DISS. ETH NO. -

Nonlinear ensemble models to predict oil reserves and stock market returns in the presence of inherent uncertainty

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES of ETH ZURICH

(Dr. sc. ETH Zurich)

presented by

LUCAS FIEVET

MSc ETH Physics

born on 01.01.1989

citizen of Germany

accepted on the recommendation of

Prof. Dr. Didier Sornette,

Prof. Dr. Marc Paolella

2017

Acknowledgments

First of all, I would like to thank my mother, Lieselotte, who has generously supported my bachelor and master studies. Without her support, I could never have dedicated myself to the study of theoretical physics, and get a glimpse of the theories that attempt to explain our universe. The mathematical tools, I have acquired in this quest of understanding our universe, have been and will be invaluable.

After this short disclaimer, I particularly want to thank my supervisor, Prof. Didier Sornette, for taking me as his PhD student. During these three years he has been an endless source of inspiration, and relentlessly followed up with detailed insights and important literature. He unconditionally took the time to review my lengthy drafts, even in the early stages. Most of all, I am grateful for his patience and support in the moments where my work wandered a bit away from the core topic, as this freedom was crucial in acquiring the interdisciplinary knowledge that allowed me to make subsequent progress on the main research questions. Also, I wish to express my thanks to Prof. Marc Paolella for being my co-examiner, and Prof. Sebastian Rausch for chairing my defense. I appreciate their time and disposition to evaluate my thesis.

I would also like to thank current and former members of the Chair of Entrepreneurial Risks and MTEC department at ETHZ. In particular, I would like to mention Diego Ardila, Peter Cauwels, Vladimir Filimonov, Herman Hellenes, Bjarni Jónsson, Sindri Kolbeinsson, Yannick Malevergne, Benjamin Vandermarliere, Qunzhi Zhang, Donnacha Daly, Guilherme Demos, Zalàn Forró, Zhuli He, Afina Inina, Tatyana Kovalenko, Dorsa Sanadgol, Hyun-U Sohn, Spencer Wheatley, Richard Senner, Ralf Kohrt, Rebekka Burkholz, Tobias Huber, and Michael Schatz. It has always been a pleasure to exchange ideas, give and receive input, and provide each other with motivation. Special thanks goes as well to Isabella Bieri, Heidi Demuth, and Judith Holzheimer, for their wonderful support in administrative matters.

Finally, I want to thank my girlfriend Zhiying, who stood by me during the PhD, and provided important statistical knowledge that shaped the later stages of the thesis.

Abstract

This is a cumulative thesis of three self-contained research papers and some additional unpublished material. The first paper has been published, while the second and third have been submitted and are awaiting review. The research revolves around the calibration of nonlinear ensemble models to forecast oil production and daily stock market returns. A particular emphasize is put on the validation through back-testing for the obtained forecasts, including the computation of statistical significance levels when possible.

Chapter two presents the generalized methodology we developed to forecast the oil production of the UK and Norway from 2014 to 2030, based on the publication Fievet et al. (2015). The methodology extrapolates the production of individual oil fields using a stretched exponential, and estimates the extrapolation error from back-tests. The production resulting from future discoveries of giant and dwarf oil fields is estimated using Monte-Carlo simulations. The forecast validation, using the time period 2008-2014 as out-of-sample, showed that the ensemble method, aggregating individual fields, was much more robust then the Hubbert model applied to aggregate production.

Chapter three defines the efficient market hypothesis, and introduces the performance benchmarks subsequently used to assess the evaluated statistical learning and agent based strategies. A robust test for the statistical significance of directional accuracy is presented, as well as an approximation to the relationship between directional accuracy and trading performance. The break-even transaction cost for unit and compounded investment is put into relationship with trading frequency and directional accuracy, and compared to typical transaction costs on equity indices. The computation of multiple testing adjusted p-values of trading strategies is presented in detail. A simple best Sharpe ratio portfolio strategy is presented, which is used in the assessment of the ex-ante performance of a universe of available strategies. Last, the four-factor regression model is introduced to regress newly found anomalies on known anomalous factors.

