. Figure 3.2 shows an example using k=4. *Leave-one-out* cross validation is k-fold cross validation taken to its logical extreme, with k equal to N, the number of data points in the set. It should be noted that $k - fold$ cross validation is generally more effective on small data sets than the simple $1 - fold$ hold out method.

Figure 3.2: K-fold Cross Validation (K=4)



This should result in models that are not biased to any particular part of the data. It would also have the advantage of producing models that have been trained on the most recent part of the data (the validation set) just before the final out-of-sample data. This could go some way in improving the model as data from the recent past may be more relevant to predicting the future. We choose not to perform an $k - fold$ cross-validation as it would require us to create and train $n-1$ extra models, and it was decided that potential benefits would not outweigh the costs at this stage.

3.2.3 Independent and Dependent variables.

Finding a good representation of the data to use as inputs (independent) variables and output or target (dependent) variables is very important, especially when building financial forecasting models. The objective is to find a representation that will render the signal (if one exists) more explicit and/or attenuate the noise component. The number of possible transformations of price to arrive at potential input candidates is of course infinite. Probably the most common price transformation used in econometrics is that of log returns, whereby the autocorrelation of a number of lagged log returns is analysed and those lags that are deemed statistically significant are included in the model.

In this study we use a different transformation. The rationale behind this is that log returns can be seen in a signal processing context as a high pass filter that passes higher frequencies and attenuates the lower frequencies contained within the data (Abarbanel 1995). In this sense there is a loss of information that may be detrimental to the model. Moreover, in the case of financial forecasting models there is the ever present issue of transaction costs. The greater the number of the trades the greater the transaction costs, hence models that trade less frequently can be more desirable. If one's price sampling frequency is daily and the price of the asset in question does not move by very much in an absolute sense, compared with the associated transaction cost, then building a forecasting model that is biased towards higher frequencies (using log returns as either the inputs or target, or both) means that trade frequency is concomitantly higher, thus potentially causing any resultant model to be overwhelmed by transaction costs. Taking this into account we choose to use raw logs of prices but transformed in such a way as to render them stationary and thus allowable in a regression model.

3.2.4 The Input Variables.

At time T, raw log price data (not returns) from $T - 15$ to T inclusive are detrended by subtracting an end-to-end linear trend line from the prices. This zeros the values at T and $T - 15$, and the resulting 14 remaining values are then scaled to lie between 0 and 1. This process renders the data stationary and also has the effect of causing the inputs to be amplitude invariant due to the scaling procedure that is, using this transformation, a highly volatile price period and a relatively calm period that share the same "waveshape" will be rendered equivalent. Finally the gradient from the detrending process is included as an input.

Standard daily log returns were chosen as the target or dependent variable though later in the thesis other targets are used:

$$x_t = log(price_{t+1}/price_t) \tag{3.1}$$

3.3 Support Vector Machine Regression (SVMR) models.

3.3.1 Model Selection.

As optimum starting parameters are not known apriori we are required to build a number of SVMR models with different starting parameter values, choosing those models that fit both the in-sample and validation data sets to create the final forecasting model, the output of which takes the form of a majority vote across the best performing models. Rather than choosing the single best model this ensemble averaging process should reduce the variance of the final model. In order to estimate the generalisation error and to select models for the final SVMR ensemble for each individual stock a cross-validation scheme is used whereby twenty one SVMR models are initially trained on the in-sample set of 2500 days. To do this the following model building procedure was followed:

1. Using in-sample data build twenty one SVMR models each using dif-

ferent starting parameter values (see Table 3.10 for the values used).

2. With the Modified Sharpe Ratio (MSR) (this is simply a non-annualised Sharpe Ratio [1]) as a performance measure, rank each model's performance on the in-sample data and remove the worst seven, leaving fourteen models.

   The SVMR models produce a real output and in order to convert this into a prediction it is necessary to decide on a decision rule. In this case we used:

   Model outputs are assigned 1 if >= 0.0 and -1 if < 0.0.

   This ignores the magnitude of each forecast.

3. Evaluate the MSR on validation dataset for remaining fourteen models and remove the seven worst performing models.

4. The final seven models' forecasts are then combined using a majority vote - see below. Use majority voting rule to analyse the performance of the top seven models on the out-of-sample dataset. The combined model forecasts are then taken as the final SVMR model.

Step 2 is performed as it is not desirable to have models that under-fit the in-sample data in the final model pool, regardless of the performance over the validation set.

[1] The fact that the SR is not annualised is acceptable as it is measured over the same time periods for all models

Majority Voting Rule:

1. Each SVMR sub-model output is assigned 1 if >= 0.0 and -1 if < 0.0.

2. The result for each model on each day is then summed, producing a number representing the majority decision which can take the values -7, -5, -3, -1, 1, 3, 5, 7.

3. A long (short) forecast occurs if this majority is above (below) zero.

Each trade is then held until the majority decision signals a trade in the opposite direction. This results in what is commonly called a stop & reverse model and is always in the market - there are no flat periods.

The SVMR method has a number of *tunable* parameters that need to be determined by the user: C, a regularisation parameter that determines the trade-off between margin maximization and training error (empirical risk) minimization. C determines the trade off between the model complexity (flatness) and the degree to which deviations larger than ε are tolerated in the optimization formulation for example, if C is too large (infinity), then the objective is to minimize the empirical risk only, without regard to model complexity. The other parameter ε controls the width of the ε-insensitive tube, used to fit the training data. The value of ε can affect the number of support vectors used to construct the regression function. Larger values of ε result in the selection of fewer support vectors. On the other hand, larger ε

-values result in "flatter" estimates. Hence, both C and ε-values affect model complexity but in different ways.

In addition one has to choose the type of kernel along with any kernel related parameters. It is the kernel that defines the high dimensional feature space where the maximal hyperplane will be found. In the case of the radial basis function kernel, the width of the RBF kernel σ^2 needs to be chosen. See Table 3.10 for the parameters sets that were used in this study.

The choice of kernel determines the form of the resulting learning machine. Common kernel functions include polynomial, radial basis functions (RBF) and sigmoid kernels. In this research we used a number of different kernels as apriori we do not know which one is optimum. As an example, RBF kernels have the form:

RBF:

$$K(\mathbf{x}, \mathbf{z}) = exp\left(-\frac{\|x - z\|^2}{2\sigma^2}\right) \tag{3.2}$$

where σ^2 is the width of the kernel.

In all one thousand two hundred and eighteen separate SVMR models were built and trained. Each stock required twenty one models, each taking up to twenty minutes to train.

3.3.2 K-Nearest Neighbour models.

The parameters chosen for the KNN model can be found in table 3.11. The 13 models were combined using a majority voting procedure as follows:

Majority Voting Rule (KNN):

1. Each KNN model output is assigned 1 if >= 0.0 and -1 if < 0.0.
2. The result for each model on each day is then summed, producing a number representing the majority decision which can take values between -13 and 13.
3. A long (short) forecast occurs if this majority is above (below) zero.

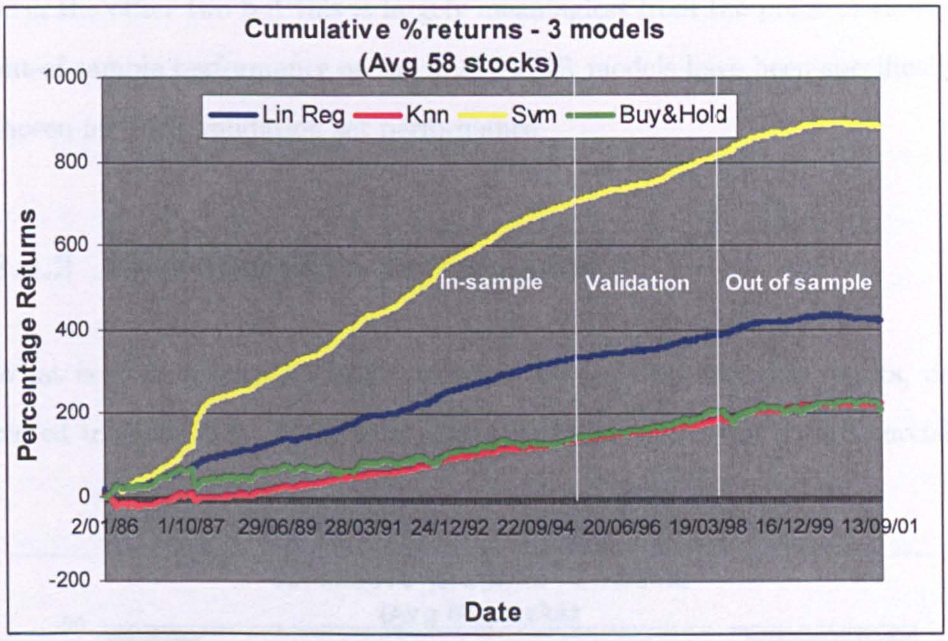
Each trade is then held until the majority decision signals a trade in the opposite direction. This results in what is commonly called a stop & reverse model and is always in the market - there are no flat periods.

3.4 Results.

3.4.1 In-sample and validation set performance.

We present the results top-down, initially as charts followed by more detailed tables. Figure 3.3 shows the cumulative percentage returns of all three models

Figure 3.3: FTSE Results over whole data set



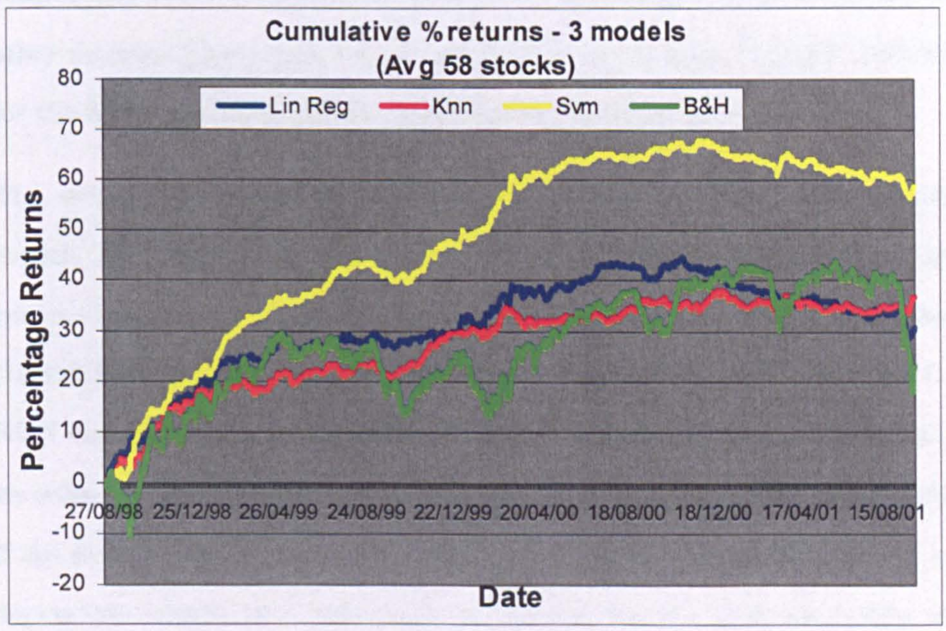
averaged across all stocks over the whole dataset and includes the buy & hold as a comparison - this is an average of the buy & hold percentage returns for each stock and is not weighted by market capitalisation. What is immediately apparent is that the SVMR models have fit the in-sample data more closely than either the regression or the KNN models. This is not surprising as the SVMR models have a greater capacity and have been explicitly trained on this part of the dataset. The regression model exhibits better performance on the in-sample period than the KNN models but this is also expected as the results from the KNN models are effectively out-of-sample forecasts over the in-sample data, except of course for the first 200 days.

Looking at the validation set we can see that again the SVMR models outperform the other two but this is largely meaningless from the point of view of out-of-sample performance as the final SVMR models have been specifically chosen for their validation set performance.

3.4.2 Out-of-sample performance.

What is more interesting and important are the out-of-sample results, depicted in figure 3.4. The cumulative percentage returns of SVMR models

Figure 3.4: Cumulative percentage returns 3 Models



shown in yellow continue to outperform the other models along with the buy

& hold (at least visually at this point - see detailed tables below). The curve rises steadily until the last quarter of the data when it then drops, along with the market as a whole during the period of the 9.11 attack on New York.

The linear and KNN models, although also exhibiting rising equity curves, do not appear to be outperforming the buy & hold except that their curves are smoother, which should result in greater MSRs. This is indeed the case (See table 3.3).

Table 3.3 shows summarised results for each method averaged over the 58 stocks in the data sample. On average the SVMR models resulted in a 58% total return over the 803 days compared to 36% and 30% respectively for the other models. This compared to 22% for the buy & hold. At 23.3 the MSR for the SVM models is almost twice that of the linear models.

The average percentage of bootstrapped trade replicates that the SVMR models' MSR statistics were greater than was 70% and 61% for the other two models. Those stocks exhibiting MSR statistics over 95% greater than the random replicates number thirteen for the SVMR models, five for the KNN models and six for the linear models. Of the 58 stocks out-of-a-sample, we would expect three stocks on average to show greater than 95% results if the forecasting models had no predictive accuracy. An assumption of independence clearly isn't reasonably in this case but if it were then using the binomial probability distribution we could work out how likely it is that the SVMR models achieve 13 successes out of 58 trials if we assume the models have no predictive ability. In this case the probability of a success is 0.05, that is, the probability that a model's MSR will be greater than the MSR of 95% of the replicates is 0.05. This indicates that as an upper bound there is a 1:250,000 chance of 13 successes out of 58 trials.

Table 3.3: Table of Results for all 3 algorithms

Linear Regression Results. Out of sample data 803 Days

	FinalEquity in %	Mod. Sharpe R	Percentage of random below	Number of trades	Average Trade %	% Trade Acuracy	% Daily Accuracy	Buy&Hold FinalEquity %	Buy&Hold Mod. Sharpe	MktCap
Average all 58	30.5	12.7	61.4%	304.2	0.11%	0.46%	50.0%	22.5	9.0	9.7

Number of stocks with M.S.R. greater than 95%	6
Number of stocks with M.S.R. greater than 90%	12

KNN Results. Out of sample data 803 Days

	FinalEquity in %	Modified Sharpe	Percentage of random below	Number of trades	Average Trade %	% Trade acuracy	Daily accuracy	B&H FinalEquity%	Buy&Hold Modified Sharpe	MktCap
Average all 58	36.4	14.2	61.5%	289.5	0.14%	0.48%	50.3%	22.5	9.0	9.7

Number of stocks with M.S.R. greater than 95%	5
Number of stocks with M.S.R. greater than 90%	13

Support Vector Regression Results. Out of sample data 803 Days

	FinalEquity in %	Mod. Sharpe R	Percentage of random below	Number of trades	Average Trade %	% Trade Acuracy	% Daily Accuracy	Buy&Hold FinalEquity %	Buy&Hold Mod. Sharpe	MktCap
Average all 58	58.1	23.2	69.3%	284.4	0.20%	0.44%	50.8%	22.5	9.0	9.7

Number of stocks with M.S.R. greater than 95%	13
Number of stocks with M.S.R. greater than 90%	16

More detailed results for each stock can be seen in tables 3.4 to 3.9. The final percentage equity and associated MSRs for each stock are shown in columns 2 and 3.

The out-of-sample results for the three modelling methods and for each stock can be seen in tables 3.4 to 3.9. Following the column of stock symbols, the second column shows the final percentage return achieved over the 803 days for each stock. This is followed by the modified sharpe ratio which is simply the final percentage equity divided by the standard deviations of daily percentage returns. The third column shows the results of the randomization tests in which the trades were bootstrapped 10000 times and the resulting MSR then measured. The figure represents where in the distribution the MSR for a particular stock falls, thus indicating the significance of the results. For example, in the case of stock ABF in table 4, the linear regression model achieved a MSR above 58.7 percent of all the bootstrapped trades' MSRs. Columns 5-8 show how accurate the forecasts were on a per trade and daily basis in addition to the magnitude of the returns. Lastly the results of the buy & hold are shown which serve as a comparison.

Based on these results we conclude that the nonlinear SVMR models do indeed outperform the linear and KNN models over the out-of-sample period, which suggests that they are able to model a certain amount of nonlinearity within the data that is not exploitable by the other two methods.

Note that transaction costs have not been included as the objective of the study is to compare the results of the different modelling procedures at this stage. Therefore there is not a claim that these results would result in above-average risk-adjusted returns. However, the target/dependent variable in the study was one day log returns. By incorporating a target over a longer period (say 5 days returns) would likely reduce the number of trades and hence transaction costs. Whether the returns would experience a less than proportionate reduction in tandem is an open question.

3.5 Summary and Conclusion.

In this chapter we have created forecasting models for 58 FTSE 100 stocks using two nonlinear methods and one linear. Using a dataset of 4100 days including an out-of-sample period of 803 days, the nonlinear SVMR models on average are shown to outperform the other two methods, in addition to the buy & hold - though no transactions costs were included. The SVMR models resulted in thirteen stocks exhibiting a MSR figure that was greater than 95% of the bootstrapped replicates in the out-of-sample period.

The results for the KNN and linear regression models were approximately equal and also showed performance above that that would be expected if they had no forecasting power. It is also apparent from the average return per trade - 0.20% for the SVMR models - that if transaction costs were included, all models on average would under-perform the buy and hold.

3.5.1 Contributions of Study 3

1. We contribute to the literature on forecasting financial markets by conducting a robust study comparing the performance of three different forecasting methods namely, support vector machines (SVMs), k-nearest neighbours and linear regression and our results suggest that the nonlinear SVM models outperform the others.

2. We contribute the literature on SVMs by providing an incremental addition to the expanding literature that show SVMs to be a competitive methodology when applied with real world datasets.

3. We show for the first time in the literature the results of applying the relatively new method of SVMs applied to FTSE 100 stocks.

4. We apply support vector machine regression to the largest financial data-set used in the literature.

Table 3.4: Linear Regression Results out-of-sample data 1-30

Linear Regression Results. Out of sample data 803 Days (stocks 1 - 30)										
Stock Symbol	FinalEquity in %	Mod. Sharpe R	Percentage of random below	Number of trades	Average Trade %	% Trade Acuracy	% Daily Accuracy	Buy&Hold FinalEquity %	Buy&Hold Mod. Sharpe	MktCap
ABF.L	30.6	8.5	57.8%	274	0.11%	0.46%	50.5%	90.0	25.1	4.101
ALLD.L	6.1	2.6	44.4%	339	0.02%	0.45%	48.7%	43.6	18.3	4.208
AHM.L	-34.2	-10.9	27.7%	235	-0.15%	0.37%	46.8%	68.2	21.8	3.601
AVZ.L	-107.5	-42.2	3.3%	341	-0.32%	0.44%	47.1%	11.3	4.4	3.943
AV.L	98.4	39.3	92.6%	367	0.27%	0.47%	52.1%	9.1	3.6	11.319
BA.L	82.2	28.2	84.0%	483	0.17%	0.46%	51.6%	15.4	5.3	5.225
BARC.L	105.7	40.3	90.4%	294	0.36%	0.48%	50.6%	57.8	22.1	30.126
BATS.L	-77.6	-24.4	10.4%	409	-0.19%	0.43%	46.2%	92.3	29.1	13.626
BOC.L	-45.9	-23.9	15.1%	371	-0.12%	0.45%	48.3%	18.1	9.4	4.297
BOOT.L	73.2	34.4	91.5%	252	0.29%	0.49%	51.4%	-34.9	-16.4	5.297
BP.L	95.6	47.9	95.6%	291	0.33%	0.48%	53.0%	47.0	23.5	94.816
BT.L	-75.2	-24.4	19.4%	293	-0.26%	0.41%	48.0%	-32.1	-10.4	15.954
BNZL.L	190.3	125.2	100.0%	283	0.67%	0.55%	55.1%	73.4	48.1	2.107
CBRY.L	68.0	33.4	88.5%	224	0.30%	0.53%	51.8%	14.9	7.3	7.857
CW.L	31.8	9.1	64.7%	325	0.10%	0.51%	50.7%	-28.2	-8.1	2.836
DGE.L	12.0	5.2	53.5%	265	0.05%	0.45%	50.8%	33.9	14.6	21.02
DXNS.L	89.4	26.5	83.2%	236	0.38%	0.45%	51.9%	63.7	18.9	2.682
DMGOa.L	-1.3	-0.5	46.2%	356	0.00%	0.48%	48.2%	22.5	8.2	2.127
EMA.L	106.2	39.0	92.1%	379	0.28%	0.47%	50.8%	-33.0	-12.1	2.191
EXL.L	80.6	34.9	86.4%	298	0.27%	0.47%	54.6%	1.0	0.4	2.011
FRCL.L	26.7	26.2	75.9%	209	0.13%	0.48%	52.2%	17.7	17.3	1.739
GKN.L	-169.2	-68.9	0.6%	287	-0.59%	0.37%	46.5%	8.2	3.3	1.923
GSK.L	95.1	45.0	94.8%	309	0.31%	0.55%	53.3%	18.5	8.8	72.194
GAA.L	34.9	12.2	67.7%	258	0.14%	0.50%	51.1%	-8.5	-3.0	2.837
GUS.L	-141.2	-50.7	1.3%	209	-0.68%	0.43%	44.3%	-10.9	-3.9	6.916
HG.L	-73.8	-24.9	18.5%	399	-0.18%	0.44%	47.6%	6.6	2.2	3.068
HNS.L	62.0	25.0	78.8%	283	0.22%	0.48%	50.7%	63.9	25.8	3.03
ICI.L	35.0	13.1	71.0%	304	0.12%	0.51%	51.4%	-38.8	-14.6	2.311
JMAT.L	50.3	20.1	67.3%	228	0.22%	0.48%	51.8%	110.2	43.9	2.144
Average all 58	30.5	12.7	61.4%	304.2	0.11%	0.46%	50.0%	22.5	9.0	9.7

Table 3.5: Linear Regression Results out-of-sample data 31-58

Linear Regression Results. Out of sample data 803 Days (stocks 30-58)										
Stock Symbol	FinalEquity in %	Mod. Sharpe R	Percentage of random below	Number of trades	Average Trade %	% Trade Acuracy	% Daily Accuracy	Buy&Hold FinalEquity %	Buy&Hold Mod. Sharpe	MktCap
KGF.L	61.1	22.9	79.5%	248	0.25%	0.45%	51.1%	-30.5	-11.5	6.335
LGEN.L	63.9	25.2	82.3%	380	0.17%	0.48%	52.1%	6.6	2.6	6.47
LAND.L	0.3	0.2	51.1%	343	0.00%	0.46%	49.1%	-1.0	-0.7	3.981
MKS.L	-38.9	-14.6	33.6%	419	-0.09%	0.47%	49.1%	-53.7	-20.2	6.913
MRW.L	-1.4	-0.6	32.9%	259	-0.01%	0.46%	50.4%	62.7	27.5	3.289
PFG.L	-28.9	-12.5	34.4%	262	-0.11%	0.38%	47.2%	-13.3	-5.8	1.624
PRU.L	35.6	14.1	71.0%	258	0.14%	0.47%	51.6%	-1.0	-0.4	8.761
PSON.L	69.1	22.7	78.0%	308	0.22%	0.46%	49.3%	19.3	6.3	4.954
RB.L	-2.6	-1.0	45.1%	283	-0.01%	0.48%	50.6%	18.2	7.1	8.399
RBS.L	103.5	37.8	85.8%	277	0.37%	0.47%	49.5%	85.5	31.2	45.715
REL.L	66.3	25.2	77.2%	287	0.23%	0.47%	49.8%	50.4	19.1	6.135
RTO.L	73.3	26.4	85.3%	238	0.31%	0.44%	48.9%	-9.1	-3.3	3.843
REX.L	110.6	43.2	91.3%	306	0.36%	0.45%	48.7%	79.3	31.0	1.877
RSA.L	35.0	12.6	67.8%	313	0.11%	0.44%	49.5%	-12.9	-4.6	1.994
RTR.L	202.0	54.8	96.4%	265	0.76%	0.51%	51.4%	67.3	18.2	3.595
SN.L	-34.8	-16.1	24.7%	384	-0.09%	0.43%	47.1%	101.9	47.3	3.775
SFW.L	-107.9	-49.8	2.8%	334	-0.32%	0.41%	47.1%	22.4	10.3	2.998
SBRY.L	-16.9	-7.1	41.5%	331	-0.05%	0.45%	51.1%	-18.7	-7.9	5.365
SDR.L	53.7	17.3	82.8%	221	0.24%	0.53%	49.8%	-10.0	-3.2	1.602
SCTN.L	110.8	48.6	97.5%	302	0.37%	0.51%	52.8%	-17.6	-7.7	3.422
SHEL.L	128.2	60.7	98.0%	323	0.40%	0.52%	52.1%	50.8	24.1	38.332
SMIN.L	14.1	6.1	57.7%	299	0.05%	0.43%	49.7%	14.3	6.2	3.939
STAN.L	32.5	10.7	59.8%	292	0.11%	0.49%	49.6%	69.8	22.9	9.903
TOMK.L	-35.5	-14.4	33.2%	313	-0.11%	0.45%	47.3%	-27.8	-11.3	2.033
TSCO.L	28.7	13.2	63.7%	413	0.07%	0.47%	49.5%	51.3	23.5	15.191
ULVR.L	34.6	15.3	70.2%	282	0.12%	0.48%	50.0%	2.1	1.0	14.979
WTB.L	-38.2	-19.0	29.4%	273	-0.14%	0.44%	47.8%	-30.7	-15.3	2.25
WOS.L	44.7	17.5	68.0%	272	0.16%	0.42%	50.6%	48.7	19.1	4.201
WPP.L	156.7	49.3	96.4%	358	0.44%	0.47%	51.8%	78.7	24.7	6.659
Average all 58	30.5	12.7	61.4%	304.2	0.11%	0.46%	50.0%	22.5	9.0	9.7