Chapter four presents the particular type of agent based models studied in this thesis. The four different types of games played by the agents, namely the majority game, delayed majority game, minority game, and delayed minority game are analyzed separately in a first step. To evaluate the predictive performance without multi-agent interactions, each type of game is reverse engineered on major equity indices for a single agent model. In a second step, the reverse-engineering results of the mixed game ABM with many agents are presented. The multi-agent model was found to have no predictive superiority to the single agent model. Therefore, the analysis of the predictive power of single agent models, and the stepwise construction of two agent models, was the natural path for the further research in this thesis.

Chapter five formalizes the relationship between the binary strategies of the agents and decision trees. The abnormal predictive performance of a single decision tree motivated the analysis of further statistical learning methods such as logistic regression, support vector machine, gradient boosting, nearest neighbors, linear discriminant analysis, and quadratic

discriminant analysis. Theoretical data generating processes are used to illustrate the limitations of regressive models and advantages of statistical learning models. Some of the statistical learning models were found to have statistically significant predictive performance on the S&P 500, FTSE and CSI 300. In particular, the predictability was found to be particularly high during the crash phase of the dotcom bubble, the financial crisis, the European debt crisis, and the recent Chinese stock market turbulence. The break-even transaction costs of up to 20 bps per round trip, on a 20 year time period, question the weak efficient market hypothesis.

Finally, chapter six introduces an extension of statistical learning towards agent based modeling. In particular, it is shown how multiple instances (agents) of an arbitrary predictor (e.g. a decision tree or binary strategy) can be combined to predict a given time series. The experiments are limited to two agent models, which can still meaningfully be calibrated. Models with more agents would require more data, which would be available in the context of intraday trading or order book prediction. However, the later cases are out-of-scope for this thesis.

Chapter seven gathers ideas for future research and concludes the thesis.

Zusammenfassung

Dies ist eine kumulative Doktorarbeit, zusammengestellt aus drei eigenständigen Forschungsarbeiten und zusätzlichem unveröffentlichtem Inhalt. Die erste Forschungsarbeit wurde bereits veröffentlicht während die zweite und dritte Forschungsarbeit eingereicht wurden. Die Forschung befasst sich mit der Kalibrierung von nichtlinearen Ensemblemodellen zur Vorhersage von Ölproduktion und Börsenmärkten. Ein besonderer Schwerpunkt wird auf die Validierung der Prognosen gelegt, insbesondere der Berechnung der statistischen Signifikanz, wenn möglich.

Kapitel zwei stellt die verallgemeinerte Methodik vor, die wir entwickelt haben, um die Ölproduktion von Großbritannien und Norwegen von 2014 bis 2030 vorherzusagen, basierend auf der Veröffentlichung Fievet et al. (2015). Die Methodik extrapoliert die Produktion von einzelnen Ölfeldern unter Verwendung einer gestreckten Exponentialfunktion und schätzt den Extrapolationsfehler ab anhand von vergangenen Daten. Die Produktion, die aus zukünftigen Entdeckungen von Riesen- und Zwergölfeldern resultiert, wird mit Monte-Carlo-Simulationen geschätzt. Die Validierung über den Zeitraum 2008-2014 zeigte, dass die Ensemble-Methode, die einzelne Felder aggregierte, viel zuverlässiger ist als das Hubbert-Modell angewandt auf die aggregierte Produktion.

Kapitel drei definiert die effiziente Markthypothese und führt die Performance Benchmarks ein, die später zur Bewertung der evaluierten statistischen lern- und agentenbasierten Strategien verwendet wurden. Es wird ein robuster Test für die statistische Signifikanz der Richtungsgenauigkeit sowie eine Annäherung an die Beziehung zwischen Richtungsgenauigkeit und Handelsleistung dargestellt. Die Nullsumme-Transaktionskosten der Strategien werden in Beziehung mit Handelsfrequenz und Richtungsgenauigkeit gebracht, und mit typischen Transaktionskosten verglichen. Die Berechnung von p-Values von Handelsstrategien angepasst für multiple Tests wird detailliert dargestellt. Eine einfache Portfolio-Strategie wird vorgestellt, um die Ex-Ante-Leistung eines Universums verfügbarer Strategien zu ermitteln. Zuletzt wird das Vier-Faktor-Regressionsmodell präsentiert, um neue Anomalien auf Zusammenhänge mit bekannten anomalen Faktoren zu prüfen.

Kapitel vier präsentiert die spezifische Art von agentenbasierten Modellen, die in dieser Arbeit untersucht werden. Die vier verschiedenen Arten von Spielen, die von den Agenten gespielt werden, nämlich das Mehrheitsspiel, das verzögerte Mehrheitsspiel, das Minderheitsspiel und das verzögerte Minderheitsspiel werden in einem ersten Schritt separat analysiert. Jede Art von Spiel wird auf realen Daten für ein Ein-Agent-Modell kalibriert, um die prädiktive Leistung ohne Multi-Agenten-Interaktionen zu bewerten. In einem zweiten Schritt werden die prädiktive-Ergebnisse des Multi-Agenten Modell mit gemischten Spielen präsentiert. Es wurde festgestellt, dass das Multi-Agenten-Modell keine Verbesserung gegenüber dem Ein-Agent-Modell aufweist. Daher waren die Analyse der prädiktiven Fähigkeiten von Ein-Agent-Modellen und der schrittweise Aufbau von zwei Agentenmodellen der natürliche Weg für die weitere Forschungsarbeit.

Kapitel fünf formalisiert die Beziehung zwischen den binären Strategien der Agenten und Entscheidungsbäume. Die abnorme prädiktive Leistung eines einzelnen Entscheidungsbaums motivierte die Analyse weiterer statistischer Lernmethoden wie logistische Regression, support vector machine, gradient boosting, nearest neighbors, lineare und quadratische Diskriminanzanalyse. Theoretische Datenerzeugungsprozesse dienen der Veranschaulichung der Einschränkungen von regressiven Modellen und Vorteilen statistischer Lernmodelle. Einige der statistischen Lernmodelle zeigten eine statistisch signifikante prädiktive Leistung auf dem S&P 500, FTSE und CSI 300. Besonders signifikant war die Vorhersagbarkeit während der Dotcom-Blase, der Finanzkrise, und der Europäischen Schuldenkrise und der jüngsten chinesischen Börsen-Turbulenzen. Die Nullsummen-Transaktionskosten von bis zu 20 bps, in einem Zeitraum von 20 Jahren, stellen die schwache effiziente Markthypothese in Frage.

Das Kapitel sechs erstellt zwei mögliche Erweiterungen der statistischen Lern-Methoden die man als agentenbasierten Modellieren bezeichnen kann. Insbesondere wird gezeigt, wie mehrere Instanzen (Agenten) eines beliebigen Prädiktors (z. B. ein Entscheidungsbaum oder eine binäre Strategie) kombiniert werden können, um eine gegebene Zeitreihe vorherzusagen. Die Kalibrierungs-Experimente beschränken sich auf die zwei und drei Agenten-Modelle. Modelle mit mehr Agenten können nicht sinnvoll kalibriert werden da sie mehr Daten benötigen.

Kapitel sieben sammelt Forschung-Ideen und schließt die Doktorarbeit ab.

Contents

Acknowledgments i
Abstract ii
Zusammenfassung iv

1 Introduction 1
1.1 Predictive limitations of regression models 4
1.2 A critique of statistical learning in finance 6
1.3 Calibration issues in agent based models 7
1.4 Mitigating the replication crisis 8
1.5 Chapter contributions 10

2 General methodology to forecast oil production 13
2.1 Introduction 13
2.2 State of the art 14
  2.2.1 Hubbert's Model for peak oil forecasting 14
  2.2.2 Back-testing Hubbert's Model 15
  2.2.3 Beyond Hubbert's Model 17
2.3 Methodology 17
  2.3.1 Extending the oil production of individual fields 17
    2.3.1.1 Regular, irregular and new fields 20
    2.3.1.2 Back-testing & Error 21
    2.3.1.3 Aggregate error 22
  2.3.2 Discovery rate of new fields 23
    2.3.2.1 Discoveries modeled as logistic growth 23
    2.3.2.2 Likelihood function for the number of discoveries 27
    2.3.2.3 Future production from discoveries 28
  2.3.3 Contributions and applicability of the methodology 30
    2.3.3.1 Contributions 30
    2.3.3.2 Applicability 31
2.4 Results 31
2.5 Validation 33
2.6 Conclusion 36
2.7 Three years later 37