Table 3.6: KNN Results out-of-sample data 1-29

KNN Results. Out of sample data 803 Days (stocks 1-29)										
Stock Symbol	FinalEquity%	Modified Sharpe	Percentage of random below	Number of trades	Average Trade %	% Trade acuracy	Daily accuracy	B&H FinalEquity%	Buy&Hold Modified Sharpe	MktCap
ABF.L	-51.5	-14.3	18.8%	259	-0.20%	0.39%	47.3%	90.0	25.1	4.101
ALLD.L	51.0	21.5	73.8%	277	0.18%	0.52%	49.3%	43.6	18.3	4.208
AHM.L	-99.2	-31.7	6.8%	325	-0.31%	0.43%	44.3%	68.2	21.8	3.601
AVZ.L	88.6	34.8	92.0%	334	0.27%	0.51%	53.0%	11.3	4.4	3.943
AV.L	141.4	56.6	98.5%	259	0.55%	0.51%	52.4%	9.1	3.6	11.319
BA.L	190.3	65.3	99.2%	259	0.73%	0.53%	54.5%	15.4	5.3	5.225
BARC.L	121.4	46.3	93.8%	279	0.44%	0.45%	51.0%	57.8	22.1	30.126
BATS.L	21.2	6.7	52.7%	316	0.07%	0.47%	50.1%	92.3	29.1	13.626
BOC.L	-47.3	-24.6	17.8%	326	-0.15%	0.44%	48.0%	18.1	9.4	4.297
BOOT.L	-9.1	-4.3	43.9%	323	-0.03%	0.51%	51.2%	-34.9	-16.4	5.297
BP.L	26.7	13.4	60.9%	295	0.09%	0.46%	52.0%	47.0	23.5	94.816
BT.L	83.6	27.1	86.6%	221	0.38%	0.52%	51.1%	-32.1	-10.4	15.954
BNZL.L	161.2	105.9	99.8%	237	0.68%	0.51%	55.1%	73.4	48.1	2.107
CBRY.L	-28.1	-13.8	29.7%	328	-0.09%	0.44%	49.5%	14.9	7.3	7.857
CW.L	99.1	28.5	86.8%	277	0.36%	0.54%	51.9%	-28.2	-8.1	2.836
DGE.L	30.2	13.0	65.1%	319	0.09%	0.49%	50.3%	33.9	14.6	21.02
DXNS.L	89.2	26.5	79.4%	272	0.33%	0.47%	50.8%	63.7	18.9	2.682
DMGOa.L	-10.3	-3.8	40.5%	266	-0.04%	0.45%	48.6%	22.5	8.2	2.127
EMA.L	28.7	10.5	65.5%	261	0.11%	0.43%	48.9%	-33.0	-12.1	2.191
EXL.L	181.8	78.8	99.3%	197	0.92%	0.51%	54.0%	1.0	0.4	2.011
FRCL.L	-19.6	-19.2	16.3%	243	-0.08%	0.39%	48.8%	17.7	17.3	1.739
GKN.L	101.9	41.5	93.5%	300	0.34%	0.49%	48.9%	8.2	3.3	1.923
GSK.L	90.6	42.9	94.4%	251	0.36%	0.53%	52.2%	18.5	8.8	72.194
GAA.L	11.4	4.0	55.2%	280	0.04%	0.46%	49.9%	-8.5	-3.0	2.837
GUS.L	-142.7	-51.2	3.3%	315	-0.45%	0.45%	48.2%	-10.9	-3.9	6.916
HG.L	76.8	25.9	82.6%	294	0.26%	0.55%	52.3%	6.6	2.2	3.068
HNS.L	12.0	4.9	49.8%	272	0.04%	0.49%	50.9%	63.9	25.8	3.03
ICI.L	-36.0	-13.5	29.3%	328	-0.11%	0.41%	48.7%	-38.8	-14.6	2.311
JMAT.L	34.2	13.6	63.5%	244	0.14%	0.45%	52.0%	110.2	43.9	2.144
Average all 58	36.4	14.2	61.5%	289.5	0.14%	0.48%	50.3%	22.5	9.0	9.7

Table 3.7: KNN Results out-of-sample data 30-58

KNN Results. Out of sample data 803 Days (stocks 30-58)										
Stock Symbol	FinalEquity%	Modified Sharpe	Percentage of random below	Number of trades	Average Trade %	% Trade acuracy	Daily accuracy	B&H FinalEquity%	Buy&Hold Modified Sharpe	MktCap
KGF.L	42.8	16.1	74.6%	319	0.13%	0.45%	50.9%	-30.5	-11.5	6.335
LGEN.L	28.6	11.3	67.4%	263	0.11%	0.52%	47.3%	6.6	2.6	6.47
LAND.L	10.8	7.6	61.5%	301	0.04%	0.47%	50.3%	-1.0	-0.7	3.981
MKS.L	56.0	21.1	75.7%	331	0.17%	0.47%	50.9%	-53.7	-20.2	6.913
MRW.L	4.7	2.1	43.8%	277	0.02%	0.42%	48.6%	62.7	27.5	3.289
PFG.L	-1.6	-0.7	51.9%	296	-0.01%	0.44%	50.7%	-13.3	-5.8	1.624
PRU.L	52.2	20.7	78.3%	234	0.22%	0.48%	50.1%	-1.0	-0.4	8.761
PSON.L	16.0	5.3	53.4%	259	0.06%	0.52%	49.6%	19.3	6.3	4.954
RB.L	43.6	17.0	70.5%	317	0.14%	0.49%	50.6%	18.2	7.1	8.399
RBS.L	123.9	45.2	89.6%	263	0.47%	0.50%	53.0%	85.5	31.2	45.715
REL.L	-79.3	-30.1	12.0%	318	-0.25%	0.46%	49.7%	50.4	19.1	6.135
RTO.L	105.2	37.8	90.1%	313	0.34%	0.49%	48.8%	-9.1	-3.3	3.843
REX.L	19.4	7.6	62.5%	268	0.07%	0.46%	49.7%	79.3	31.0	1.877
RSA.L	146.1	52.7	96.8%	333	0.44%	0.49%	51.4%	-12.9	-4.6	1.994
RTR.L	173.5	47.1	93.9%	248	0.70%	0.51%	51.6%	67.3	18.2	3.595
SN.L	6.1	2.8	42.1%	337	0.02%	0.49%	48.9%	101.9	47.3	3.775
SFW.L	-98.0	-45.2	4.6%	337	-0.29%	0.47%	48.6%	22.4	10.3	2.998
SBRY.L	23.9	10.1	66.0%	317	0.08%	0.47%	50.8%	-18.7	-7.9	5.365
SDR.L	-54.1	-17.4	24.2%	307	-0.18%	0.49%	49.7%	-10.0	-3.2	1.602
SCTN.L	-36.9	-16.2	27.8%	269	-0.14%	0.48%	49.7%	-17.6	-7.7	3.422
SHEL.L	56.2	26.6	76.1%	275	0.20%	0.50%	50.4%	50.8	24.1	38.332
SMIN.L	-8.1	-3.5	45.1%	299	-0.03%	0.44%	49.9%	14.3	6.2	3.939
STAN.L	17.7	5.8	54.8%	309	0.06%	0.46%	49.3%	69.8	22.9	9.903
TOMK.L	-37.5	-15.2	27.0%	323	-0.12%	0.46%	50.5%	-27.8	-11.3	2.033
TSCO.L	87.8	40.4	92.1%	319	0.28%	0.52%	50.4%	51.3	23.5	15.191
ULVR.L	6.4	2.8	52.0%	297	0.02%	0.46%	48.8%	2.1	1.0	14.979
WTB.L	47.4	23.6	77.6%	279	0.17%	0.54%	50.2%	-30.7	-15.3	2.25
WOS.L	118.8	46.6	94.4%	314	0.38%	0.48%	51.8%	48.7	19.1	4.201
WPP.L	39.6	12.4	61.9%	311	0.13%	0.46%	49.5%	78.7	24.7	6.659
Average all 58	36.4	14.2	61.5%	289.5	0.14%	0.48%	50.3%	22.5	9.0	9.7

Table 3.8: Support Vector Regression Results out-of-sample data 1-29

Support Vector Regression Results. Out of sample data 803 Days (stocks1-29)										
Stock Symbol	FinalEquity in %	Mod. Sharpe R	Percentage of random below	Number of trades	Average Trade %	% Trade Acuracy	% Daily Accuracy	Buy&Hold FinalEquity %	Buy&Hold Mod. Sharpe	MktCap
ABF.L	111.5	31.1	85.5%	287	0.39%	0.40%	49.7%	90.0	25.1	4.101
ALLD.L	-24.2	-10.2	40.0%	283	-0.09%	0.43%	46.0%	43.6	18.3	4.208
AHM.L	-4.1	-1.3	48.4%	304	-0.01%	0.33%	49.5%	68.2	21.8	3.601
AVZ.L	-167.5	-65.8	0.3%	306	-0.55%	0.38%	47.0%	11.3	4.4	3.943
AV.L	51.2	20.5	78.3%	281	0.18%	0.44%	49.7%	9.1	3.6	11.319
BA.L	44.5	15.3	71.1%	350	0.13%	0.43%	51.1%	15.4	5.3	5.225
BARC.L	227.4	86.9	99.9%	261	0.87%	0.52%	52.4%	57.8	22.1	30.126
BATS.L	39.2	12.4	65.7%	270	0.15%	0.49%	52.6%	92.3	29.1	13.626
BOC.L	-85.2	-44.4	2.8%	217	-0.39%	0.41%	47.7%	18.1	9.4	4.297
BOOT.L	79.6	37.4	93.0%	235	0.34%	0.50%	51.3%	-34.9	-16.4	5.297
BP.L	8.2	4.1	53.2%	322	0.03%	0.43%	50.5%	47.0	23.5	94.816
BT.L	292.6	95.0	100.0%	251	1.17%	0.53%	55.2%	-32.1	-10.4	15.954
BNZL.L	254.3	167.8	100.0%	283	0.90%	0.49%	59.8%	73.4	48.1	2.107
CBRY.L	16.0	7.9	58.6%	245	0.07%	0.45%	49.4%	14.9	7.3	7.857
CW.L	-11.1	-3.2	49.4%	177	-0.06%	0.55%	51.2%	-28.2	-8.1	2.836
DGE.L	136.6	59.0	98.4%	279	0.49%	0.47%	51.9%	33.9	14.6	21.02
DXNS.L	-21.9	-6.5	42.4%	265	-0.08%	0.43%	51.7%	63.7	18.9	2.682
DMGOa.L	70.3	25.7	82.5%	230	0.31%	0.40%	51.2%	22.5	8.2	2.127
EMA.L	171.7	63.0	98.0%	277	0.62%	0.39%	51.9%	-33.0	-12.1	2.191
EXL.L	133.0	57.6	96.7%	295	0.45%	0.40%	55.1%	1.0	0.4	2.011
FRCL.L	23.7	23.2	66.5%	160	0.15%	0.48%	52.2%	17.7	17.3	1.739
GKN.L	21.6	8.8	63.0%	307	0.07%	0.43%	48.9%	8.2	3.3	1.923
GSK.L	40.0	18.9	74.9%	312	0.13%	0.49%	51.2%	18.5	8.8	72.194
GAA.L	83.7	29.2	84.9%	315	0.27%	0.42%	52.2%	-8.5	-3.0	2.837
GUS.L	-24.5	-8.8	37.1%	296	-0.08%	0.48%	49.7%	-10.9	-3.9	6.916
HG.L	56.4	19.1	75.9%	320	0.18%	0.47%	49.2%	6.6	2.2	3.068
HNS.L	146.6	59.3	98.0%	261	0.56%	0.46%	52.0%	63.9	25.8	3.03
ICI.L	137.8	51.7	96.6%	304	0.45%	0.49%	53.0%	-38.8	-14.6	2.311
JMAT.L	3.6	1.4	50.7%	348	0.01%	0.34%	51.5%	110.2	43.9	2.144
Average all 58	58.1	23.2	69.3%	284.4	0.20%	0.44%	50.8%	22.5	9.0	9.7

Table 3.9: Support Vector Regression Results out-of-sample data 30-58

Support Vector Regression Results. Out of sample data 803 Days (stocks30-58)										
Stock Symbol	FinalEquity in %	Mod. Sharpe R	Percentage of random below	Number of trades	Average Trade %	% Trade Acuracy	% Daily Accuracy	Buy&Hold FinalEquity %	Buy&Hold Mod. Sharpe	MktCap
KGF.L	93.5	35.1	89.8%	293	0.32%	0.48%	50.5%	-30.5	-11.5	6.335
LGEN.L	71.0	28.0	84.8%	356	0.20%	0.43%	51.3%	6.6	2.6	6.47
LAND.L	41.0	28.9	84.6%	326	0.13%	0.43%	50.2%	-1.0	-0.7	3.981
MKS.L	101.9	38.3	94.7%	256	0.40%	0.51%	52.7%	-53.7	-20.2	6.913
MRW.L	-47.8	-21.0	18.8%	341	-0.14%	0.33%	48.2%	62.7	27.5	3.289
PFG.L	30.9	13.4	68.8%	304	0.10%	0.35%	49.7%	-13.3	-5.8	1.624
PRU.L	73.5	29.1	86.9%	214	0.34%	0.49%	51.8%	-1.0	-0.4	8.761
PSON.L	-23.6	-7.8	37.4%	270	-0.09%	0.47%	49.2%	19.3	6.3	4.954
RB.L	37.5	14.6	69.5%	290	0.13%	0.40%	51.0%	18.2	7.1	8.399
RBS.L	117.9	43.0	89.8%	278	0.42%	0.42%	51.1%	85.5	31.2	45.715
REL.L	185.5	70.5	99.2%	297	0.62%	0.47%	51.0%	50.4	19.1	6.135
RTO.L	-53.4	-19.2	26.0%	256	-0.21%	0.38%	48.9%	-9.1	-3.3	3.843
REX.L	24.5	9.6	60.7%	306	0.08%	0.36%	50.8%	79.3	31.0	1.877
RSA.L	-10.0	-3.6	46.0%	312	-0.03%	0.41%	47.7%	-12.9	-4.6	1.994
RTR.L	286.2	77.7	99.6%	288	0.99%	0.49%	54.5%	67.3	18.2	3.595
SN.L	18.9	8.8	75.7%	353	0.05%	0.40%	49.7%	101.9	47.3	3.775
SFW.L	-25.5	-11.8	32.2%	329	-0.08%	0.40%	48.9%	22.4	10.3	2.998
SBRY.L	10.0	4.2	55.5%	325	0.03%	0.40%	50.5%	-18.7	-7.9	5.365
SDR.L	-86.5	-27.8	11.5%	253	-0.34%	0.43%	49.3%	-10.0	-3.2	1.602
SCTN.L	33.7	14.8	74.2%	263	0.13%	0.43%	49.2%	-17.6	-7.7	3.422
SHEL.L	106.9	50.6	96.1%	267	0.40%	0.51%	53.0%	50.8	24.1	38.332
SMIN.L	137.8	59.7	98.1%	302	0.46%	0.43%	52.5%	14.3	6.2	3.939
STAN.L	45.4	14.9	69.2%	252	0.18%	0.42%	50.8%	69.8	22.9	9.903
TOMK.L	51.0	20.7	77.9%	281	0.18%	0.47%	48.9%	-27.8	-11.3	2.033
TSCO.L	42.4	19.5	72.8%	301	0.14%	0.47%	49.9%	51.3	23.5	15.191
ULVR.L	68.4	30.3	84.7%	292	0.23%	0.41%	49.3%	2.1	1.0	14.979
WTB.L	-71.1	-35.4	13.8%	185	-0.38%	0.42%	47.1%	-30.7	-15.3	2.25
WOS.L	159.9	62.8	97.9%	301	0.53%	0.38%	52.3%	48.7	19.1	4.201
WPP.L	139.3	43.8	93.9%	365	0.38%	0.39%	51.2%	78.7	24.7	6.659
Average all 58	58.1	23.2	69.3%	284.4	0.20%	0.44%	50.8%	22.5	9.0	9.7

Table 3.10: Support Vector Regression Paramcters

C 1000 EPSILON 0.01 TYPE RADIAL GAMMA .01	C 100000 EPSILON 0.001 TYPE RADIAL GAMMA .01	C 100 EPSILON 0.001 TYPE RADIAL GAMMA .01
C 1000 EPSILON 0.001 TYPE RADIAL GAMMA .001	C 100 EPSILON 0.01 TYPE RADIAL GAMMA .001	C 100000 EPSILON 0.001 TYPE RADIAL GAMMA .001
C 1000 EPSILON 0.01 TYPE RADIAL GAMMA .01	C 10 EPSILON 0.01 TYPE RADIAL GAMMA .01	C 1000 EPSILON 0.0001 TYPE RADIAL GAMMA .001
C 100 EPSILON 0.01 TYPE POLYNOMIAL DEGREE 2	C 1000 EPSILON 0.01 TYPE POLYNOMIAL DEGREE 2	C 1000 EPSILON 0.01 TYPE POLYNOMIAL DEGREE 1
C 50 EPSILON 0.01 TYPE POLYNOMIAL DEGREE 2	C 1000 EPSILON 0.01 TYPE POLYNOMIAL DEGREE 3	C 10000 EPSILON 0.01 TYPE POLYNOMIAL DEGREE 1
C 1000 EPSILON 0.001 TYPE POLYNOMIAL DEGREE 1	C 1000000 EPSILON 0.0001 TYPE NEURAL A 0.0001 B 0.0000001	C 1000 EPSILON 0.01 TYPE NEURAL A 0.0001 B 0.0000001
C 100000 EPSILON 0.001 TYPE NEURAL A 0.0001 B 0.0000001	C 1000000 EPSILON 0.01 TYPE NEURAL A 0.0001 B 0.0000001	C 10000000 EPSILON 0.01 TYPE NEURAL A 0.0001 B 0.0000001

Table 3.11: KNN Parameter values.

Set	Start	Lookback	NumKnn
1	200	4500	50
2	200	4500	10
3	200	2500	800
4	200	2500	50
5	200	1000	250
6	200	1000	50
7	200	500	250
8	200	500	100
9	200	500	10
10	200	300	250
11	200	300	50
12	200	100	10
13	200	100	4

Chapter 4

The more the merrier? Forecasting FTSE 100 stocks with a random subspace ensemble of 62500 models.

This chapter investigates the application of linear ensemble models to forecast and trade 65 component stocks within the FTSE 100, using daily data over the years 1991-2006. The primary ensemble consists of 62500 component models built using the random subspace method in which randomly sampled subsets of the feature set are used to estimate each linear regression model with the final result combined via a majority vote. The performance is compared to a number of benchmarks including a single AR model and it is found that this ensemble methodology improves the overall results both in terms of consistency across time periods and economic significance. It is also found that model selection or thinning improves performance further.

4.1 Introduction

The aim of this study is to investigate the application of linear ensemble models to forecast and trade 65 component stocks within the FTSE 100, using daily data over the years 1991-2006. Specifically, we are interested in what benefits in performance if any are accrued by using ensemble models - models that combine the forecasts of a number of individual component models - over single model specifications. We do this by comparing the performance of a single linear AR model and three ensembles with differing specifications.

While single model least-squares linear prediction of closing daily stock returns rarely produces remarkable results, even if markets are not strictly random walks[1], we might well enquire where the limits of predictability lie. Improvement beyond that of single model classical least-squares is suggested by the literature; firstly, the common approach of using just a single model results in additional information in alternatively parameterised models not

[1] The Efficient Markets Hypothesis can still hold even if markets are not random walks (Lucas 1978),(LeRoy 1973).

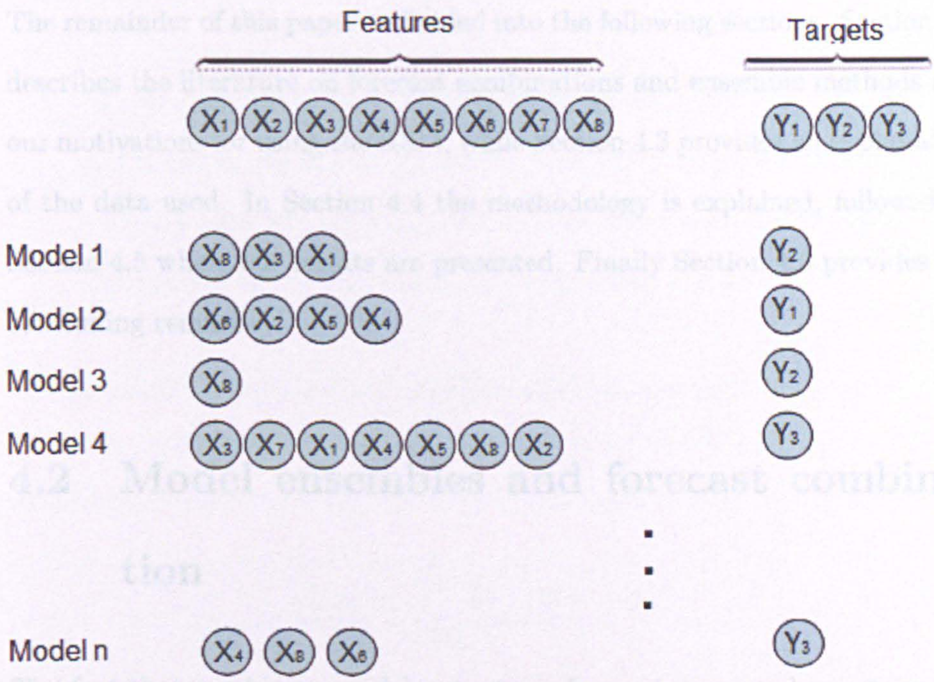
being utilized, an issue ensemble models may be able to exploit. Secondly, augmenting the feature space by including frequently neglected open, high and low price data over and above the close may increase the discriminatory power of a model. Thirdly, research into nonlinearity is suggestive of nonlinear structure within financial times-series data, which might be implicitly approximated by combinations of linear models that specialize on local areas of the problem space, thus producing a final nonlinear ensemble function that is irreproducible by its the component linear models (Opitz & Maclin 1999)(Zenobi & Cunningham 2001).

The approach taken involves estimating four models, a single AR(p) in addition to three ensemble models to forecast the returns of component stocks within the FTSE 100 index on a daily basis.

1. AR(p) model.
2. Predetermined Ensemble (PE) - uses a restricted information set.
3. Primary Random Subspace Ensemble (RSM1) - constructed using the unrestricted information set.
4. Secondary Random Subspace Ensemble (RSM2) - built using the top 2% of RSM1 models.

The three ensemble models can be differentiated both by their information set and method of construction. The *Predetermined Ensemble* (PE) is restricted

Figure 4.1: Random Subspace Method



to using a subset of the information superset and is mainly for comparative purposes. The remaining two ensembles are less restricted and are given access to the full information set which includes daily price open, high, low and close data and use a different method of construction namely, the random subspace method (RSM) of Ho (1998), whereby the component models are constructed by estimating them on pseudo-randomly selected subsets of the feature set. The primary RSM model (RMS1) consists of 62500 models while the secondary RSM model (RSM2) ensemble is built by conducting a model selection procedure on the former, that reduces the number of component models to 1250.

The remainder of this paper is divided into the following sections. Section 4.2 describes the literature on forecast combinations and ensemble methods and our motivations for using the RSM, while Section 4.3 provides an explanation of the data used. In Section 4.4 the methodology is explained, followed by Section 4.5 where our results are presented. Finally Section 4.6 provides our concluding remarks.

4.2 Model ensembles and forecast combination

The fact that combining model outputs or forecasts can, under certain conditions, increase accuracy has been exploited in many fields from meteorology to politics and has a long history. Stigler (1973) writes that Laplace suggested combining models as far back as 1818 (La Place 1820). Cunningham & Zenobi (2001) write of the Codorcet Jury Theorem (Condorcet NC de 1785), inferred by Black (1958), which predates the French Revolution:

> *"If each voter has a probability p of being correct and the probability of a majority of voters being correct is M, then* $p > 0.5$ *implies* $M > p$*. In the limit, M approaches 1, for all* $p > 0.5$*, as the number of voters approaches infinity"*

This is further explicated with an example pertaining to classification by Dietterich (2002),

> "if the error rates of L hypotheses h_l are all equal to $p < 1/2$ and if the errors are independent, then the probability that the majority vote will be wrong will be the area under the binomial distribution where more than $L/2$ hypothesis are wrong."

The probability that the final ensemble will make an error is given by (4.1).

$$P_{error} = \sum_{i=L/2}^{L} \binom{L}{i} p^i (1-p)^{L-i} \tag{4.1}$$

So given a dichotomic classification task with H hypotheses that have a *greater than 50%* classification accuracy, an ensemble using majority voting will have an accuracy that is greater than any one component model *if the errors of the component models are uncorrelated.*

In the forecasting literature Barnard (1963) showed that averaging outperformed individual forecasts. Bates & Granger (1969) developed methods for combining forecasts using linear weighted averaging. More recent work by Breiman (1996), Freund & Schapire (1996), Wolpert (1992) and Zhang, Mesirov & Waltz (1992) has extended these ideas by looking at more complex combinatorial methods, though this is somewhat of a double-edge sword; potential improvements in accuracy over simple methods often results in increased complexity which can be concomitant with the problem of overfitting.

For problems for which certain strict assumptions hold true model combination will not be beneficial. Hendry & Clements (2004) state that aggregating forecasts is not beneficial when using the correct conditional expectation in a weakly stationary process, but that this is an unlikely scenario found in practice and that mis-specification, mis-estimation or non-stationarities are behind any improvement from model combination. In a similar vein Neftci (1991) makes the point that forecasts based on Wiener-Kolmogorov prediction theory will dominate in a mean square error sense for a linear stochastic process, hence model combination would not improve on this.

Restricting choice to a single best model results in information contained within discarded models not being utilised. Additionally, when making the final choice the researcher is frequently faced with many statistically indistinguishable models without enough evidence to prefer one over another. A choice of best single model then has to made which often results in a subjective decision. Moreover, even if the model selection criteria points to one best model it is far from certain this is optimal. The discriminatory power of widely used model selection criteria such as AIC and BIC is reduced in environments with low signal to noise ratios DellÀquila & Ronchetti (2006).

Like bagging and boosting the Random Subspace Method of Ho (1998) also works by modifying the training data but it does so in feature space. Given a training set of N feature set variables, a learning machine is repeatedly applied to randomly selected feature subsets of size $M < N$. Model outputs are then aggregated via majority voting to determine the final ensemble output. Although Boosting has proven to be a very effective combining algorithm it has a tendency to overfit the data in the presence of noise (Opitz & Maclin 1999) which makes it less appropriate in the context of financial market data where the signal to noise ratio is low.

The RSM is derived from the method of Stochastic Discrimination(SD) introduced by Kleinberg (1990),Kleinberg (2000). Its effectiveness is shown in Kleinberg (2000) where it outperforms bagging and boosting on a large number of benchmark problems. It has been shown to outperform single models in a number of papers; (Chawla & Bowyer 2005), (Zhao et al. 2005), (Bertoni et al. 2005),(Rooney et al. 2004). Skurichina & Duin (2002) compared RSM, bagging and boosting using regression and nearest neighbours models on 15 data sets selected from the WEKA (Witten & Frank 1999) repository and found RSM to be most effective.

Ho (2000) found that RSM exhibited better performance when the discriminatory power of the input space is distributed evenly over many features. Skurichina & Duin (2002) confirmed this result and further noted that RSM was more effective in this case than when the discriminatory power is condensed to a few features. This is likely the case with this application in that it is difficult to argue that some specific lags contain all the relevant information at the expense of others, a situation complicated by the time varying nature of feature discriminatory power. Lastly, due to the sheer size of the feature space in this application it isn't feasible to use bagging or boosting using all features, nor is it easy to conduct the NP hard problem of multivariate feature selection (Amaldi & Kann 1998), (Guyon & Elisseeff 2003).

The use of the random subspace ensemble method with linear regression as its base learner is motivated by a number of factors. Linear regression models are fairly easy to interpret as their properties are based on firm theoretical foundations. This makes it easier to analyse a single model LR benchmark against its ensemble counterpart, teasing out those aspects of performance due to one over the other. Moreover, LR lends itself to the RSM; it has been shown that generative methods are most effective when the base learner is unstable to perturbations to the training set. Linear regression coefficients are relatively stable to the kinds of changes to the training data induced by algorithms such as bagging which uses subsets of training instances, resulting in reduced diversity of output predictions. However, RSM builds models based on subsets of features (individual independent variables) which induces instability in regression models.