3 Benchmarks for the Efficient Market Hypothesis 39
- 3.1 Defining the efficient market hypothesis 39
- 3.2 The difficulty of distinguishing true skill from luck 43
- 3.3 Performance metrics 46
  - 3.3.1 Directional accuracy 46
  - 3.3.2 Directional accuracy & compounded returns 51
  - 3.3.3 Directional accuracy & mean return 53
  - 3.3.4 Sharpe ratio 54
    - 3.3.4.1 Studentized Sharpe ratio test statistic 55
    - 3.3.4.2 Robust covariance estimation 55
- 3.4 Statistical significance & multiple testing 56
  - 3.4.1 Methodology setup 56
  - 3.4.2 Null bootstrap of returns 57
  - 3.4.3 Adjusting p-values for multiple testing 58
  - 3.4.4 Null bootstrap of trading signals 58
  - 3.4.5 P-value bias of the mean return 60
- 3.5 Portfolio of strategies 61
- 3.6 Factor regression tests 63
- 3.7 Testing the efficient market hypothesis: an overview 63

4 Agent Based Models with Binary Strategies 65
- 4.1 Motivation & review 65
- 4.2 Agent behavior 68
  - 4.2.1 Strategy space 68
  - 4.2.2 Minority & majority payoff 69
  - 4.2.3 Strategy performance 69
  - 4.2.4 Memory length 70
  - 4.2.5 Predictor summary 70
  - 4.2.6 De Bruijn representation 71
  - 4.2.7 Calibration optimization 73
  - 4.2.8 Relations between games 73
  - 4.2.9 Viable cycle predictors 74
- 4.3 Mixed game agent based model 75
  - 4.3.1 Agent heterogeneity 75
  - 4.3.2 Agent trading threshold 76
  - 4.3.3 Aggregate prediction 77
  - 4.3.4 Mixed-game model summary 77
  - 4.3.5 Calibration 77
- 4.4 Results 80
- 4.5 Conclusion 86

5 From Agent Based Modeling to Time Series Learning 87
5.1 Agent as decision tree . . . 87
5.2 Motivating tree based forecasting . . . 89
5.2.1 Linear filter models & technical trading rules . . . 89
5.2.2 From autoregressive to tree based models . . . 90
5.2.3 Regression tree algorithm . . . 92
5.2.4 Classification tree . . . 93
5.2.5 Fixed tree . . . 94
5.2.6 Probit models & polynomial features . . . 94
5.3 Comparing the forecasting power of autoregressive & tree models . . . 95
5.3.1 Connecting fixed trees to Markov chains . . . 95
5.3.2 Binary Markov based processes . . . 96
5.3.3 Expected autoregressive forecast for binary Markov processes . . . 97
5.3.4 Expected fixed tree forecast for autoregressive processes . . . 99
5.3.5 Simulation study . . . 100
5.3.5.1 Setup . . . 101
5.3.5.2 Results . . . 101
5.3.6 Empirical results . . . 105
5.3.6.1 Data & parameters . . . 105
5.3.6.2 Performance of the top strategies . . . 105
5.3.6.3 Further analysis . . . 109
5.3.6.4 Conclusion . . . 111
5.4 Towards time series learning . . . 113
5.4.1 Statistical learning & nonlinear finite lag processes . . . 113
5.4.1.1 Issues in cross-validating time series data . . . 114
5.4.1.2 Calibration window length and lag selection . . . 115
5.4.1.3 Regression vs classification . . . 116
5.4.2 Constructing the universe of strategies . . . 117
5.4.2.1 Models discussion . . . 120
5.4.3 Empirical results . . . 121
5.4.3.1 Best models . . . 121
5.4.3.2 Model mean position duration . . . 126
5.4.3.3 Wealth correlation with directional accuracy . . . 127
5.4.3.4 Model correlations . . . 130
5.4.3.5 Periods of abnormal performance . . . 132
5.4.3.6 Ex-ante performance . . . 132
5.5 Conclusion . . . 135

6 Emergent Predictability in Two Agent Models 137
6.1 Introduction . . . 137
6.2 Representing Agents as Decision Trees . . . 138

6.2.1 Strategy Space 138
6.2.2 Delay 139
6.2.3 Agent as Optimal Decision Tree 139
6.2.4 Proof of Equivalence Between Agents and Decision Trees 140
6.3 Single Agent Experiments 141
6.3.1 Trading Strategies 141
6.3.2 Experiments 141
6.3.3 Anomalous Predictability 142
6.3.4 Empirical Results 143
6.3.5 Principal Component Analysis 143
6.4 Heterogeneous Two Agent Models 149
6.5 Implementation 152
6.6 Conclusion 156

7 Conclusion & Outlook **157**

Bibliography **160**

Curriculum Vitae **172**

Chapter 1

Introduction

The **E**fficient **M**arket **H**ypothesis (EMH) defined by Fama (1970) is a pillar of neoclassical financial theory stating that asset prices instantaneously and fully reflect the available information. A consequence of this assumption is that asset returns follow a martingale process after discounting for equilibrium expected returns. In other terms, no investment strategy can generate returns in excess of the equilibrium expected returns. The hypothesis exists for several degrees of available information. In its weak form, the hypothesis states that no excess profits can be made by trading based on past prices. The semi-strong form extends this statement to publicly available information such as dividends, interest rates, annual earnings or stock splits. The strong form extends the hypothesis to all existing information, in particular private information available to only a small number of insiders.

Sufficient conditions for the EMH to hold are: no transaction costs; instantaneous and cost-less availability of information to all market participants; and an identical interpretation of current information by all market participants. Needless to say that none of these assumptions hold for real markets where transaction costs can prevent arbitrage, information takes time to reach all participants and believes about future returns based on current information are heterogeneous. Nonetheless, real markets can still be efficient as long as current information is fully reflected when transactions take place, information is available to a large enough number of market participants and no possible interpretation of current information can consistently yield returns above the equilibrium expected returns.

Despite the convincing case that real markets can be efficient, Grossman and Stiglitz (1980) prove that efficiency is not an equilibrium state for markets with costly information and transactions, as there would be no compensation to the costly activities of market participants. In addition, Milgrom and Stokey (1982) prove that the participants in an efficient market would never trade because a bid must come from a participant with better knowledge and accepting it would necessarily result in a loss. In fact, the cost of information, and its heterogeneous distribution and interpretation among market participants are the drivers of the market activity. The market allows each participant to take positions based on his current believes. The constant flow of new events is quickly reflected in asset prices by the trading activities of the fastest and best informed traders. The market is in an equilibrium state of disequilibrium that maintains the constant trading activity of the market participants.

These aspects preventing real markets from reaching the theoretical definition of an efficient market have been acknowledged by Fama (1991), but he points out the large support for the EMH. A growing body of research on corporate events shows that prices efficiently adjust to new information (e.g. dividend changes). The research on return predictability based on past returns or macroeconomic factors shows little predictability (Fama, 1998). The first statistically significant anomalies, found by Fama and French (1993), origin in the size and book to market value of stocks. The stocks of companies with small market capitalization outperform on average the stocks of companies with large market capitalization. As well, stocks with high book to market value outperform on average stocks with low book to market value. A further major anomaly found by Jegadeesh and Titman (1993) and Carhart (1997) is the momentum of stocks and funds, the best performing assets continue to outperform the worst performing assets in a persistent manner. However, when using the four factor model as the benchmark for equilibrium returns, the number of funds having statistically significant excess return is vanishing (Fama and French, 2009). Therefore, the four factor model is considered to be a good benchmark with respect to which the weak EMH holds true.

The literature up to 2007, reviewed by Subrahmanyam (2008), already provides many pieces of evidence challenging the semi-strong form and strong form of the EMH. In particular, the repeating occurrence of bubbles and crashes such as the dotcom bubble, the financial crisis, ensuing “great recession” and the on-going European sovereign debt crisis, accompanied by strong bullish markets suggests even more prominently the existence of significant anomalies occurring in financial markets (Malkiel, 2003a). In particular, asset returns may exhibit transient dependence structures that are incompatible with the EMH.

The precise origin of financial bubbles is still subject to debate in classical economics, but in any case a bubble implies a significant difference between the fundamental value of an asset and the value attributed by investors. Within the scope of the EMH, such a difference should be arbitraged away by the rational market participants, as the market price always reflects the fundamental asset value. To explain such discrepancies between the fundamental value and market value, the theory of heterogeneous belief bubbles studied markets where agents disagree about the fundamental value. The heterogeneity in beliefs can origin from psychological biases or inherent uncertainty in estimating future cash-flow. In a framework that limits short selling, the disagreement among market participants about the fundamental value is sufficient to generate equilibrium prices higher than the average expected value (Miller, 1977). The above average equilibrium price results from the pessimists staying out of the market, as they cannot short sell the asset. Further on, Harrison and Kreps (1978) showed that in a dynamic framework of heterogeneous believes the asset price can exceed the valuation of the most optimistic market participant, as a result of speculative behavior. Scheinkman and Xiong (2003) extend this dynamic framework to continuous time, and find that bubble are characterized by high trading volumes, a phenomena often observed on stock markets.