4.3 The Data

The data used in this study consist of the daily open, high, low and closing prices of 65 UK stocks covering the years 1991 to 2006 and were contained within the FTSE 100 index as of January 2006 - see table 1. As a sizeable data history was desirable only those stocks that had price histories going as far back as 1991 - approximately 3900 days - were chosen hence the choice of 65 stocks of varying market capitalisations. This has the effect of excluding the more recent entries in the index such as technology stocks and placing greater emphasis on index stalwarts, though this is unlikely to have much impact on the results. All data was obtained from Datastream. As expected, based on the Jarque-Bera statistic, normality is rejected at the 1% level for all returns and all but four stocks have a positive mean over the period of study.

As linear models are the focus of this study the Qstat p-value column shows the significance of the p-values from computing the Ljung-Box Q-statistic for high-order serial correlation for each stock up to order 15. Many of the stocks exhibit significant serial correlation in log returns with 77% significant at the 5% level..

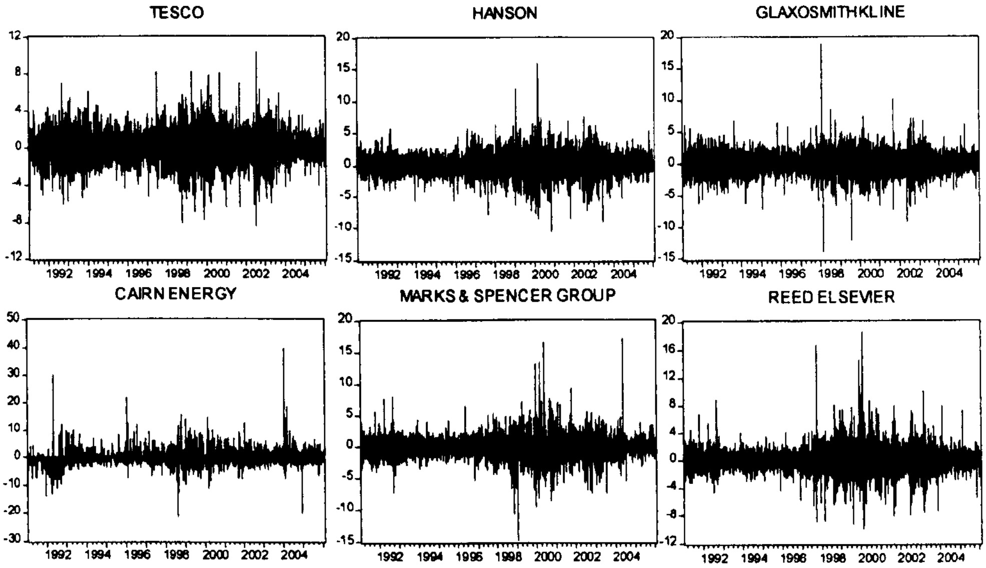
To further explore the data we include are the results of the BDS test (Brock et al. 1996). The BDS test is one the of most popular tests for nonlinear structure and tests the null that a time series is independent and identically

Table 4.1: Closing price return statistics for 65 FTSE100 component stocks from 1991 to 2006.

Stock	Mkt Cap (mil)	Mean	Std. Dev.	# Skew	# Kurt	Ljung-Box Qstat	bds	Stock	Mkt Cap (mil)	Mean	Std. Dev.	Skew.*	Kurt.*	Ljung-Box Qstat	bds
AUN	3164	0.053	1.4	1.0	22.0	**	**	LGEN	8445	0.035	2.0	0.1	6.4	**	***
AVZ	4361	0.056	2.7	0.2	8.3	***	***	MKS	10564	0.020	1.8	0.4	11.4	**	***
ANTO	4052	0.092	1.6	0.8	11.5	***		MRW	5657	0.048	1.8	-0.1	13.2	***	
ABF	6930	0.033	1.6	0.4	10.3	***	***	NXT	4154	0.118	2.1	-0.5	18.6		***
AV	18864	0.012	2.1	0.1	7.2	***	***	PSON	5644	0.026	2.1	0.2	7.7	***	***
BAA	8612	0.035	1.5	-0.6	20.1		**	PO	3880	0.025	2.0	1.0	15.8		**
BA	13484	0.030	2.6	-3.6	74.6	***		PSN	4071	0.050	1.8	0.2	9.4	***	*
BARC	43259	0.050	2.0	0.1	6.0	***	**	PRU	14622	0.028	2.1	-0.3	8.8	**	
BG	23638	0.047	1.8	0.0	5.8	***	*	RB	14691	0.035	1.7	-0.2	16.5		***
BOC	7611	0.027	1.5	0.2	15.0			REL	6550	0.029	1.8	0.7	11.5	***	**
BOOT	3446	0.018	1.5	0.0	9.0		***	RTO	2846	0.035	2.0	-0.4	12.5	**	***
BP	129540	0.039	1.6	-0.2	7.0	***		RTR	5276	0.022	2.6	-0.1	11.2		***
BAY	3718	0.026	2.4	-0.1	9.1	**	**	REX	2799	0.026	1.8	0.0	8.6		***
BATS	28547	0.053	2.0	1.4	26.4	***	*	RIO	28431	0.050	1.8	0.1	5.3	**	
BLND	6270	0.038	1.7	0.4	10.1	**	***	RR	7714	0.030	2.2	0.0	11.2	***	***
BT.A	17289	0.001	2.1	-0.1	8.4	***	**	RSA	3786	-0.021	2.5	-0.2	10.3	**	*
CW	2492	-0.019	2.7	-3.3	73.5	***	**	RBS	60525	0.064	2.0	0.0	7.0	***	***
CBRY	12056	0.033	1.5	0.2	6.2	**	**	RDSA	118437	0.036	1.6	0.0	6.4	**	**
CNE	3068	0.058	2.5	1.9	33.2	***	**	SGE	3549	0.111	2.6	0.1	11.5	**	
CPI	3098	0.113	2.2	0.4	9.7	***	*	SBRY	6132	-0.001	1.8	-0.3	8.1	**	***
DMGT	2643	0.055	1.9	0.2	11.0	**	**	SDR	3333	0.060	2.2	0.0	8.1	***	**
DGE	25243	0.021	1.7	0.4	8.8	***	**	SDRC	3333	0.070	2.5	-0.3	9.6	***	
DSGI	3190	0.041	2.4	-0.5	12.8	**	***	SCTN	4560	0.010	1.6	0.0	6.4	**	***
GSK	84210	0.032	1.8	0.2	8.7	**	**	SN	4768	0.044	1.6	0.2	7.8	**	***
GUS	9194	0.041	1.8	-0.4	14.6	***	***	SMIN	5269	0.039	1.7	-0.1	10.2		**
HMSO	3246	0.018	1.4	0.3	9.7	***	***	STAN	19540	0.086	2.2	0.1	7.0		***
HNS	5031	0.015	1.8	0.2	7.1		***	TATE	2911	0.018	1.8	-0.1	13.9	***	***
IHG	3829	0.016	2.3	0.1	11.3		**	TSCO	26514	0.036	1.7	0.1	5.1	***	***
ICI	3994	-0.001	2.2	-2.5	69.2			ULVR	17102	0.032	1.5	-0.5	10.6		***
ITV	4415	0.034	2.3	0.1	6.5	**	***	VOD	67170	0.048	2.3	0.2	5.8	***	**
JMAT	3140	0.046	1.7	0.3	8.1		***	WOS	8344	0.057	1.8	0.2	9.6	**	***
KGF	5346	0.014	1.9	-0.1	7.2	**	**	WPP	8379	0.057	3.2	0.8	20.9	**	***
LAND	8562	0.030	1.2	0.1	4.9	**	*								

distributed (iid) against a non-specified alternative. Usually an appropriate linear model is estimated on the data and the residuals, which lack linear structure by design, are then tested. If the iid null is rejected it is concluded the data contains some form of nonlinear structure. Given that much of the nonlinear structure in financial data is generated by GARCH effects

Figure 4.2: Log returns of six of 65 stocks.



(nonlinearity in the second moment) it is useful, especially if one is considering models that attempt to capture nonlinearities that are a result of other causes, to perform the BDS test on the standardised residuals of a GARCH model. In this case a GARCH (1,1) model was fitted to each stock and the resultant standardised residuals tested.

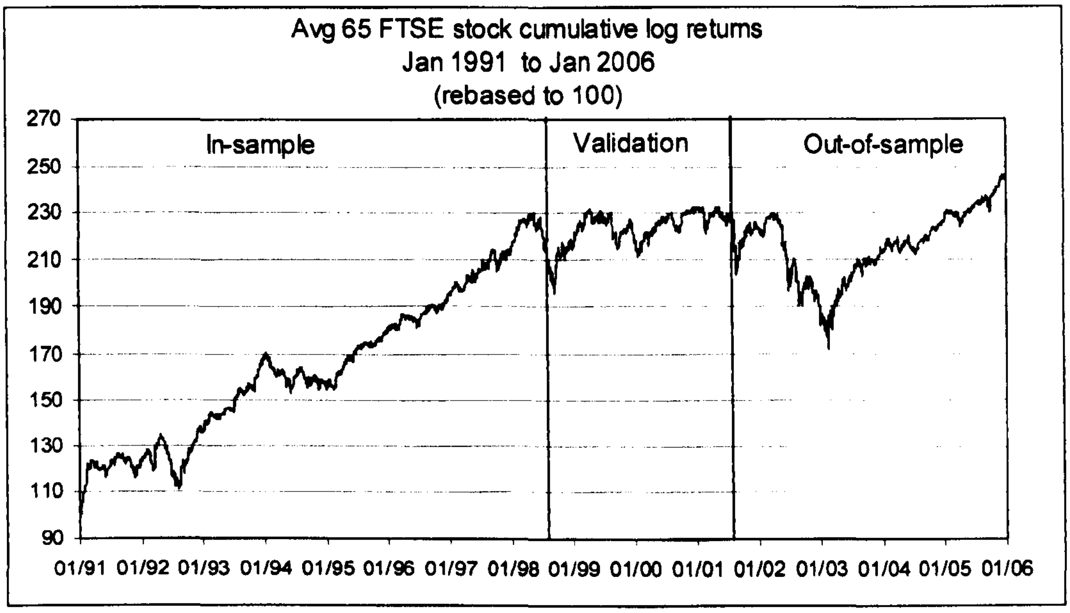
It can be seen from table 4.1 that the iid null was rejected at the 5% level for 75% (49 out of 65) of the stocks indicating that these stocks may contain nonlinear structure beyond expected GARCH effects. This conclusion is necessarily tentative being complicated by issues involving possible model miss-specification and the low power of the BDS test in detecting neglected second moment asymmetries (Brooks & Henry 1999). What is notable is that some stocks which had no significant linear structure did contain non-linearities such as, BOOTS and NEXT.

The type of forecasting model used and the manner in which the model parameters are estimated will dictate how the dataset is partitioned in order to get meaningful out-of-sample results. In the case of a linear regression forecasting model it is usually only necessary to partition the data into an in-sample and out-of-sample set, whereby the in-sample data is used to estimate the regression coefficients and the out-of-sample data to measure subsequent performance. This is due to there being only one global linear least squares fit through the data so model selection is unnecessary once the feature set has been fixed. In this study all models are autoregressive in nature and the information set is restricted to price data. Although it might be reasonable to expect additional information (volume, macro-economic variables etc) to improve model accuracy, this possibility is not investigated here. The data are divided into three different subsets, in-sample, validation and out-of-sample - see table 4.2. This is necessary as some of the ensembles require a model selection procedure based on in-sample and validation data performance to reduce problems of overfitting and datasnooping - a perennial issue in forecasting financial data. The final results are then tested out-of-sample to give an indication of potential real-time performance.

Table 4.2: Data set partitioning - 65 FTSE Stocks.

Set	Dates	Length
In-sample set	24 Jan 1991 to 31 Aug 1998	1983 days
Validation set	01 Sep 1998 to 26 Sep 2001	801 days
Out-of-sample set	26 Sep 2001 to 24 Jan 2006	1130 days

Figure 4.3: Averaged cumulative returns for all 65 stocks rebased to 100.



It is worth noting the type of market conditions that occurred during each partition as it can have a bearing on the final results. It is generally the case that a model that allows both long and short positions is less likely to perform well in a bull market when compared with the buy and hold (B&H) as often simply staying long is the best strategy. The average cumulative log returns during the in-sample period shown has a relatively high Sharpe Ratio

(SR) (Sharpe 1965) of 2.3 (see results section), clearly identifiable from figure 4.3, where the period consisted of a smoothly upward sloping B&H curve or bull market. The validation period consisted of what practitioners might refer to as a consolidated or sideways market - which is reflected in a lower SR of 0.89. The beginning of the out-of-sample period contains a large dip followed by an upwardly trending bull market and has a SR of 0.86.

4.4 Methodology

We attempt to predict and trade the daily returns of 65 FTSE 100 constituent stocks using linear regression models as our base learners or component models, which are then combined into ensembles. Each component model is estimated on and will predict the returns of each constituent stock, the results then combined into a portfolio for purposes of diversification. It is the portfolio results that are considered relevant and that are reported as they provide a more robust indication of overall model performance than do the results from individual stocks. To this end we create and compare four different models that differ both in their method of construction and by the information set used:

1. AR(p) model.

2. Predetermined Ensemble (PE) - uses a restricted information set.

3. Primary Random Subspace Ensemble (RSM1) - constructed using the unrestricted information set.

4. Secondary Random Subspace Ensemble (RSM2) - built using the top 2% of RSM1 models.

4.4.1 Model Specification.

Dependent/Target variables:

$$y_{t,t+k} = (p_{t+k}/c_t) \tag{4.2}$$

Independent/Driver variables:

$$y_{t-i,t} = (c_t/p_{t-i}) \tag{4.3}$$

where y = log returns over various horizons, c_t is the closing price on day t , and p may be an open, high, low or close from another day.

General model specification is:

$$y_{t,t+k} = a_0 + \sum_{i=1}^{15} a_i d_i y_{t-i,t} + u_{t,t+k}. \tag{4.4}$$

where a_i are estimated regression coefficients, d_i the driver variables and u_t the error term.

Model 1: AR(5)

Autoregressive model AR(5) with a fixed 500 day estimation window, rolled (re-estimated) daily. Model 1 is included as a benchmark and is typical of the form of model applied to this class of forecasting problem. The use of 5 lags is to capture possible weekly effects within the data.

$p =$ closing price c,

$k = 1, d = [1111100\ldots0]$

estimated on data from $t - 499, t$

Model 2: Predetermined Ensemble (PE)

The PE model serves as a benchmark with which to compare the more complex RSM1 and RSM2 ensembles. Trades are based on the forecasts of the following 1170 regression models:

$p =$ closing price c, $k = 1, 2, 3, \ldots, 9$ (9 horizons)

$d = [111000\ldots0], [111100\ldots0], [111110\ldots0]$ up to $[111111\ldots1]$

(13 driver variable sets)

estimated on data from windows $t-n, t$ where n= 100,200,300,...,1000. (10 samples). Models are rolled every 30 days.

To clarify, there are 13 different sets of independent variables, 9 dependent variables and 10 different rolling estimation windows. The PE consists of a fairly simple structure and there is no attempt to optimise performance via variable or model selection.

Clearly the PE's component models contain some redundancy across the independent variables for example, all models contain at least one to three day lagged returns. This is intentional and is in contrast to the RSM ensembles of which an important element of the methodology is the construction of (if not truly) orthogonal models, models which are by design more diverse than the PE models.

Model 3. Primary Random Subspace Ensemble (RSM1)

This ensemble uses a larger information set than the PE Model. The independent variables can consist of any members of the price information set including the open, high, low and close. Forecasts and thus trades are based on 62500 models generated using the following process:

p= randomly selected open, high, low or close. $k = 1, 2, 3, \ldots, 10$, randomly selected.

d= randomly selected 15-vector, with between 1 and 8 values =1,

estimated on data from windows $t - n, t$ where n= randomly selected in the range 110, 1000.

In all cases random sampling is from the uniform distribution and models are rolled every 30 days. It is evident from the model specification that there is a greater capacity for model diversity than previous models for example, at a minimum a model using only one independent variable is possible. This capacity also results in the creation of clearly miss-specified models, defined as any model that results in a negative overall return on the in-sample data. These models are removed - a form of real time model pruning. To clarify, miss-specified models can result when the target is randomly selected as:

$$y_{t,t+1} = log(Low_{t+1}/Close_t)$$

The 1-day close to low target return is frequently negative when compared to the 1-day close to close return (in the region of 70% negative for the former and 40% the latter). This produces negatively biased forecasts and, given that measured performance is based on close to close returns, can result in a negative overall return, even though the model may perform well on the basis of close to low returns.

Model 4: RSM2

The RSM2 ensemble comprises of the top 2% of RSM1 component models, chosen on the basis of in-sample total returns. This is a form of model skimming and is similar to pruning or thinning except that only a small percentage of models are retained rather than dispensed with. The rationale behind this method of model selection is explained in the results section.

Converting model outputs into forecasts for trade simulation.

Every component regression model built outputs a forecast (a real number) for each day in the dataset. This needs to be converted into a trading decision rule (TDR) to hold either a long (1), short (-1) or, in some cases, a flat (0) position in the underlying share. The first stage is to assign the output of each model either 1 or -1 according to the following rule:

If model output ≥ 0, then output = 1, Else output = -1.

This results in a vector of 1s and -1s for each of the 65 stocks for each day. In the case of Model 1 this is all that is needed and the model is tested by simulating holding a long or short position from t to t+1 depending on the sign of the model output. In the case of the ensembles we compare two TDRs based on thresholding the summed output of the whole ensemble. For each day the integer output of all component models is summed, which in the case of Model 2 (PE) consists of summing the output of all 1170 models.

Figure 4.3 shows this in more detail where, on one particular day for one of the 65 stocks, there are 1170 model outputs of either 1 or -1. The summed value is taken as the ensemble's forecast for that particular share and day.

Table 4.3: Hypothetical ensemble forecast for PE model.

Estimation window	k (forecast horizon)									
	1	2	3	4	5	6	7	8	9	
100	13	5	5	-6	-6	-3	1	3	3	
200	-3	-3	-7	5	13	3	5	-5	5	
300	-3	5	-3	5	1	-6	7	-3	6	
400	-3	7	-3	1	7	1	-11	1	-1	
500	-3	-5	-3	5	-5	1	-5	-3	-7	
600	9	-7	-3	13	7	9	-6	5	-9	
700	-9	11	12	-1	-13	-11	1	8	5	
800	-5	6	1	13	5	-11	1	3	-7	
900	-7	9	-3	5	11	5	-5	13	1	
1000	-5	-5	9	11	3	5	-9	-11	-5	
								Sum=		60

In the example shown pertaining to the PE model the forecast is +60, though it can range between 1170 and -1170. If we assume a higher value is reflective of greater confidence then it might be the case that only taking a position when the forecast is above or below a certain threshold produces better performance.

Denote by $F_1, \ldots, F_L$ the forecast (1,-1) of each component model for one stock on day_t, where L is the number of component models in the ensemble:

Trading Decision Rule 1 (TDR 1):

$$If \sum_{i=1}^{L} F_i > long\_thres, position_t = 1,$$

$$Else \quad If \sum_{i=1}^{L} F_i < short\_thres, position_t = -1,$$

$$Else \quad position_t = 0.$$

Trading Decision Rule 2 (TDR 2):

$$If \sum_{i=1}^{L} F_i > long\_thres, position_t = 1,$$

$$Else \quad If \sum_{i=1}^{L} F_i < short\_thres, position_t = -1,$$

$$Else \quad position_t = position_{t-1}.$$

When using a threshold of 0 for both rules TRD1=TDR2. For thresholds greater than zero (the threshold for short positions is simply opposite in sign to the long threshold) TDR1 results in either long, short or flat positions i.e., it is not always in the market. On the other hand TDR2 results in either long or short positions and is always in the market, as once a long (short) position is taken it is only switched to short (long) when the model output is less (greater) than the chosen short (long) threshold. The rationale behind TDR2 is to reduce transaction costs by reducing trading frequency, based on the notion that there is no strong reason to switch positions until the model is strongly confident of a move in the opposite direction.

4.5 Results

In this section the results of the four models are presented across the three data partitions for the portfolio of 65 stocks. Note that the in-sample period shown is based on the last 983 days of the original in-sample data of 1983 days and is that period beyond the initial maximum historical window length of 1000 days used to estimate the component models. This renders the in-sample results shown effectively ex-ante and allows more meaningful comparisons with the validation and out-of-sample results.

The reported SRs do not include the risk free rate in their calculation as the resulting simultaneous long and short positions will largely offset each other, though not entirely, as the portfolio is not 100% market neutral. As a result the need to choose an appropriate risk free rate is dispensed with.

4.5.1 Model 1: AR(5)

As there is only one AR model per stock the threshold value is zero and TDR1 is examined[2]. A number of performance statistics which relate to

[2] In this case TDR1=TDR2

over 80,000 simulated trades over the three separate in-sample, validation and out-of-sample periods are shown in table 4.4, including the equivalent buy and hold SR figure for each period. Figure 4.4 depicts the resulting cumulative return curves with transaction costs of 0% and 0.5% included, in addition to the equally weighted portfolio average [3]. What is immediately clear is that without transaction costs the model looks promising, exhibiting a smooth upwardly sloping cumulative return curve and SRs of between 2.2 and 5. Once transaction costs of 0.5% are taken into consideration model performance significantly degrades. The out-of-sample performance is markedly worse than that over the other data partitions, most notably the average trade figure which almost halves. Overall the results are not surprising in terms of market efficiency and are indeed what would be expected, highlighting the point that although there exists a certain amount of predictability within the price data, it is not enough to beat the assumed transaction costs using this model and information set.

When interpreting these results over the periods analysed it's important to consider the transaction cost environment over time. The in-sample SR with no transaction costs is the largest at 5.8 but then declines steadily across the data, dropping to 2.2 in the out-of-sample period. A large component of this decline will likely be the result of steadily decreasing transaction costs since the early 1990s. Whether it is also a result of the market becoming more efficient to linear modelling of closing price data is unclear, as linear

[3]This represents the return achieved via buy and hold.

Table 4.4: Model 1. AR(5) results.

In-sample

Number of models per stock = 1 Number of days = 983				Sharpe Ratio given level of transaction costs TDR1					Avg Final Equity per stock given level of trans costs					Daily Ret	Daily Ret
Thres	Avg Trade % TDR1	Avg Trade in days	Num Trades TDR1	0.0%	0.25%	0.50%	0.75%	1.00%	0.0%	0.25%	0.50%	0.75%	1.00%	0.25%	0.50%
0	0.29	2.5	25981	5.8	0.7	-4.3	-9.0	-13.3	117%	17%	-83%	-183%	-283%	0.017%	-0.084%
Buy and hold Sharpe Ratio = 2.3															

Validation

Number of models per stock = 1 Number of days = 801				Sharpe Ratio given level of transaction costs TDR1					Avg Final Equity per stock given level of trans costs					Daily Ret	Daily Ret
Thres	Avg Trade % TDR1	Avg Trade in days	Num Trades TDR1	0.0%	0.25%	0.50%	0.75%	1.00%	0.0%	0.25%	0.50%	0.75%	1.00%	0.25%	0.50%
0	0.29	2.2	23188	4.5	0.7	-3.1	-6.8	-10.4	103%	14%	-75%	-165%	-254%	0.02%	-0.09%
Buy and hold Sharpe Ratio = 0.89															

Out-of-sample

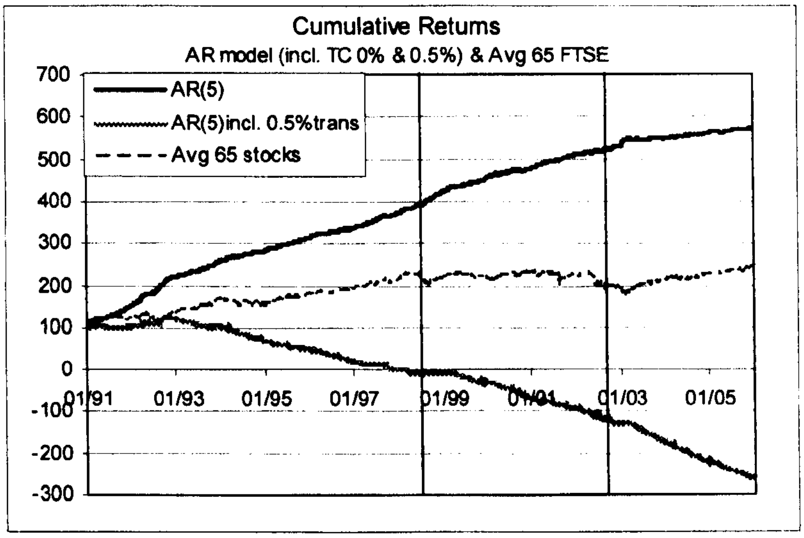
Number of models per stock = 1 Number of days = 1130				Sharpe Ratio given level of transaction costs TDR1					Avg Final Equity per stock given level of trans costs					Daily Ret	Daily Ret
Thres	Avg Trade % TDR1	Avg Trade in days	Num Trades TDR1	0.0%	0.25%	0.50%	0.75%	1.00%	0.0%	0.25%	0.50%	0.75%	1.00%	0.25%	0.50%
0	0.15	2.2	32609	2.2	-1.5	-5.2	-8.7	-11.9	74%	-51%	-177%	-302%	-428%	-0.05%	-0.16%
Buy and hold Sharpe Ratio = 0.86															

regression was, after all, clearly available at the time.

The literature confirms a steady decline in trading costs; Gemmill (1996) reports that spreads for block trades dropped from 0.78% in 1988 to 0.43% in 1992. Levin & Wright (2004) found the average spread for FTSE 100 stocks to be 0.55% during '94 to '95. (Gemmill 1998) reports a 0.39% spread for large companies before the introduction of SETS[4] in October 1997 and 0.32% shortly afterwards. The London Stock Exchange reported average spreads of 0.2% in their guide to trading the FTSE. We calculated an average spread

[4]The London Stock Exchange's trading service for UK blue chip securities

Figure 4.4: Model 1 AR(5) cumulative returns.



of 0.15% in 2006 for FTSE 100 stocks from SETS prices. Our results account for transaction costs between 0% and 1%, though we mainly assume 0.25% and 0.5% which probably ranges from a minimum to an amount reasonably attainable by institutional traders.Of course transaction costs also comprise of other components such as financing costs, estimated at 0.16% for institutional investors (Naik & Yadav 1999*b*).

On another note it's important when making any inferences regarding economic significance that it was actually possible to trade at, or close to, the price reported. For example, numerous authors have found the bid-ask spread to be at its widest during the opening, and then declining throughout the rest of the day (Levin & Wright 1999) (Naik & Yadav 1999*a*). This suggests that the opening price is less reliable in terms of simulated trading than the close. To investigate this we re-estimated the PE model using open to open prices and found the average trade figure increased by around 20%. We interpret this as being more likely reflective of the aforementioned problem of trade price unattainability than an opportunity to improve results.

The SRs reported don't account for the risk free rate as some positions offset each other so are self-financing. Additionally, traders are rarely required to trade at the observed spread. Traders face what is termed the effective spread which is generally reported to be lower than the average spread. Moreover Taylor (2002) showed that it was possible to reduce the effective spread by predicting and trading when the spread was at a minimum. As a result the SRs are only slightly inflated, if at all, when compared to exogenous research that includes the risk free rate.

4.5.2 Model 2. Predetermined ensemble.

These results are shown in Tables 4.5 to 4.7 and figure 4.5. The tables show the results of applying different thresholds to the output of the 1170 model ensemble for both TDR1 and TDR2, in addition to the effect of applying different levels of transaction costs to TDR2. In figure 4.5 the cumulative return curves for four different combinations of thresholds and transaction costs can be seen, in addition to the equivalent buy and hold.

The first two columns show the percentage and absolute value of the threshold used (the threshold for short positions is merely of opposite sign). The third and fourth column show the average percentage return of each trade given the threshold level for TDR2 and TDR1 and columns five and six show the total number of trades across all stocks for TDR2 and TDR1 respectively. The next set of five columns show the SRs based on five different levels of transaction costs for each threshold level for TDR2 only. Given that TDR2 results in higher SRs in-sample once transaction costs are taken into account this is the trading decision rule chosen for simulated trades. The equivalent results for TRD1 can be largely inferred given the average trade and transaction costs. The next seven columns show figures for the average total percentage return for each stock over the given data partition for each level of transaction costs. The last two columns show the average daily percentage return per stock assuming transaction costs of 0.25% and 0.5% which, in addition to the SR, allows comparison across different time periods. Finally, at the bottom of each table, the SR for buying and holding an equally weighted portfolio of all stocks is included for comparison.

As expected, when the threshold increases the number of trades for both TDRs decreases, though more so for TDR2, for which positions are only switched if a forecast (ensemble output) follows in the opposite direction beyond the relevant threshold value. The percentage average trade increases in tandem with the threshold for TDR2 though not for TDR1. Note that with regards to TRD1, although the average percentage return per trade isn't rising with the threshold it may be that the average *daily* percentage return *is* rising.

Table 4.5: Predetermined ensemble results over the in-sample data period.

Number of models per stock = 1170 Number of days = 983						Sharpe Ratio given level of transaction costs TDR2					Avg Final Equityper stock given level of trans costs					Daily Ret	Daily Ret
Thres	Thres Val	Avg Trade % TDR2	Avg Trade % TDR1	Num Trades TDR2	Num Trades TDR1	0.0%	0.25%	0.50%	0.75%	1.00%	0.0%	0.25%	0.50%	0.75%	1.00%	0.25%	0.50%
0	0	0.54	0.54	11081	11081	3.8	2.1	0.3	-1.5	-3.3	92%	49%	7%	-36%	-79%	0.05%	0.01%
20%	234	0.83	0.54	7303	10975	3.8	2.7	1.5	0.4	-0.8	93%	65%	37%	9%	-19%	0.066%	0.04%
40%	468	1.13	0.53	4923	10676	3.5	2.7	2.0	1.2	0.4	86%	67%	48%	29%	10%	0.068%	0.05%
60%	702	1.65	0.51	3141	9919	3.5	3.0	2.4	1.9	1.4	80%	68%	56%	43%	31%	0.069%	0.06%
80%	936	2.98	0.45	1483	8145	3.0	2.8	2.6	2.3	2.1	68%	62%	57%	51%	45%	0.063%	0.06%
99%	1169	21.22	0.58	146	1652	2.3	2.3	2.3	2.3	2.2	48%	47%	47%	46%	45%	0.048%	0.05%

Buy and hold Sharpe Ratio = 2.3

Table 4.6: Predetermined ensemble results over the validation data period.

Number of models per stock = 1170 Number of days = 801						Sharpe Ratio given level of transaction costs TDR2					Avg Final Equityper stock given level of trans costs					Daily Ret	Daily Ret
Thres %	Thres Val	Avg Trade % TDR2	Avg Trade % TDR1	Num Trades TDR2	Num Trades TDR1	0.0%	0.25%	0.50%	0.75%	1.00%	0.0%	0.25%	0.50%	0.75%	1.00%	0.25%	0.50%
0	0	0.49	0.49	11892	11892	4.3	2.3	0.2	-1.9	-4.0	90%	44%	-2%	-47%	-93%	0.06%	0.00%
20%	234	0.70	0.48	8437	11669	4.0	2.7	1.4	0.1	-1.4	90%	58%	25%	-7%	-40%	0.072%	0.03%
40%	468	0.85	0.48	5958	10923	3.5	2.6	1.7	0.7	-0.3	77%	55%	32%	9%	-14%	0.068%	0.04%
60%	702	0.94	0.47	3923	9380	2.7	2.1	1.4	0.8	0.1	57%	42%	27%	11%	-4%	0.052%	0.03%
80%	936	1.48	0.52	2084	6819	2.2	1.8	1.5	1.1	0.8	47%	39%	31%	23%	15%	0.049%	0.04%
99%	1169	0.66	0.70	199	1008	0.1	0.0	0.0	0.0	-0.1	2%	1%	0%	0%	-1%	0.002%	0.00%

Buy and hold Sharpe Ratio = 0.89

The average trade figure of 0.29% in-sample for the AR(5) model rises to 0.54% at a threshold of zero for the PE model, with an approximate halving of the number of trades to just over 37,000, which is due in a large part to the inclusion of lower frequency targets (1 to 9 days ahead as opposed to 1

Table 4.7: Predetermined ensemble results over the out-of-sample data period.

Number of models per stock = 1170 Number of days = 1130						Sharpe Ratio given level of transaction costs TDR2					Avg Final Equityper stock given level of trans costs					Daily Ret	Daily Ret
Thres %	Thres Val	Avg Trade % TDR2	Avg Trade % TDR1	Num Trades TDR2	Num Trades TDR1	0.0%	0.25%	0.50%	0.75%	1.00%	0.0%	0.25%	0.50%	0.75%	1.00%	0.25%	0.50%
0	0	0.22	0.22	14987	14987	1.5	-0.2	-1.9	-3.5	-5.0	51%	-6%	-64%	-122%	-179%	-0.01%	-0.06%
20%	234	0.36	0.21	9105	14736	1.5	0.4	-0.6	-1.6	-2.6	50%	15%	-20%	-55%	-90%	0.014%	-0.02%
40%	468	0.50	0.21	5468	13381	1.2	0.6	0.0	-0.6	-1.2	42%	21%	0%	-21%	-42%	0.019%	0.00%
60%	702	0.71	0.20	2865	10896	0.9	0.6	0.3	0.0	-0.3	31%	20%	9%	-2%	-13%	0.018%	0.01%
80%	936	1.00	0.21	1177	7271	0.5	0.4	0.2	0.1	0.0	18%	14%	9%	5%	0%	0.012%	0.01%
99%	1169	-0.16	0.26	89	648	0.0	0.0	0.0	0.0	-0.1	0%	-1%	-1%	-1%	-2%	0.000%	0.00%

Buy and hold Sharpe Ratio = 0.86

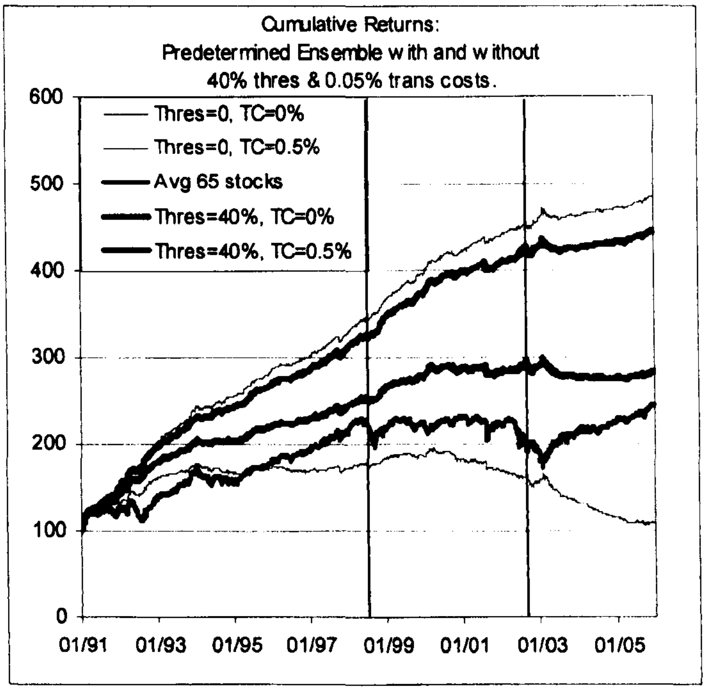
day for AR(5)) and inputs in the models.

Initially, for some threshold levels the results look promising. Assuming a threshold level of 80% and transaction costs of 0.5% the model exhibits a SR of 2.6 in-sample and 1.5 during the validation period. However, the performance degradation continues out-of-sample, where it drops to 0.2 and is representative of a general downward trend in all performance statistics when moving from validation to out-of-sample. Given that all the results shown are effectively *ex-ante*[5] the overall decline in performance is less likely to be explained by model overfitting or datasnooping and again, like the AR(5) model, results are more likely to be explained by the steadily declining rate of transaction costs. Note also that at a threshold of 80% the number of trades in-sample for TDR1 is only reduced to 73% of the value for a threshold of zero. This suggests there is low diversity in the ensemble as the individual

[5]The in-sample results are reported for the 983 days after the maximum 1000 days estimation period.

model's forecasts are correlated.

Figure 4.5: Cumulative equity curves for the PE model.



In figure 4.5 the very top and bottom curves are the results of using a threshold of zero and 0% and 0.5% transaction costs respectively. Their shape is comparable to the AR model's cumulative equity curves and similarly reflect the inability of the ensemble to beat the market once transaction costs are considered. The two grey curves above the black B&H curve depict the results of using a 40% threshold with and without 0.5% transaction costs. It is clear overall that the addition of costs is having a lesser effect due to larger average trades but still large enough to underperform the buy and hold.

4.5.3 Model 3. Random Subspace Ensemble (RSM1)

The results for Model 3 are shown in tables 4.8 to 4.10. The number of thresholds for which results are reported has been increased as much of the change occurs at the lower end of threshold values. This points to one method of gaining insight into the extent of model component diversity; compare between models the reduction in the total number of trades for a given percentage threshold. The RSM1 model was expected to have greater diversity than the previous models and this is indeed the case in that a threshold of 80% reduces the number of trades to 48% of the amount at a threshold of zero, compared to 73% for the PE model.

Table 4.8: RSM1 Ensemble results over the in-sample data period.

Number of models per stock = 62500 Number of days = 983						Sharpe Ratio given level of transaction costs TDR2					Avg Final Equityper stock given level of trans costs					Daily Ret	Daily Ret
Thres	Thres Val	Avg Trade % TDR2	Avg Trade % TDR1	Num Trades TDR2	Num Trades TDR1	0.0%	0.25%	0.50%	0.75%	1.00%	0.0%	0.25%	0.50%	0.75%	1.00%	0.25%	0.50%
0	0	0.98	0.98	4386	4386	2.7	2.1	1.5	0.9	0.2	66%	49%	32%	16%	59%	0.050%	0.033%
5%	3125	1.26	0.99	3481	4324	2.8	2.3	1.8	1.3	0.8	67%	54%	40%	27%	62%	0.055%	0.041%
10%	6250	1.50	0.99	2885	4331	2.8	2.4	2.0	1.6	1.2	66%	55%	44%	33%	62%	0.056%	0.045%
15%	9375	1.74	0.98	2454	4328	2.8	2.5	2.2	1.8	1.4	66%	56%	47%	37%	62%	0.057%	0.048%
20%	12500	2.05	1.00	2058	4203	2.9	2.6	2.3	2.0	1.7	65%	57%	49%	41%	62%	0.058%	0.050%
25%	15625	2.47	1.02	1706	4160	3.0	2.7	2.5	2.2	1.9	65%	58%	52%	45%	62%	0.059%	0.053%
30%	18750	2.87	1.01	1450	4144	3.0	2.8	2.6	2.4	2.1	64%	59%	53%	47%	62%	0.060%	0.054%
40%	25000	4.24	1.01	1006	3952	2.9	2.8	2.7	2.5	2.4	66%	62%	58%	54%	64%	0.063%	0.059%
50%	31250	6.30	1.01	663	3734	2.8	2.7	2.6	2.5	2.4	64%	62%	59%	57%	63%	0.063%	0.060%
60%	37500	9.94	0.96	400	3415	2.7	2.6	2.5	2.5	2.4	61%	60%	58%	57%	61%	0.061%	0.059%
70%	43750	19.83	0.94	205	2860	2.5	2.5	2.5	2.4	2.4	63%	62%	61%	60%	62%	0.063%	0.062%
80%	50000	63.24	0.95	66	2116	2.5	2.5	2.5	2.5	2.5	64%	64%	64%	63%	64%	0.065%	0.065%
90%	56250	1310.71	1.08	3	1182	2.3	2.3	2.3	2.3	2.3	60%	60%	60%	60%	60%	0.062%	0.062%
99%	61875	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Buy and hold Sharpe Ratio = 2.3

What is immediately of note is the average trade for TDR1 has increased from

Table 4.9: RSM1 Ensemble results over the validation data period.

Number of models per stock = 62500 Number of days = 801						Sharpe Ratio given level of transaction costs TDR2					Avg Final Equityper stock given level of trans costs					Daily Ret	Daily Ret
Thres	Thres Val	Avg Trade % TDR2	Avg Trade % TDR1	Num Trades TDR2	Num Trades TDR1	0.0%	0.25%	0.50%	0.75%	1.00%	0.0%	0.25%	0.50%	0.75%	1.00%	0.25%	0.50%
0	0	0.60	0.60	5579	5579	2.2	1.4	0.7	-0.2	-1.0	51%	30%	8%	-13%	43%	0.037%	0.010%
5%	3125	0.79	0.67	4397	5599	2.2	1.6	1.0	0.4	-0.2	53%	36%	20%	3%	47%	0.046%	0.024%
10%	6250	0.96	0.69	3635	5534	2.2	1.7	1.3	0.8	0.3	54%	40%	26%	12%	48%	0.049%	0.032%
15%	9375	1.06	0.65	2972	5328	1.9	1.6	1.2	0.8	0.4	49%	37%	26%	14%	44%	0.047%	0.032%
20%	12500	1.06	0.68	2462	5076	1.7	1.4	1.1	0.8	0.4	40%	31%	21%	12%	36%	0.038%	0.026%
25%	15625	1.19	0.64	2053	4826	1.7	1.4	1.2	0.9	0.6	37%	30%	22%	14%	34%	0.037%	0.027%
30%	18750	1.35	0.69	1691	4478	1.7	1.5	1.2	1.0	0.8	35%	29%	22%	16%	33%	0.036%	0.028%
40%	25000	1.50	0.70	1087	3786	1.2	1.1	1.0	0.8	0.7	25%	21%	17%	12%	23%	0.026%	0.021%
50%	31250	0.88	0.74	644	3039	0.8	0.7	0.6	0.5	0.4	9%	6%	4%	1%	8%	0.008%	0.005%
60%	37500	1.66	0.78	357	2425	0.7	0.7	0.6	0.6	0.5	9%	8%	6%	5%	9%	0.010%	0.008%
70%	43750	3.43	0.91	233	1709	0.8	0.7	0.7	0.7	0.7	12%	11%	11%	10%	12%	0.014%	0.013%
80%	50000	7.37	1.17	87	937	0.7	0.6	0.6	0.6	0.6	10%	10%	9%	9%	10%	0.012%	0.011%
90%	56250	133.88	1.53	11	423	0.8	0.8	0.8	0.8	0.8	23%	23%	23%	23%	23%	0.028%	0.028%
99%	61875	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Buy and hold Sharpe Ratio = 0.89

Table 4.10: RSM1 Ensemble results over the out-of-sample data period.

Number of models per stock = 62500 Number of days = 1130						Sharpe Ratio given level of transaction costs TDR2					Avg Final Equityper stock given level of trans costs					Daily Ret	Daily Ret
Thres	Thres Val	Avg Trade % TDR2	Avg Trade % TDR1	Num Trades TDR2	Num Trades TDR1	0.0%	0.25%	0.50%	0.75%	1.00%	0.0%	0.25%	0.50%	0.75%	1.00%	0.25%	0.50%
0	0	0.88	0.88	6123	6123	2.3	1.7	1.0	0.3	-0.3	83%	59%	36%	12%	73%	0.052%	0.032%
5%	3125	1.25	0.92	4397	6148	2.3	1.9	1.4	0.9	0.4	85%	68%	51%	34%	78%	0.060%	0.045%
10%	6250	1.56	0.90	3319	6025	2.1	1.8	1.4	1.1	0.7	80%	67%	54%	41%	75%	0.059%	0.048%
15%	9375	1.81	0.91	2550	5911	1.8	1.6	1.3	1.1	0.8	71%	61%	51%	41%	67%	0.054%	0.045%
20%	12500	2.23	0.91	1988	5629	1.7	1.6	1.4	1.2	1.0	68%	60%	53%	45%	65%	0.053%	0.047%
25%	15625	2.91	0.93	1524	5273	1.7	1.6	1.4	1.3	1.1	68%	62%	56%	51%	66%	0.055%	0.050%
30%	18750	3.65	0.97	1144	4881	1.6	1.5	1.4	1.3	1.2	64%	60%	55%	51%	63%	0.053%	0.049%
40%	25000	6.83	0.96	601	4122	1.6	1.5	1.5	1.4	1.4	63%	61%	59%	56%	62%	0.054%	0.052%
50%	31250	11.96	0.97	291	3406	1.4	1.4	1.4	1.3	1.3	54%	52%	51%	50%	53%	0.046%	0.045%
60%	37500	31.47	0.97	111	2558	1.4	1.4	1.3	1.3	1.3	54%	53%	53%	52%	54%	0.047%	0.047%
70%	43750	80.70	1.11	39	1531	0.9	0.9	0.9	0.9	0.9	48%	48%	48%	48%	48%	0.043%	0.043%
80%	50000	302.93	1.18	11	626	0.9	0.9	0.9	0.9	0.9	51%	51%	51%	51%	51%	0.045%	0.045%
90%	56250	3669.86	1.61	1	138	0.8	0.8	0.8	0.8	0.8	56%	56%	56%	56%	56%	0.050%	0.050%
99%	61875	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Buy and hold Sharpe Ratio = 0.86

about 0.25-.45 for the PE model to a range of about 0.65 to 1 for RSM1. Similar to the PE model the number and size of trades increases in line with the size of the threshold. The SRs also increase as the average trade increases across all data partitions. What is markedly different is that the performance of the model remains more consistent across the whole dataset than the PE model. The in-sample SRs for thresholds greater than 20% tends to equal or exceed that of the B&H (2.3) even when transaction costs of 0.5% are taken into account, which is interesting given the relatively advantage of the B&H during a bull market, as previously noted.

4.5.4 Model 4. Random Subspace Ensemble (RSM2)

Given that all constructed models are included in the final RSM1 ensemble (there is no model selection[6]) it is likely that a proportion of the component RSM1 models perform poorly enough to actually reduce the performance of the ensemble as a whole. RSM2 uses a relatively simple form of model selection in an attempt to render a higher performing ensemble than RSM1. Removing just a small proportion of the weak performing models is unlikely to change RSM1's performance as many of them will just be contributing noise, with their near random outputs simply out-voted by the higher quality models. RSM2 is built by skimming the top 2% of RSM1 models based on a measure of their in-sample performance.

[6]Apart from the initial pruning of models that result in negative in-sample returns.

There are many criteria that could be used to define top performing models but in this case we choose what is probably the simplest; those models with the highest in-sample total returns, not including transaction costs. This criteria is unlikely to be optimal if the ultimate objective is a model exhibiting a high SR net of transaction costs and will be affected by a number of factors:

1. By not including transaction costs the selection procedure is biased to those models that capture a larger component of the daily price movement which tends to be models that trade more frequently and therefore increase transaction costs.

2. A good ensemble will contain models that are not only accurate but are also diverse. This type of selection will tend to reduce diversity as those models that perform well individually in terms of total returns will tend to be correlated.

Examples of other methods include choosing those models exhibiting the highest SR after accounting for transaction costs or choosing the top 2% of models based on some portfolio optimisation method. Whatever the method chosen there will be drawbacks. For example, the latter method will tend to exclude those models that are very accurate on a short term basis but result in many transactions, even though they may actually benefit the ensemble when combined with longer term models. We choose a simple method as first a approximation and any recorded improvement might be increased by using a more sophisticated selection procedure.

The results of this skimming procedure can be seen in figures 4.6 to 4.8. What is notable is the consistency of performance of the average trade figures across all three data partitions, ranging from 0.62% to 0.66% at a threshold of zero. Except for the validation period these are lower than RSM1, something we might expect as the model selection method favoured models that traded more frequently and hence with lower average percentage trades. The SRs across all periods for transaction costs of 0.5% are in the region of 2, with figures dipping slightly in the validation set but subsequently improving out-of-sample. The in-sample results are good by construction so we would expect nothing less.

Figure 4.6: Top 1250 RSM2 model results over the in-sample data period.

Number of models per stock = 1250 Number of days = 983						Sharpe Ratio given level of transaction costs TDR2					Avg Final Equityper stock given level of trans costs					Daily Ret	Daily Ret
Thres	Thres Val	Avg Trade % TDR2	Avg Trade % TDR1	Num Trades TDR2	Num Trades TDR1	0.0%	0.25%	0.50%	0.75%	1.00%	0.0%	0.25%	0.50%	0.75%	1.00%	0.25%	0.50%
0	0	0.62	0.62	11892	11892	5.8	3.5	1.1	-1.2	-3.5	114%	68%	22%	-24%	-69%	0.069%	0.022%
20%	250	0.83	0.64	8179	11679	5.4	3.8	2.2	0.5	-1.2	104%	73%	41%	10%	-22%	0.074%	0.042%
40%	500	1.02	0.65	5716	11097	4.7	3.6	2.5	1.3	0.2	90%	68%	46%	24%	2%	0.069%	0.047%
60%	750	1.40	0.63	3792	9985	4.1	3.4	2.7	2.0	1.3	82%	67%	53%	38%	24%	0.068%	0.054%
80%	1000	2.35	0.64	1974	7528	3.4	3.0	2.7	2.4	2.1	71%	64%	56%	49%	41%	0.065%	0.057%
99%	1237	36.13	0.77	105.0	1561	2.4	2.4	2.4	2.4	2.4	58%	58%	58%	57%	57%	0.059%	0.059%

Buy and hold Sharpe Ratio = 2.3

A reasonable choice of threshold of between 40% and 60% based on in-sample and validation results yields SRs in the region of 2.5, 2 and 2.4 over the three data partitions, the consistency of which suggests the ensemble is managing to model some of the structure within the data.

Figure 4.7: Top 1250 RSM2 model results over the validation data period.

Number of models per stock = 1250 Number of days = 801						Sharpe Ratio given level of transaction costs TDR2					Avg Final Equityper stock given level of trans costs					Daily Ret	Daily Ret
Thres %	Thres Val	Avg Trade % TDR2	Avg Trade % TDR1	Num Trades TDR2	Num Trades TDR1	0.0%	0.25%	0.50%	0.75%	1.00%	0.0%	0.25%	0.50%	0.75%	1.00%	0.25%	0.50%
0	0	0.63	0.63	10341	10341	4.3	2.8	1.2	-0.6	-2.4	100%	60%	20%	-19%	-59%	0.075%	0.025%
20%	250	0.78	0.62	7242	10079	4.0	2.9	1.7	0.5	-0.7	87%	59%	31%	3%	-24%	0.074%	0.039%
40%	500	1.01	0.65	5077	9120	3.7	3.0	2.1	1.3	0.4	79%	59%	40%	20%	0%	0.074%	0.049%
60%	750	1.10	0.64	3463	7684	3.1	2.5	1.9	1.2	0.6	59%	46%	32%	19%	6%	0.057%	0.040%
80%	1000	1.31	0.63	1958	5409	2.1	1.8	1.5	1.1	0.8	39%	32%	24%	17%	9%	0.040%	0.030%
99%	1237	8.31	0.84	161.0	979	0.9	0.9	0.8	0.8	0.8	21%	20%	19%	19%	18%	0.025%	0.024%

Buy and hold Sharpe Ratio = 0.89

Figure 4.8: Top 1250 RSM2 model results over the out-of-sample data period.

Number of models per stock = 1250 Number of days = 1130						Sharpe Ratio given level of transaction costs TDR2					Avg Final Equityper stock given level of trans costs					Daily Ret	Daily Ret
Thres %	Thres Val	Avg Trade % TDR2	Avg Trade % TDR1	Num Trades TDR2	Num Trades TDR1	0.0%	0.25%	0.50%	0.75%	1.00%	0.0%	0.25%	0.50%	0.75%	1.00%	0.25%	0.50%
0	0	0.66	0.66	14038	14038	4.5	2.9	1.2	-0.7	-2.8	142%	88%	34%	-20%	-74%	0.077%	0.030%
20%	250	0.95	0.65	8874	13642	3.9	3.0	2.0	0.9	-0.2	130%	96%	62%	27%	-7%	0.085%	0.055%
40%	500	1.33	0.68	5675	12448	3.9	3.3	2.6	1.8	1.1	116%	94%	73%	51%	29%	0.084%	0.064%
60%	750	1.84	0.65	3314	9991	3.0	2.7	2.3	1.9	1.4	94%	81%	68%	56%	43%	0.072%	0.061%
80%	1000	3.59	0.68	1456	6430	2.4	2.2	2.1	1.9	1.8	80%	75%	69%	64%	58%	0.066%	0.061%
99%	1237	35.59	0.75	76.0	597	1.3	1.3	1.3	1.3	1.3	42%	41%	41%	41%	40%	0.037%	0.036%

Buy and hold Sharpe Ratio = 0.86

Further details of RSM2's component models are provided in figures 4.9 to 5.9. Figure 4.9 depicts the frequency of occurrence of forecast horizons between 1 and 10. Most models use a shorter term horizon of between 1 to 4 days, as was expected, given that the model selection criteria favoured high frequency trading. Figure 4.10 depicts the frequency of occurrence of historical lag length and prime facie indicates a lack of dominant feature lags as each is approximately equally represented. However it could simply be that the models are robust against uninformative features - good performance is derived from informative lags even in the presence of noisy features.

Figure 5.8 shows clearly that the targets used were confined to close-to-close and open-to-close only, at all lags. This is also not surprising as short term forecast horizons (1-4 days) models which use either high-to-close or low-to-close targets will be penalized more heavily when ultimate performance is measured using close-to-close prices, as the shorter term horizons suffer from the bias mentioned in section 4.5.4. However, it may be that a different model selection method would produce different results in that one which favoured longer forecast horizons (by including transaction costs and thus reducing trading frequency) would be less affected by either high or low to close targets as they tend have a higher correlation to their close-to-close counterparts. Figure 5.8 and 5.9 show there to be little in terms of preference for particular price classes in the lag structure as each appears more or less equally.

Figure 4.9: Forecast horizon frequency.

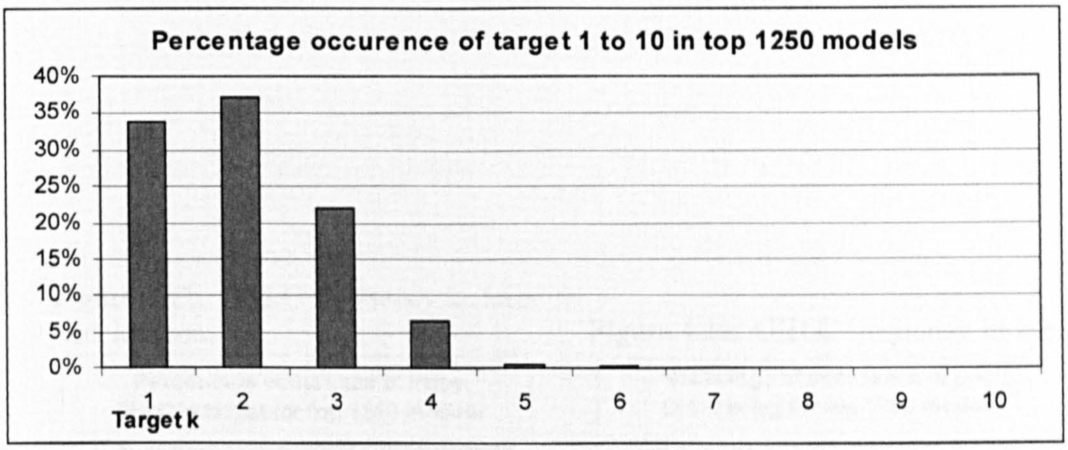


Figure 4.10: Lag frequency.

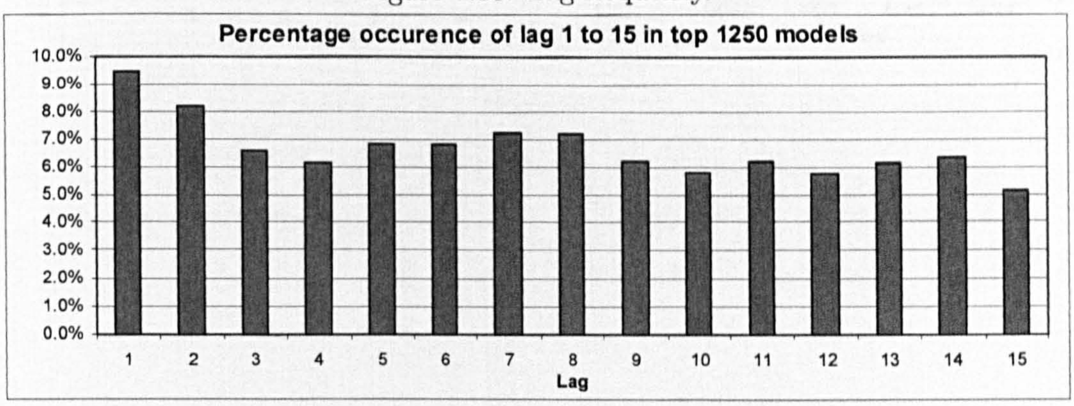


Figure 4.11: OHLC frequency in forecast horizon.

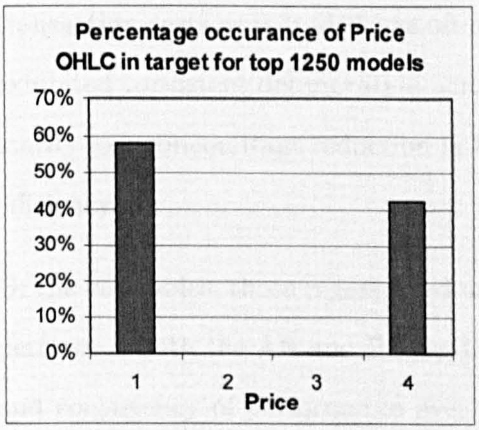
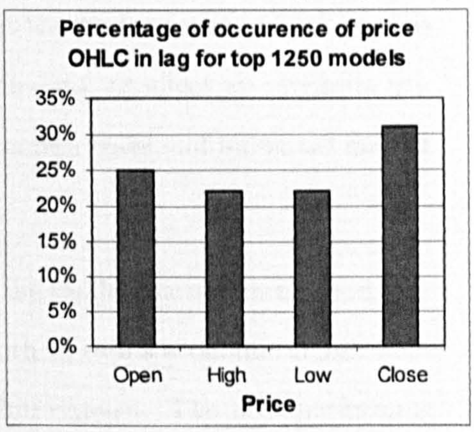


Figure 4.12: OHLC frequency in lags.



4.6 Summary and Conclusions

The objective of this study was to investigate the value of using ensembles over single models in the context of forecasting and trading 65 FTSE 100 components stocks. To this end the performance of four different models, comprising of a single AR model and three ensembles, was compared.

We found that the single AR model performed well until transaction costs were taken into account. The PE model outperformed the AR model net of transaction costs as it traded less often, but the performance of both models exhibited consistent deterioration across the data, an effect we attribute primarily to a concomitant reduction in transaction costs and increased market efficiency.

Of the ensembles, those constructed using the random subspace method outperformed both the AR and PE models both in terms economic significance and consistency of performance over the full dataset. The best performing model was obtained by selecting the top 2% of the RSM1 ensemble models and resulted in SRs of over 2.0 across the all data partitions. We conclude that creating ensembles of forecasting models adds value beyond that of single models and that the inclusion of extra price information in the form of the open, high and low improves the results.

4.6.1 Contributions of Study 4

1. For the first time in the literature we apply the RSM method in the context of financial forecasting models and find it to be an effective method for constructing ensembles of linear models.

2. We shed further light on the ability of standard AR models in forecasting stock prices.

3. The majority of literature on forecasting financial markets uses closing price data. We show that the daily open, high, low prices are important contributors to forecasts of asset returns.

4. The study is based on a large data sample which lends additional confidence to the results.

5. We contribute to the literature on nonlinearity in financial markets by reporting the results of the BDS test 65 FTSE 100 stocks.

Chapter 5

An investigation of a nonlinear technical trading rule over a portfolio of futures markets.

5.1 Introduction

Given that the persistent use of technical analysis is still an open question and that it contradicts the very foundation of modern finance, the Efficient Markets Hypothesis, it remains an important area of research. There now exists a large body of literature investigating the efficacy of various forms of technical analysis, with the number of studies increasing in recent times

(Park & Irwin 2004). A sizeable proportion of this work suggests this form of analysis may have value (Brock et al. 1992) though the evidence is countered by opposing studies, some suggesting positive results may be due to inadvertent data-snooping (Sullivan et al. 1999) thus, despite these efforts, the academic jury is still out.

In this study we concern ourselves with a subset of the technical analysis literature that relates to technical trading rules; mathematically definable trading indicators, usually based on price, that signal when a trader should enter and or exit the market. Many of these rules can be seen as nonlinear forecasting methods, some of which use the same underlying price transformation functions as those found in econometrics. Simple and exponential moving averages form the basis of ARIMA models and are also used for simple trading rules [1] , though the manner and intent in which they are applied differs. This chapter contributes to the extant literature by taking a somewhat different approach to many studies, which frequently analyse the performance of a selection of trading rules over either a single market or a restricted set of markets, see (Tomek & Querin 1984), (Qi & Wu 2002) for representative examples. Rather than looking at the performance of many different rules applied to a few markets we apply one simple incarnation of a popular technical trading rule, the channel break-out or trading range break-out, to a diverse portfolio of thirty seven futures markets. The rule in question has only one parameter - the length of the historical estimation

[1] For example, moving average crossover rules

window - which is estimated over the entire portfolio. Additionally, we test another commonly used rule - the moving average cross-over - for comparison and to shed further light on the generality of our results.

The channel break-out technical trading rule is examined over a portfolio of thirty seven futures markets from 1982 to 2005. We look not only at a large portfolio of markets, which few studies attempt, but also in a manner which closely resembles how practitioners implement these rules by including a trade management strategy (TMS) using stop-loss and profit-limit orders and accounting for transaction costs. Bootstrap tests are used to test the significance of the results and it is found that both the vanilla trading rule and the rule with the TMS added consistently realise significant net returns over the whole data sample.

Managed futures funds predominantly use technical trading as their primary methodology (Billingsley & Chance 1996) and our approach is motivated by the observation that many of them implement this type of methodology in the context of a diverse portfolio (Lukac, Brorsen & Irwin 1988), as combining over many markets reduces the effect of individual market idiosyncrasies across the portfolio. Moreover, it is rare for practitioners to simply follow trading rules in their basic form, they will often construct some form of trade management strategy, combining stop-loss and profit-limit orders, in an attempt to improve performance and reduce risk. In fact it is claimed that more emphasis should be placed on trade position management and allocation issues than should be placed on the specific trading rule as returns from different trend-following rules can be highly correlated (Vince 1999). Therefore it is arguably expedient to investigate trading rules in a manner resembling how they are actually implemented and as such we also look at the effect of adding a trade management strategy to the vanilla rule using stop-loss and profit-limit orders.

Much of literature on technical analysis has tended to focus on equity markets (Brock et al. 1992) but as we're investigating its efficacy within the context of a portfolio of futures markets, the majority of which are classed a commodities (70% of the portfolio), the literature that investigates commodity futures and their characteristics as an investment class is clearly relevant. Miffre & Rallis (2007) report on Jegadeesh & Titman (1993) style momentum profits in commodity futures. Gorton, Hayashi & Rouwenhorst (2007) show that commodity futures can be an effective vehicle for diversification with stock and bond portfolios. Marshall, Cahan & Cahan (2008) use the bootstrap reality test methodology of Sullivan et al. (1999) to test the profitability of trading rules in commodity futures and find that individually and when controlled for datasnooping trading rues are not profitable.

The channel breakout rule was chosen for a number of reasons; it has been shown historically to have value (Brock et al. 1992), (Taylor 1994),(Qi & Wu 2002) providing an imperative to investigate whether performance continues on recent data. The possibility of such a parsimonious formulation, using only one parameter, should make it less likely to over-fit the data and more likely to generalise out-of-sample.

Whilst it is true that main rule we have chosen has been tested on market data over many years by both practitioners and academics, we ameliorate the data-snooping issue to an extent by considering a final out-of-sample test of data not analysed by many of the key papers in this subject area, for the simple reason that they were written before the start of this period. Moreover, to further guard against over-fitting and data-snooping we divide the data into appropriate in-sample, validation and out-of-sample sets in an attempt to render as true an out-of-sample performance test as possible.

Bootstrap re-sampling tests are used to evaluate the results and it is found that the performance of the trading rule is both statistically and economically significant over the whole data period, even when accounting for transaction costs. The addition of a trade management strategy to the vanilla rule increases performance in terms of the Sharpe Ratio (SR) (Sharpe 1965) over the in-sample and validation sets but results are inconclusive over the final out-of-sample data. The remaining sections of this chapter are organised as follows: in the next section we present the relevant literature pertaining to this form of trading rule. Section 5.3 explains data and methodology and section 5.4 the results and lastly, section 5.5 our conclusions.

5.2 Literature Survey

Technical Analysis is the use of historical market generated data such as price, volume and, in some cases, open interest to forecast and trade financial markets. Derived from these three primary data streams there exist a multiplicity of methods devised by practitioners to form expectations conditional mainly on price which include technical indicators such as moving averages (Pring 2002), patterns based on chart formations (Bulkowski 2000), cycle analysis (Dewey 1971), candle stick charting (Nilson 1991), to name but a few.

The channel break-out is a so-called technical trading rule, a mechanical rule designed (alleged) to give indications of price direction. It is these mechanical trading rules that have received most attention from academics as they are easy to test in isolation - though some authors question the wisdom of this approach arguing that indicators play only a small part in forming practitioners' investment decisions.

The channel break-out rule has a long history and its underlying logic is relatively straight forward, relying on the concept of support and resistance, references to which can be found in (Wyckoff 1910). The idea is that market prices are driven by demand and supply (Murphy 1999) and that recent minimum price levels form a level of support at which demand is thought to be strong enough to prevent the price from declining further. Conversely, recent price maximums are thought to signify a level of resistance at which sellers are willing to supply (sell) the market and thus prevent the price from rising above this level. These support and resistance levels create a channel, or trading range, around the current price which, when broken, is said to signify a change in market sentiment and the beginning of a possible trend. There are a variety of methods for defining these levels though the simplest is to use a recent price maximum and minimum over a given historical data window. Donchian (1960) is usually credited with actually using the concept within a trading rule in which the preceding fortnight was used to define the channels.

The channel breakout belongs to class of trading rule considered by practitioners to be "trend following" (Pring 2002),(Schwager 1996) in that it is assumed that market prices are persistent to some degree ("markets trend") and by taking a position early enough in the direction of a trend the trader can gain positive returns. There exist innumerable variations, all following a common theme that involves trading when the price "breaks out" from a predefined channel which can be based on price or volatility - in which case it is a volatility breakout system (Pring 2002).

Osler (2000), provides a microstructural explanation for why trends gain momentum once predictable support and resistance levels - stop-loss and profit-limit order cluster points - are crossed, lending some credence to the notion held by adherents of technical analysis that when prices reach a certain level they can be prone to "break out" and produce trends. Brock et al. (1992) investigated both the channel and moving average trading rules and found that returns generated by applying these rules to the DJIA were not consistent with traditional asset pricing models, though transaction costs weren't accounted for and no inference regarding economic significance was made.

Lukac et al. (1988) tested twelve trading systems including the channel break out and found seven generated statistically significant returns and four generated significant risk-adjusted returns. In a more recent paper Park & Irwin (2005) replicate these results and investigate the performance on new data, finding that although positive performance over the original data is confirmed this does not persist in recent data periods. Lukac, Brorsen & Irwin (1989) tested the channel rule and the directional movement rule over fifteen futures markets and found results significantly greater than a buy and hold benchmark, but also noted that attempts to apply adaptive re-optimisation strategies using more recent data to estimate the parameters did not yield significant results. Taylor (1994) applied the channel rule to four currency futures and found significant net returns over the sample period. Lee (2001) tested moving average and channel rules on thirteen Latin American currencies and found for the channel rule significant positive net returns for three of the currencies and note that they were unable to discern which currencies would be profitable based on serial correlation statistics. For a further exposition on technical trading rules and technical analysis see appendix C

5.3 Data and Methodology.

We chose what we considered to be the simplest manifestation of this rule and define it thus:

Closing Price (CP) = the closing price at time t.

Highest Close $(HC) = max\{CP_{t-1}, \ldots, CP_{t-n+1}\}$, where CP_{t-1} is the closing price at time $t-1$ and n is the historical estimation window.

Lowest Close $(LC) = min\{CP_{t-1}, \ldots, CP_{t-n+1}\}$.

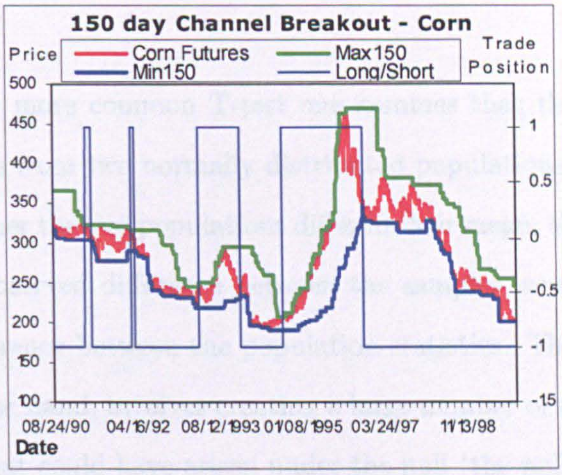
If $CP > HC$, then enter Long at tomorrow's Open.

If $CP < LC$, then enter Short at tomorrow's Open.

Figure 5.1 shows the rule applied to Corn futures using the last 150 days to define the channel.

The rule in this form results in the trader always having an open position in the market, only exiting when a subsequent trade is signalled in the opposite direction.

Figure 5.1: Corn futures and 150 day channel.



5.3.1 Bootstrap Randomisation Tests.

Measuring the performance of the rule over the validation set in terms of its SR provides some information as to its value however, it's not enough to simply observe that the trading rule signals result in an appreciable rate of return, what is more important is its performance compared to rules that have no "skill". One way of doing this is to compare the performance to what we would expect if the rule had no predictive ability.

This can be done by using randomisation tests, initially introduced by Fisher (1935), which give one the probability of the observed model's performance assuming the null hypothesis of no skill. Randomisation tests are a useful alternative to more traditional parametric tests for analysing empirical research data. They have the advantage of not making any distributional assumptions about the data, such as normality, yet still remain powerful tests.

When using the more common T-test one assumes that the data arose by drawing samples from two normally distributed populations, with the question being whether the two populations differ in their mean, that is, how likely is it that the observed difference between the samples would be realised if there is no difference between the population statistics. The randomisation test, on the other hand, involves creating a large number of randomised data set replicates that could have arisen under the null (the null in this case being no predictive ability) and then computing some statistic and examining its distribution. The empirical distribution of this statistic is then used to estimate alpha, the probability of rejecting the null hypothesis when in fact the null is true.

We can create an empirical distribution under the null by sampling with replacement the order of trades produced by the rule N times and then comparing its realised performance (SR) to the distribution of the SR under the null. We also look at two other measures of performance; total profit and the profit factor of each rule.

5.3.2 Methodology.

The data used in this study consist of the daily open, high, low and close of thirty seven futures contracts - see table 5.1. The time period covered is from January 1982 to January 2005, a total of twenty three years or eight hundred and fifty one years combined. A broad spectrum of asset classes are covered, the only criteria for inclusion was simply that the markets had price histories going back to 1982[2]. No attempt was made to choose markets that, from the literature, have shown to be amenable to trend-following systems. Moreover, the out-of-sample data was not viewed until all the results from the initial data were in.

Continuous futures contract adjustment.

When working with futures data it is frequently necessary to convert the individual contract data for each market into a continuous contract series whereby prices are adjusted to take into account the spread at rollover between the nearest contract to expiry and the next contract. This was accomplished by establishing the spread between the nearest contract and the next nearest contract at rollover and then adding the cumulative spread up to that contract to the new contract prices. Contracts are rolled over when the trading volume of the next contract is equal to, or greater than, that

[2]Note that we replaced the Deutschmark contract with the Euro when the former ceased trading

Table 5.1: 37 Futures markets.

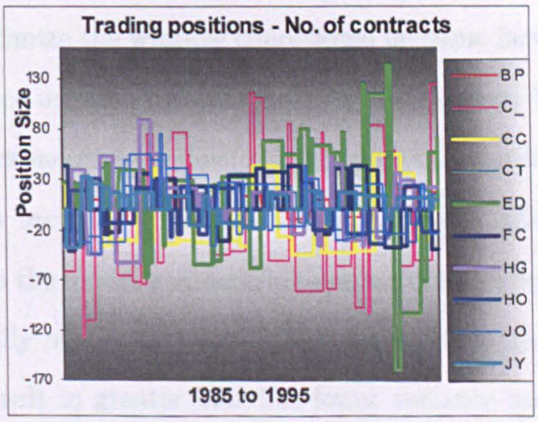
Name	Symbol	Exchange		Name	Symbol	Exchange
Soybean Oil	BO	CBT		Live Cattle	LC	CME
British Pound	BP	CME		Live Hogs	LH	CME
Corn	C	CBT		Minnesota Wheat	MW	MGE
Cocoa	CC	CS&CE		Oats	O	CBT
Canadian Dollar	CD	CME		Palladium	PA	NYME
Crude Light	CL	NYME		Pork Bellies	PB	CME
Cotton #2	CT	CTN		Platinum	PL	NYME
Deutschmark	DM	CME		Soybeans	S	CBT
Eurodollars 3MO	ED	CME		Sugar #11	SB	CS&CE
Feeder Cattle	FC	CME		Swiss Frank	SF	CME
Gold	GC	COMEX		Silver	SI	COMEX
Copper #1	HG	COMEX		Soybean Meal	SM	CBT
Heating Oil #2	HO	NYME		S&P Index	SP	CME
Unleaded Gasoline	HU	NYME		T Bilss 90 Days	TB	CME
Orange Juice	JO	CTN		Ten Year Notes	TY	CBT
Japanese Yen	JY	CME		T Bonds	US	CBT
Coffee	KC	CS&CE		Chicago Wheat	W	CBT
Kansas Wheat	KW	KCBT		NYSE Index	YX	NYFE
Lumber	LB	CME				

of the present contract. This splicing creates a new series with the contract rollover distortions removed. All exchange holidays are filled with data from the previous close.

Adhering to a policy of emulating what is used in practice we opted to use actual returns without converting to logs, though subsequent testing showed no significant difference in our results. In this case, when testing a strategy using a portfolio of futures contracts it is important to take into account the different *dollar* volatilities of the various markets traded, as the objective is for each market to contribute equally to overall portfolio performance. To futures traders the S&P 500 is considered a large contract, whereas oats is a small contract as the notional value of the former is much greater than the latter - if only one contract of both markets were traded the resulting two market portfolio would clearly be dominated by the S&P 500 futures. This can be ameliorated by normalising the number of contracts traded for each market and results in trading more contracts of those markets that move less in *dollars* per time period, and fewer contracts of those that have greater movements over the same period.

To do this the dollar value of a point was calculated for each market by dividing the dollar value of a tick - the minimum price movement possible - by its size. For example, using the CBOT wheat contract, the dollar value of a tick is \$12.50 and its size is 0.25, so the dollar value of 1 point is 12.50/0.25=\$50. A measure of volatility is then needed and in this case an average of the absolute values of daily close to close returns is taken over the last 100 days, although other measures of volatility can be used. The resulting dollar value of 1 point and the volatility value are then multiplied to obtain the daily dollar volatility (DDV). To compute the number of contracts a market should trade the DDV is divided into a constant of 6000 (chosen as it was greater than the maximum DDV of any single contract at any time during the data sample so that each market trades at least 1 whole contract), which means that the number of contracts traded in each market is such that each position will have a daily dollar volatility of about 6000. For example, in early 1997 the system was trading seven S&P contracts and twenty six coffee contracts. Figure 5.2 shows the position sizes for a number of markets over a 10 year period using the channel breakout system.

Figure 5.2: Position sizes for 10 markets.



The optimal estimation window.

A major issue when modelling data is that of over-fitting. Implicit to tests of trend-following rules is an assumption that market prices persist to some degree and, explicitly, a hypothesis that these rules may be able to exploit this general characteristic. As far as we know there is nothing from the literature to indicate how susceptible this class of model is to over-fitting and therefore it was decided to err on the side of caution. It was assumed that if indeed markets have a tendency to trend, they do so only weakly, and that this characteristic is homogenous across markets.

The choice variable that needs to be estimated in this case is the length of the historical estimation window - or equivalently the look-back period - that defines the channel. In the literature on return predictability a variety of methods for choosing the optimum window are length used, some of which attempt to re-optimize the window conditional on some factor such as recent model performance or more complex notions of structural breaks. The most popular include expanding window, rolling window, volatility and structural break estimation methods (Pesaran & Timmermann 2003), the choice of which depends on the model's underlying assumptions. The choice of window length conceptually relates to the so-called bias variance trade off in that long time windows result in greater bias but lower variance and vice versa with respect to shorter window lengths.

Another issue when testing trading rules is whether the estimation window should be optimized individually for each market (as is frequently the case in the literature), or across the whole portfolio. Lukac et al. (1988), in a study of trading rules that included the channel break-out which was conducted over a number of markets, used a methodology that involved re-optimising the estimation window every year on the preceding three years of data for each individual market.

We choose what we considered to be the simplest method which involves estimating the optimal window length across the whole portfolio. This goes some way in reducing the problem of model uncertainty in that we are restricting the size of the search space of possible models.

Trade Management Strategy - Adding Stop loss and Profit Limit orders.

Many practitioners consider a trade management strategy (TMS) to be an integral part of any trading system (Osler 2001). The objective is to balance the two competing yet equally important objectives of cutting losses short and letting profits run (Shiller 1996). It is said that when the market moves against a trade it is important to exit quickly, cutting any losses short, as it is probable that the trade was the result of a failed entry signal (Schwager 1998). In contrast, it is believed to be imperative to allow trades some room to breath as it may be that, after a brief initial reversal, the trade bounces back to produce large profits, in which case exiting early is detrimental. A properly designed TMS attempts to allow enough room for a trade to become profitable yet place a strong emphasis on controlling losses and preserving capital.

One of the more common strategies for cutting losses is the use of a stop-loss order, which in the case of a long (short) trade, involves placing a stop order at a price some distance below (above) the trade entry price. Once a trade is placed and the price hits the stop before any other exit criteria it is exited, with a loss equal to the difference between the entry price and the stop price - except when the market moves quickly through the stop, in which case the loss can be substantially greater.Theoretically it is difficult to justify the use of these forms of trade management strategies, however they have been showed to be efficacious under certain conditions. As Kaminski & Lo (2007) state,

> *"If the portfolio follows a random walk, i.e., independently and identically distributed returns, the stopping premium is always negative. This may explain why the academic and industry literature has looked askance at stop-loss policies to date. If returns are unforecastable, stop-loss rules simply force the portfolio out of higher-yielding assets on occasion, thereby lowering the overall expected return without adding any benefits. In such cases, stop-loss rules never stop losses. However, for non-random-walk portfolios, we find that stop-loss rules can stop losses. For example, if portfolio returns are characterized by "momentum" or positive serial correlation, we show that the stopping premium can be positive and is directly proportional to the magnitude of return persistence."*

It is not only important to know when to take a loss but also when to take a profit - once a trade has moved into profit it is quite possible for the price to then retrace, converting what was once a profitable trade into a losing one. One method of addressing this is to set a profit target some distance above (below) the entry price for a long (short) position so that when a trade becomes reasonably profitable this profit is locked in by exiting the position.

The vanilla channel breakout trading rule is what is known as a stop and reverse system as it exits a trade when a subsequent entry is signalled in the opposite direction, regardless of the initial trade's performance. To add the TMS, a static stop-loss and profit-limit order are added to each trade. They are static in that once the orders have been placed at a certain price they don't change and are either hit/executed or cancelled.

Normalising the TMS for multiple futures contracts.

In order to facilitate a meaningful test of the TMS over the portfolio of futures it is necessary to employ a method that can be applied across markets and is adaptive to changes in per contract volatility. It is inadvisable to use fixed dollar amounts as the volatility of individual contracts is liable to change over time, making what might have been a loose stop-loss in one period a tight one in another - a good example is that of S&P 500 futures; since this contract was first traded its volatility has increased, so to have a fixed dollar stop would make little sense, it would be hit far more often today than 15 years ago. In addition, using the same fixed dollar amount stops for both the S&P 500 and Oats would make little sense for the same reason that trading a similar number of contracts for both markets in the same portfolio isn't advisable.

To deal with this issue we use stop-losses and profit targets based on units of the average true range, as this allows us to test our exit strategy across the whole portfolio of futures.

Estimating the parameters.

We are estimating parameters for two models, the vanilla channel breakout rule (VCBO) and the channel breakout with the TMS included. The first step is to estimate the look-back period for the VCBO in-sample and then observe resulting performance on the validation set. Next the TMS is optimised in-sample using the length of estimation window chosen for the VCBO.

We partition the data into three sets rather than two in an attempt to mitigate data-snooping issues in that once we have observed the performance of the vanilla rule on the second partition any further model estimation using the same data is, in a sense, data-snooping as we already know the performance of the vanilla trading rule. Therefore, the third partition is used as a final out-of-sample set. The dataset is divided into three sets; in-sample, validation and out-of-sample consisting of 13, 5, and 5 years respectively, approximating a ratio of 60:20:20 - see Table 5.2.

The in-sample data was used to find the optimum estimation window for the VCBO using values between 10 and 210 days inclusive, in increments of 10

Table 5.2: Data set partitioning - 37 Futures Markets.

Set	Dates	Length
In-sample set	January 1982 to January 1995	13 Years
Validation set	January 1995 to November 1999	Approx. 5 years
Out-of-sample set	November 1999 to January 2005	Approx. 5 years

days. For example, the look-back parameter was set to 10 days and the rule was then applied to each individual market, results combined, and a Sharpe Ratio computed for the entire portfolio returns - this process was repeated for each parameter value. Once the optimal value was found subsequent rule performance was measured on the validation set.

To test the TMS the optimum value for the stop-loss and profit-limit orders are estimated on the in-sample data using the VCBO's optimum window and performance evaluated on the validation data. The last set of data allows us to investigate a true out-of-sample test of both systems. Finally, the performance on out-of-sample data is evaluated and reported for both the VCBO and TMS rules. Also included are transaction costs of $15 per round turn and 3 ticks slippage for each trade which is inline with what traders actually use (Katz 2000).

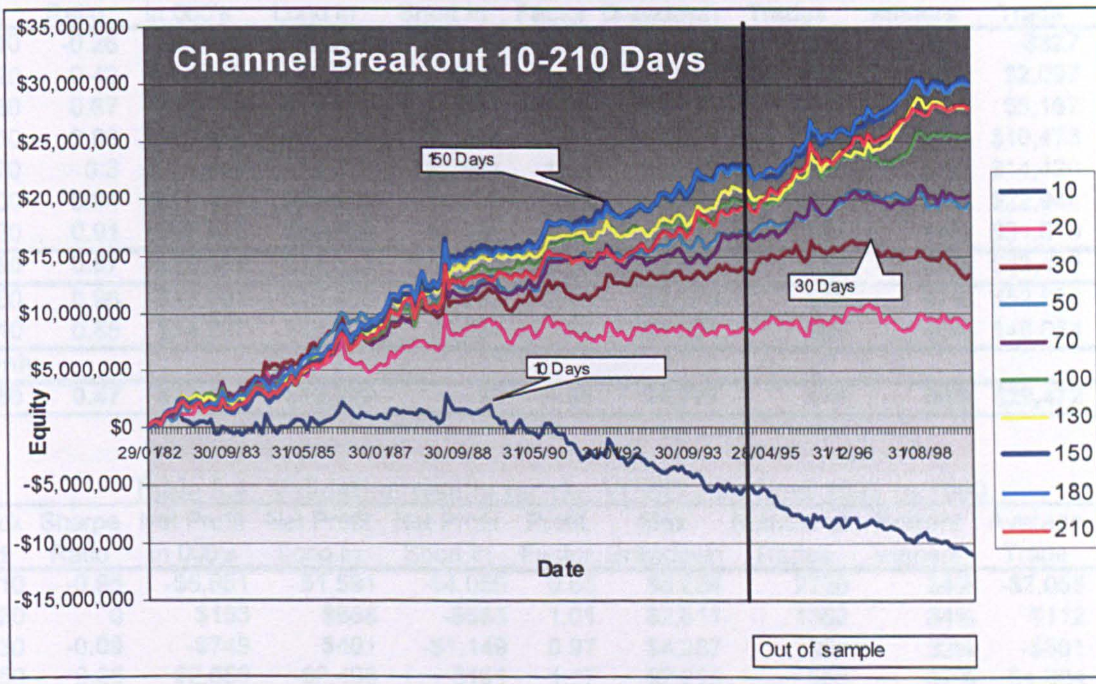
5.4 Results.

5.4.1 VCBO Results.

The results from the VCBO in-sample from 1982 to 1995 measured over the whole portfolio can be seen in Table 5.3. A visual representation of these results is shown in figure 5.3, which depicts the cumulative returns conditional on the rule for a number of the window lengths tested. It can be seen that shorter estimation windows result in lower SRs, primarily due to transaction costs. The 150 day period resulted in the highest in-sample SR though performance is fairly stable in this region of the parameter space. The results suggest a worsening of performance for those periods of less than 70 days, around the beginning of the 1990's, confirming Kidd & Brorsens' (2004) observation that returns to managed futures funds have decreased since around the same time. This cannot be said of the longer time periods where performance has been fairly consistent throughout this part of the data set. Interestingly, the results also confirm a common finding that short trades return less than long trades (Brock et. al 1992), contributing to just 16% of the total net returns at the 150 day period.

Included at the bottom of table 5.3 are the performance results of simply trading long all markets without transaction costs.

It is evident that a window of 150 days realised the highest SR in-sample so ure 5.3: Cumulative equity curves for the VCBO rule. Lookbacks from 10 to 210 days for the sample and validation periods.



this value was chosen for the rule. Performance was then measured over the validation set from 1995 to 1999, the results of which can be seen in figure 5.3 and table 5.4. These results are fairly consistent with those in-sample, with the 150 days estimation window realising a SR of 0.91, very similar to the in-sample value of 0.97. Most of the other statistics are consistent with their equivalents in-sample suggesting that the trading rule is continuing to generalise beyond the in-sample data.

Although these results are interesting it may simply be that we were fortu-

Table 5.3: In-sample results for VCBO Rule from 1982 to 1995.

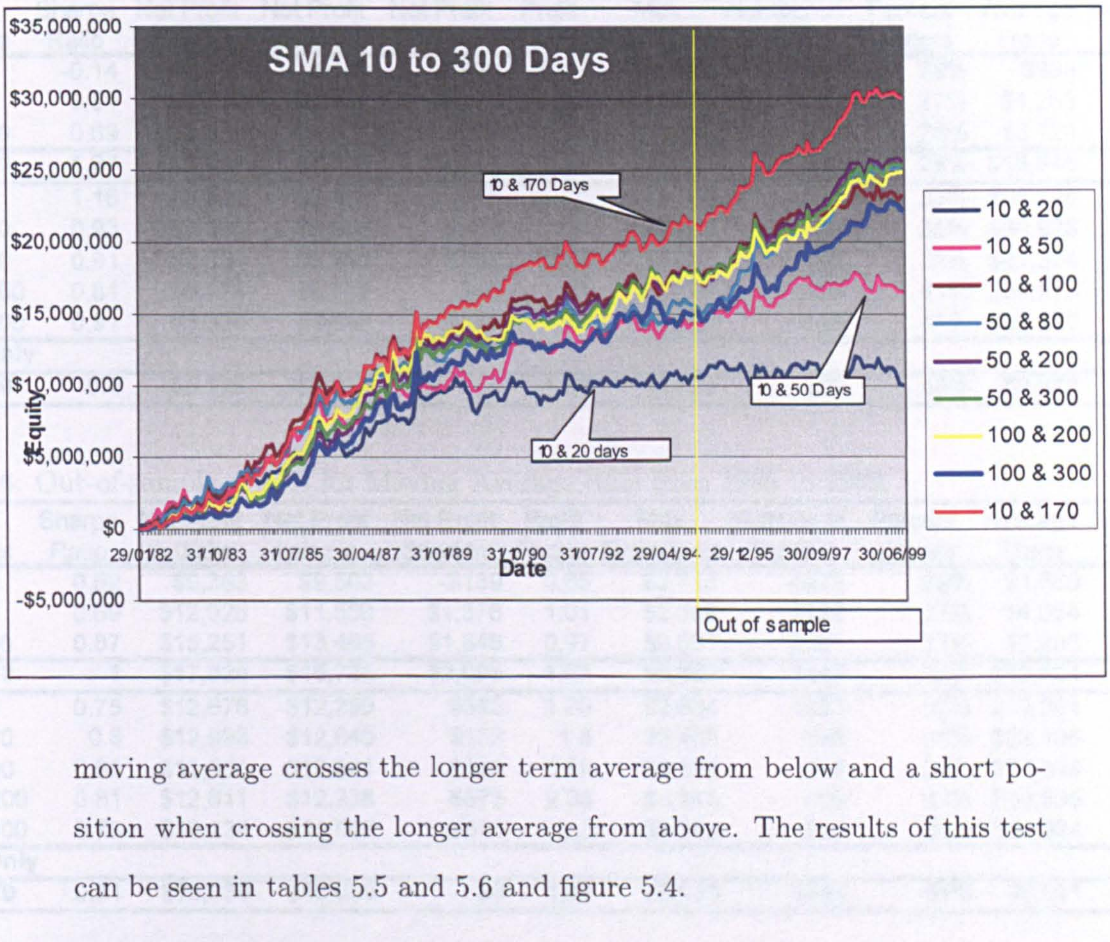
Lookback Period	Sharpe Ratio	Net Profit in 000's	Net Profit Long in	Net Profit Short in	Profit Factor	Max Drawdown	Number of Trades	Percent winners	Average Trade	Avg Trade in days
10	-0.26	-$6,071	$2,229	-$8,300	0.94	$9,480	7338	32%	-$827	18
20	0.42	$7,659	$8,657	-$998	1.12	$3,077	3652	33%	$2,097	35
30	0.67	$12,670	$11,112	$1,557	1.25	$2,596	2452	36%	$5,167	51
50	0.83	$15,061	$12,709	$2,351	1.41	$2,292	1438	38%	$10,473	85
70	0.8	$14,449	$12,887	$1,561	1.49	$2,603	1002	41%	$14,420	120
100	0.91	$15,464	$13,646	$1,817	1.63	$3,092	674	40%	$22,944	174
130	0.91	$15,936	$14,200	$1,735	1.78	$3,464	504	44%	$31,619	230
150	**0.97**	**$18,161**	**$15,323**	**$2,838**	**2.04**	**$3,392**	**418**	**44%**	**$43,449**	**277**
180	0.96	$17,791	$15,458	$2,333	2.19	$3,579	340	47%	$52,328	336
210	0.85	$14,227	$12,493	$1,734	1.97	$2,951	296	46%	$48,064	377
Long Only										
150	**0.47**	**$12,319**	**$12,319**	**$0**	**1.61**	**$4,777**	**418**	**50%**	**$29,472**	**277**

Table 5.4: Validation results for the VCBO rule from 1995 to 1999.

Lookback Period	Sharpe Ratio	Net Profit in 000's	Net Profit Long in	Net Profit Short in	Profit Factor	Max Drawdown	Number of Trades	Percent winners	Average Trade	Avg Trade in days
10	-0.95	-$5,651	-$1,591	-$4,059	0.85	$6,234	2750	34%	-$2,055	18
20	0	$153	$688	-$535	1.01	$2,611	1362	34%	$112	35
30	-0.09	-$748	$401	-$1,149	0.97	$4,287	934	33%	-$801	51
50	0.36	$2,656	$2,495	$161	1.17	$2,290	553	37%	$4,804	85
70	0.43	$3,445	$3,474	-$28	1.26	$2,391	395	36%	$8,723	118
100	0.69	$7,357	$6,518	$838	1.8	$2,631	258	39%	$28,517	183
130	0.86	$7,999	$6,823	$1,176	1.99	$2,552	192	41%	$41,664	244
150	**0.91**	**$7,402**	**$6,408**	**$993**	**2.08**	**$2,266**	**163**	**44%**	**$45,411**	**278**
180	0.89	$6,960	$6,017	$942	2.2	$2,531	134	51%	$51,942	339
210	0.94	$8,605	$7,647	$957	2.73	$2,920	117	54%	$73,548	414
Long Only										
150	**0.42**	**$5,328**	**$5,328**	**$0**	**1.68**	**$4,860**	**163**	**5400%**	**$32,688**	**278**

nate to have chosen a particular trading rule that happened to have profitable performance over this data period. To test whether these results are more general we tested another trading rule - the simple moving average crossover - using the same data. This rule signals a long position when the shorter

Figure 5.4: Cumulative equity curves for the Simple Moving Average double crossover. Subset of Crossovers from 10 to 300 days for the in-sample and out-of-sample periods.



moving average crosses the longer term average from below and a short position when crossing the longer average from above. The results of this test can be seen in tables 5.5 and 5.6 and figure 5.4.

The fact that the in-sample and out-of-sample moving average crossover rules results are somewhat similar to those of the VCBO rule leads us to conclude they are not due to just to luck. To test this more formally we conducted the aforementioned bootstrap re-sampling tests on the VCBO rule over the in-sample and validation data and then finally tested the rule on the third

Table 5.5: In-sample results for the from 10 to 300 day Simple Moving Average double crossover rule from 1982 to 1995.

SMA Lengths	Sharpe Ratio	Net Profit in 000's	Net Profit Long in	Net Profit Short in	Profit Factor	Max Drawdown	Number of Trades	Percent winners	Average Trade	Avg Trade in days
10 & 20	-0.14	-$651	$169	-$820	0.98	$2,699	1923	28%	-$338	25
10 & 50	0.31	$1,474	$1,200	$274	1.07	$1,917	1167	27%	$1,263	40
10 & 100	0.69	$4,939	$4,031	$908	1.31	$2,549	807	26%	$6,121	58
10 & 170	**1.07**	**$8,911**	**$7,175**	**$1,735**	**1.84**	**$2,313**	**529**	**26%**	**$16,845**	**87**
50 & 80	1.16	$9,402	$7,491	$1,910	2	$2,108	373	33%	$25,206	122
50 & 200	0.93	$7,361	$6,696	$664	1.98	$2,389	238	35%	$30,928	190
50 & 300	0.91	$7,654	$6,871	$782	2.25	$3,081	185	39%	$41,374	254
100 & 200	0.81	$6,728	$6,777	-$49	2.09	$2,656	170	41%	$39,578	268
100 & 300	0.91	$3,820	$3,988	-$167	1.71	$2,600	146	41%	$26,169	307
Long Only										
10 & 170	**0.4**	**$5,226**	**$5,226**	**$0**	**1.43**	**$4,429**	**529**	**48%**	**$9,879**	**87**

Table 5.6: Out-of-sample results for Moving Average Rule from 1995 to 1999.

SMA Lengths	Sharpe Ratio	Net Profit in 000's	Net Profit Long in	Net Profit Short in	Profit Factor	Max Drawdown	Number of Trades	Percent winners	Average Trade	Avg Trade in days
10 & 20	0.52	$9,363	$9,502	-$139	0.85	$2,713	4979	29%	$1,880	25
10 & 50	0.69	$12,926	$11,550	$1,376	1.01	$2,569	3188	27%	$4,054	39
10 & 100	0.87	$15,251	$13,405	$1,846	0.97	$2,591	2091	27%	$7,293	57
10 & 170	**1**	**$17,820**	**$15,725**	**$2,095**	**1.63**	**$2,862**	**1446**	**25%**	**$12,324**	**80**
50 & 80	0.75	$12,676	$12,280	$395	1.26	$2,634	1023	32%	$12,391	113
50 & 200	0.8	$12,998	$12,846	$152	1.8	$3,488	588	31%	$22,106	186
50 & 300	0.82	$12,504	$12,341	$162	1.99	$3,501	449	31%	$27,849	231
100 & 200	0.81	$12,611	$12,238	$373	2.08	$3,243	409	33%	$30,835	257
100 & 300	0.73	$10,134	$10,805	-$671	2.2	$3,293	291	36%	$34,824	338
Long Only										
10 & 170	**0.51**	**$13,030**	**$13,030**	**$0**	**1.43**	**$5,193**	**1446**	**51%**	**$9,011**	**80**

and final out-of-sample data set from 1999 to 2005. The results can be seen in table 5.7.

As might be expected the in-sample SR and Net profit figures are significant at the 1% level[3] (given that the window was optimised over the in-sample

[3] a single * denotes significance at 10% and a double * at 1%

Table 5.7: VCBO results for all data partitions.

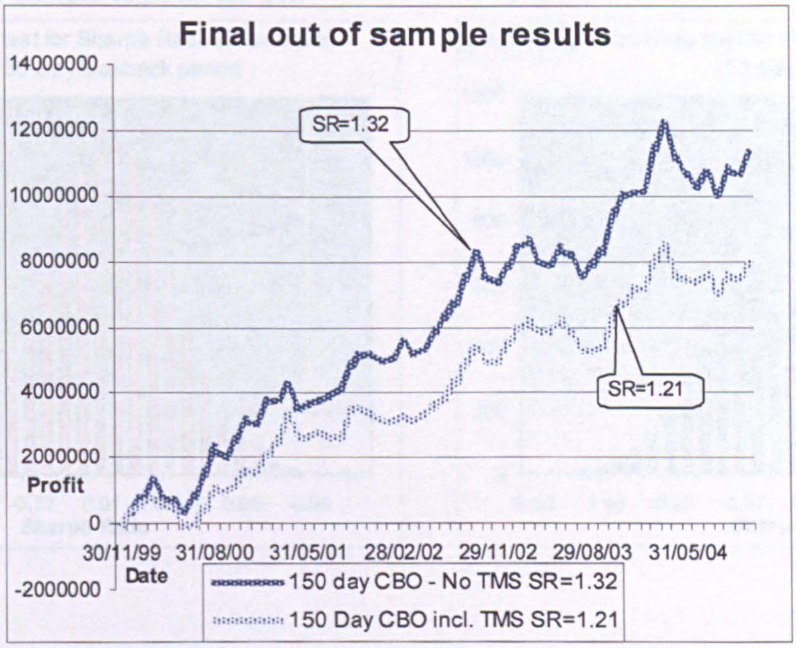
Lookback Period	Sharpe Ratio	Net Profit in 000's	Net Profit Long in	Net Profit Short in	Profit Factor	Max Drawdown	Number of Trades	Percent winners	Average Trade	Avg Trade in days
In sample results from 1982 to 1995										
150	**0.97****	**$18161****	**$15,323**	**$2,838**	**2.04**	**$3,392**	**418**	**44%**	**$43,449**	**277**
Long Only										
150	0.47	$12,319	$12,319	$0	1.61	$4,777	418	50%	$29,472	277
Validation results from 1995 to 1999										
150	**0.91***	**$7402****	**$6,408**	**$993**	**2.08**	**$2,266**	**163**	**44%**	**$45,411**	**278**
Long Only										
150	0.42	$5,328	$5,328	$0	1.68	$4,860	163	54%	$32,688	278
Final Out of sample Results from 1999 to 2005										
150	**1.32**	**$10,189**	**$7,226**	**$2,962**	**2.37**	**$2,386**	**163**	**42%**	**$62,510**	**296**
Long Only										
150	1.02	$4,182	$4,182	$0	1.4	$2,581	163	47%	$25,660	296

data). What is more interesting is that over the validation data the SR continues to be significant (but at the 10% level, whereas the net profit figure again significant at the 1%). The out-of-sample results also show signficance at the 1% level. This suggests that the rule is not performance on the basis of chance alone.

Table 5.8: TMS rule results for all data partitions.

Lookback Period	Sharpe Ratio	Net Profit in 000's	Net Profit Long in	Net Profit Short in	Profit Factor	Max Drawdown	Number of Trades	Percent winners	Average Trade	Avg Trade in days
In sample results from 1982 to 1995										
150	**1.32****	**$22,928**	**$16,332**	**$6,595**	**1.58**	**$1,587**	**1170**	**54%**	**$19,597**	**83**
Validation results from 1995 to 1999										
150	**1.19****	**$7,332**	**$4,742**	**$2,590**	**1.51**	**$1,515**	**409**	**51%**	**$17,927**	**86**
Final Out of sample Results from 1999 to 2005										
150	**1.21****	**$7,972**	**$8,000**	**-$27**	**1.48**	**$1,672**	**456**	**51%**	**$17,483**	**87**

Figure 5.5: Cumulative equity curves for VCBO and TMS rules out-of-sample.



5.4.2 TMS Results.

The TMS results can be seen in figure 5.8 (we haven't included detailed results for the sake of brevity). The total number of trades increases and the average holding period drops as a result of the TMS. It can be seen that the TMS rule exhibits a SR in-sample of 1.32 and 1.19 in validation and finally 1.21 out-of-sample. These are all significant at the 1% level. Prima facie this suggests that the TMS rule does improve the performance over the VCBO however, the VCBO exhibits a SR of 1.32 out-of-sample. Figure 5.5 depicts the out-of-sample cumulative equity curves for the VCBO and TMS rules.

Figure 5.6: In-sample Results for SR.

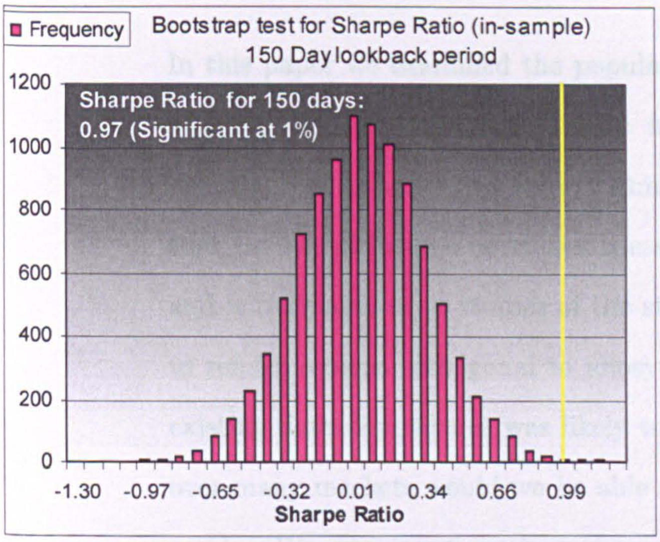


Figure 5.7: Out-of-sample Results for SR.

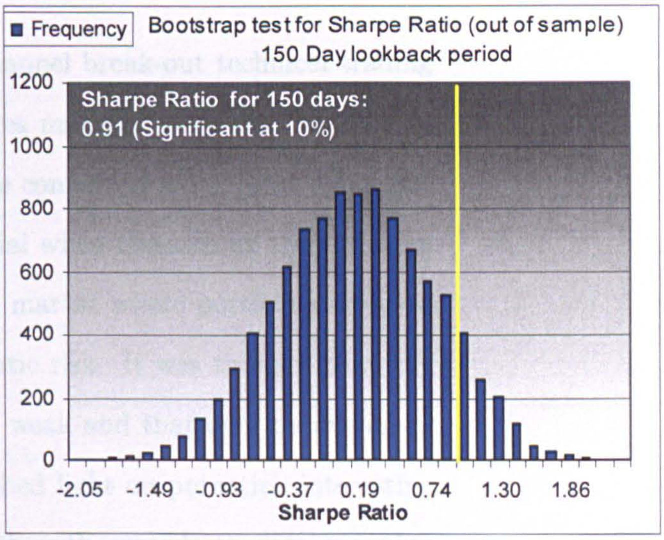


Figure 5.8: In-sample Results for Net Profit.

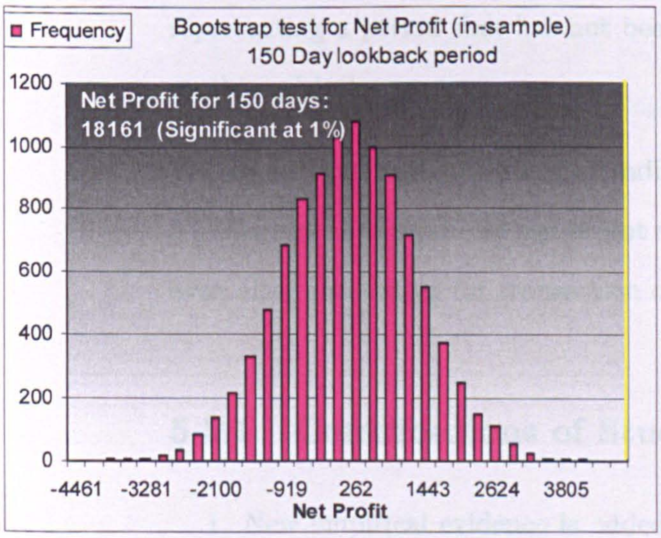
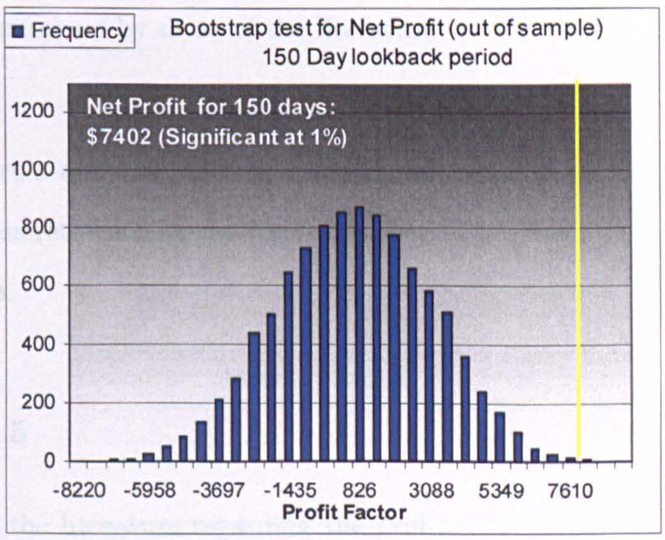


Figure 5.9: Out-of-sample Results for Net Profit.



5.5 Conclusion

In this paper we examined the popular channel break-out technical trading rule over a portfolio of thirty seven futures markets from 1982 to 2005 to investigate whether it had value within the context of a portfolio. We argue that the use of futures portfolios is essential when researching trading rules and is comparative to studies of the stock market where portfolios are used to render returns orthogonal to idiosyncratic risk. It was thought that any existing data dependence was likely to be weak and that only by averaging over many markets would we be able to shed light on potentiall interesting resuls. We were cautious in our approach with regards to data-snooping and over-fitting in that the data was divided into appropriate in-sample, validation and out-of-sample sets, with the last part of the data covered representing a period that has not been examined by most of the literature on this subject.

We found that both the vanilla trading rule and the rule with additional TMS component produced significant excess returns over the whole data set, even after accounting for transaction costs.

5.5.1 Contributions of Study 5

1. New empirical evidence is added to the literature regarding the profitability of technical trading rules and demonstrate that their use in

investment circles must primarily be their performance in the context of a portfolio.

2. We add to the extant literature by using a comparatively large dataset: The data used consists of daily close of 37 futures markets from 1982 to 2005, resulting in over 200,000 data points, more than all 56 studies that dealt with trading rules on futures and/or foreign exchange identified in Park & Irwin (2004)'s extensive review. Moreover, we also use arguably the simplest parameterisation of many of the studies reviewed, serving to increase the robustness of our results.

3. We shed further light on the prevalent use of stop and limit orders within trade management strategies by showing that they can be beneficial.

4. We cast some doubt on the finding that returns to technical trading rules have substantially diminished in recent years by showing that longer term lookback windows exhibit somewhat stable returns up to the year 2005.

Chapter 6

Summary and Conclusions

This thesis investigates the important question of asset price predictability. It does so via three self-contained studies that focus on whether:

1) Machine learning and statistical modelling techniques - both linear and nonlinear - which impose less restrictive assumptions on asset price dynamics than do classical linear methods, can be used to forecast and trade financial markets to a degree greater than that which traditional asset pricing models would lead us to expect.

2) To what extent model combination strategies can add value in this pursuit.

Based on the results of the three studies the answer to both of the above appears to be in the affirmative.

In chapter 3 we compared the performance of three different forecasting methodologies in predicting individual daily returns of 58 UK stocks in the FTSE 100. The methods used were support vector regression, k-nearest neighbours and linear regression The nonlinear support vector regression models exhibited superior performance over the linear regression models by factor of 2.

Chapter 4 investigated the application of linear ensemble models to forecast and trade 65 component stocks within the FTSE 100, using daily data over the years 1991-2006. The primary ensemble consisted of 62500 component models built using the random subspace method in which randomly sampled subsets of the feature set were used to estimate each model with the final result combined via a majority vote.

Of the ensembles, those constructed using the random subspace method outperformed both the AR and PE models both in terms economic significance and consistency of performance over the full dataset. The best performing model was obtained by selecting the top 2% of the RSM1 ensemble models and resulted in SRs of over 2.0 across the all data partitions. We conclude that creating ensembles of forecasting models adds value beyond that of single models and that the inclusion of extra price information in the form of the open, high and low improves the results. The results suggest that model combination strategies do indeed add value in forecasting financial markets.

In Chapter 5 we examined the channel break-out technical trading rule over a portfolio of thirty seven futures markets from 1982 to 2005. We look not only at a large portfolio of markets, which few studies attempt, but also in a manner which closely resembles how practitioners implement these rules by including a trade management strategy (TMS) using stop-loss and profit-limit orders and accounting for transaction costs. Bootstrap tests are used to test the significance of the results and it is found that both the vanilla trading rule and the rule with the TMS added consistently realise significant net returns over the whole data sample.

Chapter 7

Suggestions for Future Research

One of the important aspects of completing a Ph.D. is knowing when to cease doing empirical research and start writing up. Invariably there is always something extra that one could do to render the research that little bit more complete. The first suggestions are logical extensions from the central themes that were covered earlier in this thesis. Following that are some ideas that are somewhat less related.

7.1 Forecasting FTSE 100 stocks using support vector regression, linear regression and k-nearest neighbour methods

1. It would be interesting to investigate further the form of nonlinearity that the results suggest exists in the stock prices (if indeed it is nonlinearity rather than some other factor that is causing the superior performance of the SVM models). We need to investigate whether the SVMR models were superior in this chapter simply because they better able to deal with the curse of dimensionality (there were fourteen input features) and possible multi-colinearity in the training data than were the linear and knn models.

   This is important because if this is the case it may be not be that the data posses nonlinearities, but rather, that they just require a more stringent pre-processing procedure for linear models to achieve similar performance.

2. This study was carried out using a fixed historical estimation window of 2500 days to estimate the parameters of both the linear and nonlinear models. It would be interesting to investigate the result of using a rolling window approach as this would have the effect of including more recent data. One can see from figure 3.4 that the cumulative equity curves all start to flatten at the end of the out of sample period, possibly

as a result of redundant data.

3. Using a more sophisticated preprocessing stage may produce better results.

7.2 The more the merrier? Forecasting FTSE 100 stocks with a random subspace ensemble of 62500 models

1. Experimentation with different types of performance measurement would be interesting as the one used made no allowance for transaction costs.

2. The RSM2 ensemble was built via a fairly simple model selection method. There are many other approaches that might produce a more accurate result. For example one could choose say the top 100 models based on their in-sample sharpe ratio. Then each model could be added only if it increases the sharpe ratio of the ensemble as a whole. This should increase the orthogonality of the component models.

3. An interesting study would be to include volume and open interest information into a similar study performed on futures markets.

4. Replacing linear regression with a nonlinear variant would allow one to ascertain whether there were potential nonlinearities in the data which

lead to more accurate models.

7.3 An investigation of technical trading rules over a portfolio of futures markets

1. One aspect of this investigation that is striking is that the channel break-out trading rule, at least in the manner in which it was formulated here, takes very little of the available historical price information into account, yet still produces a Sharpe Ratio of around 1 for the portfolio. This is information is simply the most recent close in comparison with the maximum closing price within the historical estimation window. What might be interesting would be a study that attempted to exploit more information in terms of how the price actually moved within the window.

2. A form of classification overlay model could be built which uses the price movement within the window and tries to differentiate between those original break-out signals that resulted in a profit and those a loss. Although the 150 day window was optimal one could carry out this proposed study on all window lengths. Using shorter windows would mean larger datasets and hence potentially more significant results.

3. More recent papers in the literature test a large universe of trading

rules and test for significance using the bootstrap reality test. It might be interesting to conduct this class of study but using the portfolio of futures in this chapter.

4. What we have not touched upon is the possibility that there are structural drivers behind the results of this study. Structural in the sense that certain aspects of futures contracts themselves are behind the positive results. The literature suggests that the average return of individual commodity futures prices are not statistically different from zero (Ibbotson and Peng 2003), Dimson, Marsh and Staunton (2002), Gorton and Rouwenhorst(2005) and that they are largely uncorrelated, Erb & Harvey 2006 show the average correlation to be 0.09%.

   Prima facie this would suggest that a system that produces abnormal returns is likely to be doing well. However although the returns to individual commodity futures may be zero this is not necessarily the case in the context of a portfolio. The intuition that the return of a portfolio should mirror the weighted average of the portfolio's individual constituents holds in the case of unrebalanced portfolios of bonds. For example, a portfolio consisting of two bonds both with a return of zero would unlikely combine to produce a positive return. This is also the case with a portfolio of stocks. Siegel (2005) shows that both the S&P 500 and a weighted average of its constituents result in a 11% return from 1957 ro 2003. However, when the assets are uncorrelated and exhibit high average standard deviations the portfolio can experience a positive geometric return. Erb & Harvey (2006) put it thus:

> *"When the return of a portfolio is greater than the average return of the portfolio's constituents, and the portfolio constituents have average geometric risk premia of zero, then the portfolio weighting decision, not a geometric risk premium, is the source of the incremental return."*

Additionally it may be interesting to investigate how much the results are affected by the term structure of futures prices and the so-called roll return.

Appendix A

The Efficient Markets Hypothesis.

It was in 1978 that economist Michael Jensen famously pronounced that, "there is no other proposition in economics which has more solid empirical evidence supporting it than the Efficient Markets Hypothesis" (Jensen 1978). Even given the evidence available at the time this claim was somewhat hyperbolic. Yet Jensen's oft-quoted assertion is indicative of the extent that efficient markets theory had captured the collective imagination of a generation of economists and finance theorists by the 1970s.

It is often misunderstood which simply serves to fuel the controversy for example, it is frequently assumed that the EMH implies prices are unpredictable, that all investors are rational or that no investor will beat the market - it does not. It is necessary to this thesis to examine the EMH more closely in order to gauge its impact on the empirical forecasting applications contained herein. At the very least it provides a starting point on which to base one's ex ante chances of success.

The EMH is a central proposition of finance and much of modern financial theory rests on its far reaching implications i.e., option pricing, modern portfolio theory etc. Although its origins date back to 1565 (Hald 2003) it is Bachelier (1900) who is usually credited with having sown the initial seeds of the idea, in the context of stock returns, in his thesis entitled, "Theory of speculation":

> *The determination of these fluctuations depends on an infinite number of factors; it is, therefore, impossible to aspire to mathematical prediction of it. Contradictory opinions concerning these changes diverge so much that at the same instant buyers believe in a price increase and sellers in a price decrease (p. 17).*

Informally, the underlying rationale for this proposition lies in the idea that asset prices are impacted by supply and demand which is affected by new information or *news*. All presently known information is already incorporated in the price, so prices only change in reaction to news, which is by definition unpredictable, otherwise it wouldn't be news.

A.0.1 Classic taxonomy

It was Fama (1970) who is widely acknowledged to have laid down its modern foundations in terms of information sets, though the concept originally appeared in Roberts (1967):

Weak Form Efficiency: the information set includes only the history of prices.

Semi-strong Form Efficiency: the information set includes all information known to all market participants (publicly available information).

Strong Form Efficiency: the information set includes all information known to any market participant (private information).

It should be noted at this point that for this thesis it is the so- called weak form of the classic taxonomy that is relevant, which states that it is impossible to gain above average risk-adjusted returns on the basis of historical price information only.

During the two decades surrounding the publication of Fama's definitive paper (Fama 1970) the EMH seemed almost unassailable as, in addition to existing work, more empirical and theoretical evidence appeared to support it. Samuelson (1965) showed that the weak form was consistent with the random walk hypothesis. US and UK stocks and indices, along with some commodities, were shown to be efficient (Working 1934),(Cowles & Jonnes 1937), (Kendall 1953),(Osborne 1959),(Mandelbrot 1963),(Fama 1965). Event studies supported the semi-strong form (Ball & Brown 1968),(Fama, Fisher, Jensen & Roll 1969) and studies showing that no fund manager beat the market supported the strong form (Cowles 1934), (Jensen 1968). It was at this juncture that Jensen (1978) famously stated:

> *There is no other proposition in economics which has more solid empirical evidence supporting it that the efficient markets hypothesis.*

It was during the following decade that evidence began to emerge against the EMH as it was defined at the time, initially in the form of Shiller (1981) and LeRoy & Porter (1981), whose work on stock market volatility suggested that market prices were more volatile than were theoretically justified. Price variation was shown not to correspond with news (Cutler, Poterba & Summers 1989) and, in a paper that marked the beginning of Behavioural Finance, Bondt & Thaler (1985) showed that stock prices overreact, suggesting weak form market inefficiencies. On theoretical grounds Grossman (1976) and Grossman & Stiglitz (1980) pointed out that if markets are efficient and prices actually reflect all available information, there is no incentive for rational traders and arbitragers to collect this information in the first place. Further research elucidating the possibility that the EMH stood on less than hollowed ground includes (Fama & French 1988), (Lo & MacKinlay 1988), (Poterba & Summers 1988).

There is still a vigorous ongoing debate in the field as to what extent markets are efficient. Indeed there exists a chronological increase in clarification by various authors in the literature of what exactly is meant by *efficient* (see below) primarily in terms of information sets and in part to accommodate further empirical results casting some doubt against it (Jegadeesh & Titman 1993), (Levich & Thomas 1993), (Fama & French 1998), (Neely, Weller & Dittmar 1997).

A.0.2 Increased clarification or a movement of the goal posts?

Fama (1970)

> *A market in which prices always "fully reflect" available information is called "efficient".*

Jensen (1978)

> *A market is efficient with respect to information set θ_t if it is impossible to make economic profits by trading on the basis of information set θ_t.*

Fama (1998)

'...market efficiency (the hypothesis that prices fully reflect available information)...' '...the simple market efficiency story; that is, the expected value of abnormal returns is zero, but chance generates deviations from zero (anomalies) in both directions.'

Timmermann & Granger (2004)

'A market is efficient with respect to information set, Ω_t, search technologies, S_t, and forecasting models, M_t, if it is impossible to make economic profits by trading on the basis of signals produced from a forecasting model M_t, defined over predictor variables in the information set Ω_t, and selected using a search technology in S_t.'

A.0.3 The EMH and its impact on asset price predictability

Prima facie, attempting to predict asset returns is by most accounts liable to be a futile exercise, especially if prices move randomly in accordance to an independent and identically distributed process (*i.i.d.*).

In recent decades a large amount of effort has been expended conducting research into the behaviour of asset prices. Weak form market efficiency would suggest that prices traded in a market that is weak form efficient are not predictable using historical price information. This would imply that prices traded in such a market are serially uncorrelated. One method that has been adopted in the extant literature for testing weak form market efficiency has been an examination of asset prices for evidence of non-random behaviour. The Random walk hypothesis (RWH) posits that successive price changes in an efficient market are random. However, the RWH is somewhat restrictive in practice, as it implies that the process:

$$P_t = \beta P_{t-1} + \varepsilon_t \tag{A.1}$$

the error term ε is $i.i.d.$. If P_t represents the log of a market price, this implies that the market returns are $i.i.d.$ It is possible to relax the assumption of $i.i.d.$ returns in the context of weak form efficiency by replacing A.1 with:

$$E[P_{t+1} \mid I_t] = P_t \tag{A.2}$$

where I_t is any information set which includes $P_{t-j}, j \geq 0$. A.2 is a martingale process. It was Samuelson (1965) and Mandelbrot (1966) who formally recognised the importance of a martingale when describing efficient markets, though the seeds were sown by Bachelier (1900).

What this means is that tomorrow's price is expected to equal today's price given an asset's price history or, alternatively, an asset's expected return is zero when conditioned on its past price history. It's important to note that whilst a martingale process implies a correlation of zero between successive price changes, or the first moment of returns, it does not rule out dependence in higher moments. In fact, correlation between successive squared returns is a so-called *stylised fact* in finance literature.

For the martingale property to hold for prices investors are required to be risk neutral. However, LeRoy (1973), Cox & Ross (1976), Lucas (1978) and Harrison & Kreps (1979) pointed out that investors are in practice risk-averse and therefore there is no theoretical justification for the martingale property hence, a random walk is neither necessary nor sufficient for an efficient market. What this means is that asset price predictability does not necessarily imply *inefficiency* and conversely, asset prices that follow a random walk do not necessarily imply *efficiency*.

The EMH is further complicated by the "joint hypothesis problem". An efficient market is said to always fully reflect available information, however in order to determine how the market should fully reflect this information, investors risk preferences need to be determined. Therefore, any test of the EMH is a test of both market efficiency and investors risk preferences. So for this reason, the EMH per se is not a well-defined and empirically refutable hypothesis. This joint hypothesis problem was first pointed out by Fama (1970).

So to what extent would we expect asset prices to be predictable within the confines of EMH? We would expect predictability that doesn't refute the EMH if risk aversion or changes in risk are predictable as prices will reflect this. Additionally, asset prices may exhibit a measure of "residual" predictability that, if exploited, would not produce excess returns net of transaction costs.

A.0.4 Heterogenous Agent Models and The Efficient Markets Hypothesis

An important paradigmatic shift in finance has emerged since the 1990s, from a rational-expectations-efficient-markets (REEM) view towards a boundedly rational, heterogeneous agents, behavioral and evolutionary finance approach. The central keystones of the traditional well established Rational Representative Agent Paradigm are investor homogeneity and market efficiency. The REEM model held sway up until the 1990s where some of its inconsistencies began to be investigated more thouroughly in light of the emerging disciplines of Behavioural Finance and Heterogenous Agent Modelling (HAM). Up until the 1990s, the Efficient Markets Hypothesis (EMH) (Fama 1970) was the most widely–accepted and influential idea in financial economics. More recently, however, the concept of market efficiency has fallen into disrepute as it fails to adequately explain most of the stylized facts empirical puzzles characterizing the dynamics of financial markets.

It would be disingenuous to suggest the EMH has been usurped, it has not, and its advocates still argue vigorously in its defence and against what they consider to be heterodox. This is not surprising given the impact that the EMH and economic theory has had on modern finance, ranging from modern portfolio theory, the Capital Asset Pricing Model, Arbitrage Pricing Theory, Black-Scholes/Merton option pricing model to the Cox-Ingersoll-Ross theory of the term structure of interest rates. It remains a controversial topic ever since its inception.

Appendix B

Forecasting financial markets and asset pricing models.

B.1 Forecasting financial markets and asset pricing models

Intuitively, the higher the risk of investing in a particular asset, the higher the expected return required by an investor as compensation. The question then is how to measure and price risk such that a theoretical value or *price* can be assigned to an asset in terms of its individual risk characteristics. A widely accepted measure of risk is volatility, measured in terms of the standard deviation of an asset's returns. To solve the problem of pricing risk a number of asset pricing models have evolved. Asset pricing models provide a benchmark with which we can compare other assets in order to ascertain whether they offer fair value given their risk-return characteristics. Classic asset pricing models such as the Capital Asset Pricing Model (CAPM) (Sharpe 1964),(Lintner 1965) and (Mossin 1966) and Arbitrage Pricing Theory (APT) (Ross 1976) attempt to provide a prediction of the relationship between the risk of an asset and its expected return - the former in terms of so-called market risk and the latter in terms of unspecified risk factors that can include various fundamental, firm specific and statistical factors.

Both models differentiate between two types of risk:

1. Unsystematic or idiosyncratic risk
2. Systematic or market risk

The first type, unsystematic risk, is known a diversifiable risk as with an appropriately formed portfolio it can be virtually eliminated. The second type cannot be diversified away and hence investors are only compensated for accepting systematic risk.

Asset pricing models can be expressed in general terms:

$E(R_i) = f(F_1, F_2, F_3, ..., F_N)$

where $E(R_i)$, F_l and N denote the expected return on asset i, the l-th risk factor and the number of risk factors.

If an investor decides to invest in an asset that is not risk free they will expect a premium over and above the risk free rate

$E(R_i) = R_f + riskpremium$

where R_f is the risk free rate. The size of the risk premium will depend on those risk factors relevant to the particular asset and the manner by which the asset pricing model takes these into account. In this case the asset pricing model can be reformulated as

$$E(R_i) = R_f + g(F_1, F_2, F_3, ..., F_N) \tag{B.1}$$

B.1.1 The CAPM

The Capital Asset Pricing Model introduced by Sharpe (1964) can be expressed as:

$E(R_i) = R_f + \beta_i(E(R_M) - R_f)$

$\beta_i = \frac{Cov(R_i, R_M)}{Var(R_M}$

where $E(R_M)$ is the expected return on the portfolio and β_i the measure of systematic risk of asset i relative to the market portfolio.

In this way the CAPM can be considered a factor model that only takes into account one risk factor, the market risk.

B.1.2 The APT

$$E(R_i) = R_f + \sum_{k=1}^{k} \beta_{ik}(E(F_k) - R_f)) \tag{B.2}$$

With some minor conditions the CAPM can be considered a special case of the APT that only takes into account one risk factor, market risk.

As touched on previously, in any study of return predictability it is necessary to compare the results of any modelling procedure with what would be expected based on underlying theory. Empirical tests of the EMH are by their very nature joint hypothesis tests; one of the underlying asset pricing model and the other of whether the test results show significant departures from this assumed pricing model.

Appendix C

Technical Analysis and Trading Rules.

Technical analysis is the prediction of future asset prices based on an inductive study of historical price information and thus is contradictory with the notion of weak form efficiency. This form of market analysis covers a broad spectrum, incorporating both qualitative and quantitative techniques ranging from the positively arcane: astrology, numerology, to the somewhat sensible: linear regression.

The continued use of technical analysis by practitioners (Cheung, Chinn & Marsh 1999),(Oberlechner 2001) and the sheer amount of resources devoted to the study of its various esoteric forms is difficult to explain when contrasted against the backdrop of continued academic scepticism based on a belief in informationally efficient markets. Unlike astrology for example, a practice which holds currency in some circles regardless of general/academic ridicule, large pools of capital are invested by intelligent individuals on the premise of technical analysis. Given the Darwinian ruthlessness of markets at routing out inefficiencies it would be surprising that, at least at the extreme , its usage was simply a case of bright individuals[1] acting irrationally. If, as academic finance proclaims, it really has no value, that it persists presents something of a conundrum.

Of course, that it exists does not necessarily imply it must have value in the sense of gaining above-average returns by its usage, there are many reasons why a particular practice or belief may continue to exist in a population with scant real evidence of efficacy. Sewell (2007) describes technical analysis as representiveness, a psychological heuristic that people employ when making judgments under uncertainty posited by Tversky & Kahneman (1974). When trying to predict future events people often form their predictions by relying on a short period of historical data and extrapolating this forward. Sewell (2007) gives other psychological reasons for the persistent use of TA.

1. Communal Reinforcement. This is a social construction in which a strong belief is formed when a claim is repeatedly asserted by members

[1]In the case of well financed funds which tend to be run by those who appear in the far right tail of the IQ bell curve (Chevalier & Ellison 1999)

of a community, rather than due to the existence of empirical evidence for the validity of the claim.

2. Selective Thinking. This is the process by which one focuses on favourable evidence in order to justify a belief, ignoring unfavourable evidence.

3. Confirmation Bias. This is is a cognitive bias whereby one tends to notice and look for information that confirms one's existing beliefs, whilst ignoring anything that contradicts those beliefs. It is a type of selective thinking.

4. Self-deception. This is the process of misleading ourselves to accept as true or valid what we believe to be false or invalid by ignoring evidence of the contrary position.

Theoretical reasons that support the persistent use of technical analysis are found in the form of noisy rational expectation models that rest on the idea that price, rather than adjusting instantaneously to new information, adjusts sluggishly due to such factors as noise, market frictions, market power, investors sentiments or herding behavior, or chaos. Smidt (1965) describes two futures market, one where participants have perfect information and another which consists of two types of traders, "insiders" and "outsiders", and insiders are privy to new information before outsiders. He posits that technical analysis is unlikely to perform in the former or in the latter if insiders perfectly predict outsider's performance. In the event that insiders do not predict well there may be a situation where market rises or falls tend to persist which may provide an environment where technical analysis could result in long term profits.

There now exists a large body of literature investigating the efficacy of various forms of technical analysis with the number of studies increasing in recent times (Park & Irwin 2004). The majority focus on technical trading rules, mathematically definable trading indicators, usually based on price, that signal when a trader should enter and or exit the market. The reason for this focus is simply that easily defined trading rules are more readily testable than the more subjective forms of technical analysis, though there have been some forays into testing the latter (Lo 2000),(Osler 2000).

Despite academic scepticism a sizeable proportion of the work suggests this form of analysis may have value. In a study by Park & Irwin (2004), of 92 "modern studies" [2] dating from 1988, 58 (63%) show positive results, a somewhat surprising result given the robustness of the EMH. On that note we would not want to underestimate potential publication bias, the impact of which, some authors claim, renders most published research findings to be false (Ioannidis 2005). Refutation of positive results does not rest solely on publication bias, many of these studies are also criticised rather convincingly on a theoretical basis. Sullivan et al. (1999) suggest that positive results may

[2] They define *Modern studies* as those that account for one or more of the following; risk, transaction costs and datasnooping. By contrast *Early studies* fell short in one or more of these areas.

be due to inadvertent data-snooping (see section 2.12) whilst others have shown that many did not account sufficiently for risk (Brock et al. 1992) thus, despite these efforts in support of technical analysis, the academic jury is still out.

Much of literature on technical analysis has tended to focus on equity markets (Brock et al. 1992) but as we investigated its efficacy within the context of a portfolio of futures markets, the majority of which are classed a commodities (70% of the portfolio), the literature that investigates commodity futures and their characteristics as an investment class is also clearly relevant. Miffre & Rallis (2007) report on Jegadeesh & Titman (1993) style momentum profits in commodity futures. Gorton et al. (2007) show that commodity futures can be an effective vehicle for diversification with stock and bond portfolios. Marshall et al. (2008) use the bootstrap reality test methodology of Sullivan et al. (1999) to test the profitability of trading rules in commodity futures and find that individually and when controlled for datasnooping trading rues are not profitable.

Bibliography

Abarbanel, H. (1995), *Analysis of Observed Chaotic Data*, Springer, New York.

Abraham, A., Philip, N. S. & Saratchandran, P. (2003), 'Modeling chaotic behavior of stock indices using intelligent paradigms', *International Journal of Neural, Parallel and Scientific Computations* **11**, 143.

Aha, D., Kibler, D. & Albert, M. (1991), 'Instance-based learning algorithms', *Machine Learning* **6**, 37–66.

Akaike, H. (1973), Information theory and an extension of the maximum likelihood principle, *in* B. N. Petrov & F. Csaki, eds, 'Second International Symposium on Information Theory', Akademiai Kiado, Budapest, pp. 267–281.

Amaldi, E. & Kann, V. (1998), 'On the approximation of minimizing non zero variables or unsatisfied relations in linear systems', *Theoretical Computer Science* **1**(209), 237–260.

Antoniou, A. & Vorlow, C. (2005), 'Price clustering and discreteness: Is there chaos behind the noise?', *Physica A* **348**, 389–403.

Asuncion, A. & Newman, D. (2007), 'UCI machine learning repository'.

Bachelier, L. (1900), 'Théorie de la spéculation', *Annales Scientifiques de l'École Normale Supérieure Sér.* **3**(17), 21–86.

Ball, R. & Brown, P. (1968), 'An empirical evaluation of accounting income numbers', *Journal of Business* **6**, 159–178.

Barabino, N., Pallavicini, M., Petrolini, A., Pontil, M. & Verri, A. (1999), Support vector machines vs multi-layer perceptrons in particle identification, *in* 'Proceedings of the European Symposium on Artifical Neural Networks '99', D-Facto Press, Belgium, pp. I: 257–262.

Barber, B. & Odean, T. (2000), 'Trading is hazardous to yourwealth: The common stock investment performance of individual investors', *Journal of Finance* **55**, 773–806.

Barnard, G. A. (1963), 'New methods of quality control', *Journal of the Royal Statistical Society A* (126), 255–259.

Bates, J. & Granger, C. (1969), 'The combination of forecasts', *Operations Research Quarterly* **20**, 451–468.

Bell, D. (1982*a*), 'Regret in decision making under uncertainty', *Operations Research* **30**, 961–981.

Bell, D. (1982*b*), 'Risk premiums for decision regret', *Management Science* **29**, 1156–1166.

Bengio, Y. & Grandvalet, Y. (2004), 'No unbiased estimator of the variance of k-fold cross-validation', *Journal of Machine Learning Research* **5**, 1089–1105.

Bertoni, A., Folgieri, R. & Valentini, G. (2005), 'Bio-molecular cancer prediction with random subspace ensembles of support vector machines', *Neurocomputing* **1**, 63C:535–539.

Billingsley, R. S. & Chance, D. M. (1996), 'Benefits and limitations of diversification amoung commodity trading advisors.', *Journal of Portfolio Management* **Fall**, 65–80.

Black, D. (1958), *The Theory of Committees and Elections*, Cambridge University Press.

Bondt, W. F. M. D. & Thaler, R. (1985), 'Does the stock market overreact?', *The Journal of Finance* **40**(3), 793–805.

Boser, B. E., Guyon, I. M. & Vapnik, V. N. (1992), A training algorithm for optimal margin classifiers, *in* 'COLT '92: Proceedings of the fifth annual workshop on Computational learning theory', ACM Press, New York, NY, USA, pp. 144–152.

Breiman, L. (1996), 'Bagging predictors', *Machine Learning* **24**(2), 123–140.

Brock, W., Dechert, W., Scheinkman, J. & LeBaron, B. (1996), 'A test for independence based on the correlation dimension', *Econometrics Review* **15**(3), 197–235.

Brock, W., Lakonishok, J. & LeBaron, B. (1992), 'Simple technical trading rules and the stochastic properties of stock returns', *The Journal of Finance*.

Brooks, C. & Henry, O. (1999), Can portemanteau nonlinearity tests serve as general mis-specification tests? evidence from symmetric and asymmetric garch models, Department of Economics - Working Papers Series 723, The University of Melbourne.

Brors, B., Kohlmann, A., Schoch, C., Schnittger, S., Haferlach, T. & Eils, R. (2000), 'Aggregation of classifiers makes supervised analysis of molecular genetics data more robust statistical', *Statistical* **12**, 993–1001.

Bulkowski, T. N. (2000), *Encyclopedia of Chart Patterns.*, John Wiley & Sons.

Burges, C. (1998), 'A tutorial on support vector machines for pattern recognition', *Data Mining and Knowledge Discovery* **2**(2), 121–167.

Campbell, J., Lo, A. & MacKinlay, A. C. (1997), *The Econometrics of Financial Markets*, Princeton University Press, Princeton, New Jersey.

Chawla, N. V. & Bowyer, K. W. (2005), Random subspaces and subsampling for 2-d face recognition, *in* 'Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on', Vol. 2, IEEE, pp. 582–589.

Chen, D. & Cheng, X. (2000), 'A simple implementation of the stochastic discrimination for pattern recognition', *Advances in Pattern Recognition, Proc. Joint IAPR Int. Workshops SSPR 2000 and SPR 2000 (Francesc J. Ferri, Jose M. Inesta, et l., eds.)* pp. 882–887.

Cheung, Y.-W., Chinn, M. D. & Marsh, I. W. (1999), How do uk-based foreign exchange dealers think their market operates?, CEPR Discussion Papers 2230, C.E.P.R. Discussion Papers. available at http://ideas.repec.org/p/cpr/ceprdp/2230.html.

Chevalier, J. & Ellison, G. (1999), 'Are some mutual fund managers better than others? cross- sectional patterns in behavior and performance', *The Journal of Finance* **54**(3), 875–899.

Cho, S. & Kim, J. (1995), 'Multiple networks fusion using fuzzy logic', *IEEE Transactions on Neural Networks* (6), 497–501.

Clarke, R., Krase, S. & Statman, M. (1994), 'Tracking errors, regret, and tactical asset allocation', *Journal of Portfolio Management* **20**, 16–24.

Condorcet NC de (1785), 'Essai sur lapplication de lanalyse a la probabilite des decisions rendues a la pluralite des voix. see i. mclean and f. hewitt, translators, 1994'.

Cortes, C. & Vapnik, V. (1995), 'Support-vector networks', *Machine Learning* **20**(3), 273–297.

Cournot, A. A. (1843), 'Exposition de la theorie des chances et des probabilites'.

Cover, T. M. & Hart, P. E. (1968), 'Nearest neighbor pattern classification', *IEEE Transactions on Information Theory*.

Cowles, A. (1934), 'Can stock market forecasters forecast?', *Econometrica* **1**, 309–324.

Cowles, A. & Jonnes, H. (1937), 'Some a posteriori probabilities in stock market action', *Econometrica* **5**, 280–294.

Cox, J. C. & Ross, S. A. (1976), 'The valuation of options for alternative stochastic processes', *Journal of Financial Economics* **3**(1–2), 145–166.

Cristianini, N. & Shawe-Taylor, J. (2000), 'An introduction to support vector machines and other kernel-based learning methods'.

Cunningham, P. & Zenobi, G. (2001), 'Case representation issues for case-based reasoning from ensemble research', *Lecture Notes in Computer Science* **2080**, 146–157.

Cutler, D. M., Poterba, J. M. & Summers, L. H. (1989), 'What moves stock prices?', *Journal of Portfolio Management* **15**(3), 4–12.

De Grauwe, P. & Grimaldi, M. (2005), 'Heterogeneity of agents,transactions costs and the exchange rate', *Journal of Economic Dynamics & Control* **29**, 691–719.

DeBondt, W. & Thaler, R. (1986), 'Does the stock market overreact?', *Journal of Finance* **40**, 793–807.

DellÀquila, R. & Ronchetti, E. (2006), 'Stock and bond return predictability: The discrimination power of model selection criteria', *Computational Statistics and Data Analysis* (50), 1478–1495.

Denton, F. T. (1985), 'Data mining as an industry', *Review of Economics and Statistics* **67**, 124–127.

Dewey, E. R. (1971), *The Mysterious Forces that Trigger Events.*, Hawthorn, New York.

Dietterich, T. G. (2000*a*), 'Ensemble methods in machine learning', *In J. Kittler and F. Roli, editors, Multiple Classifier Systems, First International Workshop.*

Dietterich, T. G. (2000*b*), 'An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization', *Machine Learning* **40**(2), 139–158.

Dietterich, T. G. (2002), Ensemble learning, *in* M. A. Arbib, ed., 'The Handbook of Brain Theory and Neural Networks', second edn, The MIT Press, Cambridge, MA.

Donchian, R. D. (1960), 'High finance in copper', *Financial Analysts Journal* **Nov-Dec**, 133–142.

Dudani, S. (1975), 'The distance-weighted k-nearest neighbor rule', *IEEE Transactions on Systems, Man and Cybernetics* **6**(4), 325–327.

Duin, R. & Tax, D. (2000), Experiments with classifier combining rules, pp. 16–29.

Efron, B. (1979), 'Bootstrap methods: Another look at the jackknife', *The Annals of Statistics* **7**(1), 1–26.

Fama, E. F. (1965), 'The behavior of stock-market prices', *Journal of Business* **33**(1), 34–105.

Fama, E. F. (1970), 'Efficient capital markets: A review of theory and empirical work', *Journal of Finance.*

Fama, E. F. (1998), 'Market efficiency, long-term returns, and behavioral finance', *Journal of Financial Economics* **49**(3), 283–306.

Fama, E. F., Fisher, L., Jensen, M. C. & Roll, R. (1969), 'The adjustment of stock prices to new information', *International Economic Review* **10**, 1–21.

Fama, E. F. & French, K. R. (1988), 'Permanent and temporary components of stock prices', *The Journal of Political Economy* **96**(2), 246–273.

Fama, E. F. & French, K. R. (1998), 'Value versus growth: The international evidence', *The Journal of Finance* **53**(6), 1975–1999.

Fischoff, B. & Slovic, P. (1980), A little learning: Confidence inmulticue judgment tasks., NJ: Erlbaum.

Fisher, R. A. (1935), *The Design of Experiments*, 9th Edition 1971, reprinted 1974. Hafner Press, New York.

Fix, E. & Hodges, J. (1952), Discriminatory analysis: Nonparametric discrimination: Small sample performance, Technical Report 11, USAF School of Aviation Medicine, Randolf Field, Texas. Technical Report Project 21-49-004.

Freund, Y. & Schapire, R. E. (1996), Experiments with a new boosting algorithm, *in* L. Saitta, ed., 'Machine Learning: Proceedings of the Thirteenth International Conference (ICML '96)', Morgan Kaufmann, San Francisco, CA, pp. 148–156.

Geman, S., Bienenstock, E. & Doursat, R. (1992), 'Neural networks and the bias/variance dilemma', *Neural Computation* **4**, 1–58.

Gemmill, G. (1996), 'Transparency and liquidity: A study of block trades on the london stock exchange under different publication rules', *Journal of Finance* **51**(5), 1765–1790.

Gemmill, G. (1998), 'Touched by success', *Financial News Briefing Notes, October 5.*

Gervais, S. & Odean, T. (2001), 'Learning to be overconfident', *Review of Financial Studies* **14**, 1–27.

Gorton, G., Hayashi, F. & Rouwenhorst, K. (2007), Fundamentals of commodity futures returns, NBER Working Papers 13249, National Bureau of Economic Research, Inc.

Goutte, C. (1997), 'Note on free lunches and cross-validation', *Neural Computation* **9**(6), 1245–1249.

Granger, C. & Jeon, Y. (2004), 'Thick modeling', *Economic Modeling* (21), 323–343.

Grossman, D. & Williams, T. (2000), 'Machine learning ensembles: An empirical study and novel approach. unpublished manuscript'.

Grossman, S. J. (1976), 'On the efficiency of competitive stock mar- kets where trades have diverse information', *Journal of Finance* **31**(2), 573–585.

Grossman, S. J. & Stiglitz, J. E. (1980), 'On the impossibility of informationally efficient markets', *The American Economic Review*.

Guyon, I. & Elisseeff, A. (2003), 'An introduction to variable and feature selection', *Journal of Machine Learning Research* **3**, 1157–1182.

Guyon, I. M., Boser, B. E. & Vapnik, V. N. (1993), Automatic capacity tuning of very large vcdimension classifiers, *in* 'Advances in Neural Information Processing Systems', pp. 147–155.

Hansen, P. R. (2005), 'A test for superior predictive ability', *Journal of Business and Economics Statistics* **23**, 365–380.

Hanson, L. & Salomon, P. (1990), 'Neural network ensembles', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **12**, 993–1001.

Harrison, J. M. & Kreps, D. M. (1979), 'Martingales and arbitrage in multiperiod securities markets', *Journal of Economic Theory* **20**(3), 381–408.

Hendry, D. F. & Clements, M. P. (2004), 'Pooling of forecasts', *Econometrics Journal* **7**(1), 1–31.

Herbert, S. (1957), 'A behavioral model of rational choice', *Models of Man, Social and Rational: Mathematical Essays on Rational Human Behavior in a Social Setting.*

Ho, T. K. (1992), 'A theory of multiple classifier systems and its application to visual word recognition'.

Ho, T. K. (1998), 'The random subspace method for constructing decision forests', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **20**(8), 832–844.

Huang, Z., Chen, H., Hsu, C.-J., Chen, W.-H. & Wu, S. (2004), 'Credit rating analysis with support vector machines and neural networks: a market comparative study', *Decis. Support Syst.* **37**(4), 543–558.

Huberman, G. & Regev, T. (2001), 'Contagious speculation and a cure for cancer: A nonevent that made stock prices soar', *Journal of Finance* **56**, 387–396.

Ince, H. (2000), Support vector machine for regression and applications to financial forecasting, *in* 'IJCNN '00: Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks (IJCNN'00)-Volume 6', IEEE Computer Society, Washington, DC, USA, p. 6348.

Ioannidis, J. P. A. (2005), 'Why most published research findings are false', *PLoS Medicine* **e124**, 0696–0701.

Jegadeesh, N. & Titman, S. (1993), 'Returns to buying winners and selling losers: Implications for stock market efficiency', *Journal of Finance* **48**(1), 65–91.

Jensen, M. (1968), 'The performance of mutual funds in the period 1945-1964', *Journal of Finance* **23**, 389–416.

Jensen, M. C. (1978), 'Some anomalous evidence regarding market efficiency', *Journal of Financial Economics* **6**(2/3), 95–101.

Joachims, T. (1997), Text categorization with support vector machines, Technical Report 23, LS VIII University of Dortmund.

Kahneman, D. & Tversky, A. (1979), 'Prospect theory: An analysis of decision under risk', *Econometrica* **47**, 263–291.

Kaminski, K. & Lo, A. W. (2007), 'When Do Stop-Loss Rules Stop Losses?', *SSRN eLibrary*.

Kearns, M. (1988), 'Thoughts on hypothesis boosting', *Class Project*.

Keller, J. & Yang, H. (2000), Fuzzy aggregation networks of hybrid neurons with generalized yager operators, pp. 2270–2274.

Kendall, M. (1953), 'The analysis of economic time series-part i:prices', *Journal of the Royal Statistical Society* **116**(1), 11–34.

Kirman, A. P. (1992), 'Whom or what does the representative individual represent?', *Journal of Economic Perspectives* **6**(2), 117–136.

Kittler, J., Hatef, M., Duin, R. P. W. & Matas, J. (1998), 'On combining classifiers', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **20**(3), 226–239.

Kleinberg, E. M. (1990), 'Stochastic descrimination', *Annals of Mathematics and Artificial Intelligence*.

Kleinberg, E. M. (2000), 'On the algorithmic implementation of stochastic descrimination', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**(5), 473–490.

Krogh, A. & Vedelsby, J. (1995), 'Neural network ensembles, cross validation, and active learning'.

Kuncheva, L. & Jain, L. C. (2000), 'Designing classifier fusion systems by genetic algorithms', *IEEE Trans. on Evolutionary Computation* **4**(4), 327–336.

Kyrtsoua, C. & Serletisb, A. (2006), 'Univariate tests for nonlinear structure', *Journal of Macroeconomics* **28**(1), 154–168.

La Place, (Marquis de), P. S. (1820), *Theorie analytique des probabilites*, third edn, M.me V.e Courcier, Paris.

Laibson, D. (1997), 'Golden eggs and hyperbolic discounting', *Quarterly Journal of Eco- nomics* **62**, 443–477.

Lee, C. L. (2001), 'Trading rule profits in latin american currency spot rates.', *International Review of Financial Analysis* **10**, 135–156.

LeRoy, S. F. (1973), 'Risk aversion and the martingale property of stock prices', *International Economic Review* **14**(2), 436–446.

LeRoy, S. F. & Porter, R. D. (1981), 'The present-value relation: Tests based on implied variance bounds', *Econometrica* **49**(3), 555–574.

Levich, R. M. & Thomas, L. R. (1993), 'The significance of technical trading rules in the fx market: a bootstrap approach', *Journal of Money and Finance* (12), 451–474.

Levin, E. J. & Wright, R. E. (1999), 'Why does the bid-ask spread vary over the day?', *Applied Economic Letters* **6**, 563–567.

Levin, E. J. & Wright, R. E. (2004), 'Estimating the profit markup component of the bid-ask spread: Evidence from the london stock exchange', *Quarterly Review of Economics and Finance* **44**, 1–19.

Lichtenstein, S., Fischoff, B. & Phillips, L. (1982), 'Calibration of probabilities: The state of the art to 1980', *D. Kahneman, P. Slovic, and A. Tversky (eds.), Judgment Under Uncertainty: Heuristics and Biases.*

Lintner, J. (1965), 'The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets', *The Review of Economics and Statistics* **47**(1), 13–37.

Lo, A. (2004), 'The adaptive markets hypothesis: Market efficiency from an evolutionary perspective', *Journal of Portfolio Management* **30**, 15–29.

Lo, A. & MacKinlay, A. (1990), 'Data-snooping biases in tests of financial asset pricing models', *Review of Financial Studies.*

Lo, A. W. (2000), 'Foundations of technical analysis: Computational algorithms, statistical inference, and empirical implementation', *The Journal of Finance.*

Lo, A. W. & MacKinlay, A. C. (1988), 'Stock market prices do not follow random walks: Evidence from a simple specification test', *The Review of Financial Studies* **1**(1), 41–66.

Lucas, R. E. (1978), 'Asset prices in an exchange economy', *Econometrica* **46**(6), 1429–1445.

Lukac, L., Brorsen, B. & Irwin, S. (1989), 'The usefulness of historical data in selecting parameters for technical trading systems', *Journal of Futures Markets* **9**, 55–65.

Lukac, L. P., Brorsen, B. W. & Irwin, S. H. (1988), 'A test of futures market disequilibrium using twelve different technical trading systems.', *Applied Economics* **20**, 623–639.

Mandelbrot, B. B. (1963), 'The variation of certain speculative prices', *Journal of Business* **36**(4), 394–419.

Mandelbrot, B. B. (1966), 'Forecasts of future prices, unbiased markets, and martingale models.', *Journal of Business* **39**, 242–255.

Marshall, B., Cahan, R. & Cahan, J. (2008), 'Can commodity futures be profitably traded with quantitative market timing strategies', *Article in Press, Journal of Banking and Finance.*

Mayoraz, E. & Moreira, M. (1997), 'On the decomposition of polychotomies into dichotomies', *In The XIV International Conference on Machine Learning* pp. 219–226.

Miffre, J. & Rallis, G. (2007), 'Momentum strategies in commodity futures markets', *Journal of Banking and Finance* **31**(6), 1863–1886.

Mitchell, T. M. (1980), 'The need for biases in learning generalisations'.

Mitchell, T. M. (1997), *Machine Learning*, The McGraw-Hill Companies, Inc.

Monteiro, A. (2001), Interest rate curve estimation: A support vector regression application to finance, Technical report, Princetown University.

Mossin, J. (1966), 'Equilibrium in a capital asset market', *Econometrica* **34**(4), 768–783.

Mukherjee, S., Tamayo, P., Mesirov, J., Slonim, D., Verri, A. & Poggio, T. (1999), Support vector machine classification of microarray data, Technical Report 182, AI Memo 1676, CBCL.

Muller, K.-R., Smola, A. J., Ratsch, G., Schokopf, B., Kohlmorgen, J. & Vapnik, V. (1999), 'Using support vector machines for time series prediction', pp. 243–253.

Munro, R., Ler, D. & Patrick, J. (n.d.), 'Meta-learning orthographic and contextual models for language independent named entity recognition', *IEEE Transactions on Pattern Analysis and Machine Intelligence.*

Murphy, J. J. (1999), *Technical Analysis of the Financial Markets*, New York Institute of Finance.

Naik, N. Y. & Yadav, P. K. (1999*a*), 'The effects of market reform on trading costs of public investors: Evidence from the london stock exchange', *Institute of Finance and Accounting Working Paper No. 296. London Business School.*

Naik, N. Y. & Yadav, P. K. (1999*b*), 'Execution costs and order flow characteristics in dealership markets: Evidence from the london stock exchange', *Unpublished working paper, London Business School and University of Strathclyde.*

Neely, C., Weller, P. & Dittmar, R. (1997), 'Is technical analysis in the foreign exchange market profitable?: A genetic programming approach', *Journal of Financial and Quantitative Analysis* **32**(4), 405–426.

Neftci, S. N. (1991), 'Naive trading rules in financial markets and wiener-kolmogorov prediction theory: A study of "technical analysis."', *The Journal of Business.*

Nilson, S. (1991), *Japanese candlestick charstick charting techniques.*, NY Institute of Finance, New York.

Oberlechner, T. (2001), 'Importance of technical and fundamental analysis in the european foreign exchange market', *International Journal of Finance and Economics* **6**, 81–93.

Odean, T. (1998), 'Are investors reluctant to realize their losses?', *Journal of Finance* **53**, 1775–1798.

Opitz, D. & Maclin, R. (1999), 'Popular ensemble methods: An empirical study', *Journal of Artificial Intelligence Research* **11**, 169–198.

Osborne, M. F. M. (1959), 'Brownian motion in the stock market', *Operations Research* **7**, 145–173.

Osler, C. L. (2000), 'Support for resistance: Technical analysis and intraday exchange rates.', *Economic Policy Review* **6**, 53–65.

Park, C.-H. & Irwin, S. H. (2004), 'The profitability of technical analysis: A review', *AgMAS Project Research Report*.

Park, C.-H. & Irwin, S. H. (2005), 'The profitability of technical trading rules in us futures markets; a data snooping free test.', *AgMAS Project Research Report*.

Parmanto, B., Munro, P. W. & Doyle, H. R. (1996), 'Improving committee diagnosis with resampling techniques', *In Touretzky, D. S., Mozer, M. C., & Hesselmo, M. E. (Eds.), Advances in Neural Information Processing Systems* **8**, 882–888.

Pesaran, H. M. & Timmermann, A. (2003), How costly is it to ignore breaks when forecasting the direction of a time series?, Cambridge Working Papers in Economics 0306, Faculty of Economics (formerly DAE), University of Cambridge.

Poterba, J. M. & Summers, L. H. (1988), 'Mean reversion in stock prices: Evidence and implications', *Journal of Financial Economics* **22**(1), 27–59.

Pring, M. J. (2002), *Technical Analysis Explained: The Successful Investor's Guide to Spotting Investment Trends and Turning Points*, fourth edn, McGraw-Hill.

Qi, M. & Wu, Y. (2002), 'Technical trading-rule profitability, data snooping, and reality check: Evidence from the foreign exchange market.', *Working paper.*

Roberts, H. (1967), Statistical versus clinical prediction of the stock market, unpublished manuscript.

Roberts, H. V. (1965), 'Probabilistic prediction', *Journal of the American Sta- tistical Association* **60**, 50–60.

Rogova, G. L. (1994), 'Combining the results of several neural network classifiers', *Neural Networks* **7**(5), 777–781.

Rooney, N., Patterson, D., Tsymbal, A. & Anand, S. (2004), Bio-molecular cancer prediction with random subspace ensembles of support vector machines, Technical report.

Rosenthal, R. (1979), 'The "file drawer problem" and tolerance for null results', *Psychological Bulletin* **86**(3), 638–641.

Samuel, A. L. (1959), 'Some studies in machine learning using the game of checkers', *IBM Journal of Research and Development* **3**(3), 210–229.

Samuelson, P. A. (1965), 'Proof that properly anticipated prices fluctuate randomly', *Industrial Management Review* **6**, 41–49.

Schapire, R. E. (1990), 'The strength of weak learnability', *Machine Learning*.

Scholkopf, B., Burges, C. & V.Vapnik (1995), Extracting support data for a given task, *in* 'In Proceedings First International Conference on Knowledge Discovery and Data Mining', AAAI Press, pp. 252–257.

Schwager, J. D. (1996), *Schwager on Futures: Technica Analysis.*, John Wiley & Sons, New York.

Schwarz, G. (1978), 'Estimating the dimension of a model', *The Annals of Statistics* **6**(2), 461–464.

Sewell, M. (2007), Efficient market hypothesis - history, *in* 'Efficient Market Hypothesis - Empirical Proof From India', ICFAI University Press, Panjagutta, Hyderabad, India. In press.

Sewell, S. P., Stansell, S. R., Lee, I. & Below, S. D. (1996), 'Using chaos measures to examine international capital market integration', *Applied Financial Economics* **6**(2), 91–101.

Shaffer, J. P. (1995), 'Multiple hypothesis testing', *Annual Review of Psychology* **46**, 561–584.

Sharpe, W. F. (1964), 'Capital asset prices: A theory of market equilibrium under conditions of risk', *The Journal of Finance* **19**(3), 425–442.

Sharpe, W. F. (1965), 'Mutual fund performance', *The Journal of Business* **39**(1), 119–138.

Shefrin, M. & Statman, M. (1985), 'The disposition to sell winners too early and ride losers too long: Theory and evidence', *Journal of Finance* **40**, 777–790.

Shiller, R. J. (1981), 'Do stock prices move too much to be justified by subsequent changes in dividends?', *The American Economic Review* **71**(3), 421–436.

Simon, H. (1955), 'A behavioral model of rational choice', *Quarterly Journal of Economics* **69**, 99–118.

Skurichina, M. & Duin, R. P. W. (2002), 'Bagging, boosting and the random subspace method for linear classifiers', *Pattern Analysis and Applications* **5**(2), 121–135.

Smidt, S. (1965), 'A test of serial independence of price changes in soybean futures', *Food Research Institute Studies* **5**, 117–136.

Smola, A. & Scholkopf, B. (1998), 'A tutorial on support vector regression', *NeuroCOLT2 Technical Report NC2-TR-1998-030.*

Stigler, S. M. (1973), 'Studies in the history of probability and statistics xxxii : Laplace, fisher, and the discovery of the concept of sufficiency', *Biometrika* **60**, 439–445.

Stone, M. (1978), 'Cross-validatory choice and assessment of statistical predictions', *Journal of the Royal Statistical Society, Series B* **36**(2), 111–147.

Sullivan, R., Timmermann, A. & White, H. (1999), 'Data-snooping, technical trading rule performance, and the bootstrap', *Journal of Finance*.

Taylor, N. (2002), 'The economic and statistical significance of spread forecasts: Evidence from the london stock exchange', *Journal of Banking and Finance* **29**, 795–818.

Taylor, S. J. (1994), 'Trading futures using a channel rule: A study of the predictive power of technical analysis with currency examples.', *Journal of Futures Markets* **14**, 215–235.

Timmermann, A. (2008), 'Elusive return predictability', *International Journal of Forecasting* **24**(1), 1–18.

Timmermann, A. & Granger, C. W. J. (2004), 'Efficient market hypothesis and forecasting', *International Journal of Forecasting* **20**(1), 15–27.

Tomek, W. G. & Querin, S. F. (1984), 'Random processes in prices and technical analysis', *Journal of Futures Markets* **4**, 15–23.

Tresp, V. (2001), 'Committee machines', *In Yu Hen Hu and Jenq-Nen Hwang, editors, Handbook for Neural Network Signal Processing* pp. 993–1001.

Tumer, K. & Oza, C. N. (1999), 'Decimated input ensembles for improved generalization', *In Proc. of the Int. Joint Conf. on Neural Networks (IJCNN-99).*

Tversky, A. & Kahneman, D. (1974), 'Judgment under uncertainty: Heuristics and biases', *Science* **185**(4157), 1124–1131.

Tversky, A. & Kahneman, D. (1981), 'The framing of decisions and the psychology of choice', *Science* **211**, 453–458.

Valentini, G. & Masulli, F. (2002), Ensembles of learning machines., *in* 'WIRN', pp. 3–22.

Vanceka, J. & snd S Alya, I. V. (2009), 'Fuzzy aggregation and averaging for group decision making: A generalization and survey', *Knowledge-Based Systems* **22**, 79–84.

Vapnik, V. & Chervonenkis, A. (1964), 'A note on one class of perceptrons', *Automation and Remote Control.*

Vapnik, V., Golowich, S. & Smola, A. (1997), Support vector method for function approximation, regression estimation, and signal processing, *in* 'In Advances in Neural Information Processing Systems', pp. 281–287.

Vapnik, V. N. (1995), *The Nature of Statistical Learning Theory*, Springer-Verlag New York, Inc.

Vapnik, V. N. (1998), *Statistical Learning Theory*, John Wiley & Sons, Inc.

Vapnik, V. N. (2000), *The Nature of Statistical Learning Theory*, Springer-Verlag New York, Inc.

Vapnik, V. N. & Chervonenkis, A. Y. (1974), *Teoriya raspoznavaniya obrazov: Statisticheskie problemy obucheniya. (Russian) [Theory of pattern recognition: Statistical problems of learning]*, Nauka, Moscow.

Varshney, P. K. (1996), 'Distributed detection and data fusion'.

Vince, R. (1999), *The Mathematics of Money Management: Risk Analysis Techniques for Traders*, Wiley.

Weiss, S. M. & Kulikowski, C. A. (1991), *Computer Systems that Learn*, Morgan Kaufmann, San Mateo, CA.

Wherry, R. J. (1931), 'A new formula for predicting the shrinkage of the coefficient of multiple correlation', *The Annals of Mathematical Statistics* **2**(4), 440–457.

White, H. L. (2000), 'A reality check for data snooping', *Econometrica* **68**(5), 1097–1126.

Witten, I. & Frank, E. (1999), *Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufmann.

Wolpert, D. H. (1992), 'Stacked generalization', *Neural Networks* **5**(2), 241–259.

Wolpert, D. H. & Macready, W. G. (1995), 'No free lunch theorems for search'.

Working, H. (1934), 'A random difference series for use in the analysis of time series', *Journal of the American Statistical Association* **29**, 11–24.

Wyckoff, R. D. (1910), *Studies in Tape Reading.*, Fraser Pub. Co.

Yang, H., Chan, L. & King, I. (2002), Support vector machine regression for volatile stock market prediction., *in* 'IDEAL', pp. 391–396.

Yang, S. & Brorsen, W. (1993), 'Nonlinear dynamics of daily futures prices: Conditional heteroskedasticity or chaos?', *The Journal of Futures Markets* **13**(2), 175–191.

Zenobi, G. & Cunningham, P. (2001), Using diversity in preparing ensembles of classifiers based on different feature subsets to minimize generalization error, *in* 'EMCL '01: Proceedings of the 12th European Conference on Machine Learning', Springer-Verlag, London, UK, pp. 576–587.

Zhang, X., Mesirov, J. P. & Waltz, D. L. (1992), 'Hybrid system for protein secondary structure prediction', *Journal of Molecular Biology* **225**(30), 1049–1063.

Zhao, F., Tang, X., Lin, F., Samson, S. & Remsen, A. (2005), Binary plankton image classification using random subspace, *in* 'IEEE International Conference on Image Processing, ICIP 2005.', Vol. 1, IEEE, pp. I: 357–360.