A major obstacle in rejecting the EMH is that for a large number of trading strategies some will outperform the benchmark by pure luck. To obtain a meaningful result, the statistical significance of the top performing strategy has to be adjusted for multiple testing with respect to the full universe of considered strategies. Among the first to achieve this feat was White (2000) with the Reality Check method, which allowed him to show that in the universe of technical trading rules none yields significant trading performance on the S&P 500. This result is confirmed by Sullivan et al. (1999) for a 100 year trading period on the Dow Jones Industrial, but some significantly performing technical trading rules on emerging markets are found by Hsu et al. (2010). The multiple testing framework has become a robust methodology to compute the statistical significance of a model, adjusting for data snooping in a universe of models. The methodology has been extended by Romano and Wolf (2005a) to reject as many null hypothesis as possible without violating the family wise error rate, and by Hansen (2005) to reduce sensitivity to poorly performing strategies.

The finding of a trading strategy that significantly outperforms the benchmark is a necessary condition to reject the EMH, but not sufficient because it may have been impossible to select the winning strategy ex-ante. The definition of the EMH proposed by Timmermann and Granger (2004) circumvents this issue by limiting the efficiency statement to a finite universe of trading models and requiring the existence of a search technology that would have selected an out-performing strategy ex-ante.

These observations motivated an extensive research in recent years to understand and predict the anomalies that can occur in asset prices. Within the scope of this research, an important tool is agent-based modeling (aka ABM), also known as agent-based computational economics (ACE). ABM provides an alternative to equilibrium models, because it relaxes some of their restrictive assumptions by adopting a bounded rationality framework (Simon, 1955, Rubinstein, 1997), which allows for heterogeneity in the preferences and skills of the different agents. This characteristic of ABM is crucial to encompass the possibility of a transient out-of-equilibrium dynamical view of the world. Further on, these models give rise to complex emergent phenomena at the macro level despite well defined behavioral rules of the agents at the micro level. As explained for instance by Bonabeau (2002), the attributes of the system's emergent phenomena cannot be mapped to its individual entities. The phenomena that cannot occur in classical equilibrium models, namely financial bubbles, market instabilities, crises and regime shifts, emerge endogenously at the macro level in ABMs. By its structure, ABM is an ideal tool to study complex interactions between agents as they occur in real stock markets. We refer to Hommes (2006), Hommes and Wagener (2009), Chiarella et al. (2009) and Evstigneev et al. (2009) for reviews on agent-based models from different perspectives. The ABM applications by Satinover and Sornette (2012a,b) and Andersen and Sornette (2005) to predict stock market returns have revealed significant predictability that requires further investigation.

Unfortunately, ABM often remains a tool to qualitatively understand how phenomena emerge at the macro level, while their calibration to real data is often unsolved. The

ABMs typically have too many degrees of freedom due to their many parameters and overfit the limited amount of available data. In contrast, a field that has been thriving in calibrating complex non-parametric models to real data is statistical learning. The regularization, cross-validation and bootstrap techniques developed in statistical learning (Hastie et al., 2001) have allowed researchers to find the maximal model flexibility (i.e. degrees of freedom) that can robustly be calibrated to a limited dataset. These techniques have promising applications to the calibration of ABMs.

Statistical learning models too are well suited to describe non-parametric dependence structures potentially present in financial returns. This has sparked a large search for predictability on stock markets using support vector machine (e.g. Cao and Tay (2001)), neural networks (e.g. Guresen et al. (2011)), and many other methods. Unfortunately, these studies often lack robust statistical tests and ex-ante selection of the winning strategy, which are needed to reject the EMH at a certain significance level.

The reviewed literature shows three major research gaps that can be addressed. First, the regression and equilibrium models most commonly used in econometrics are unsuited to capture the non-linear emerging phenomena that arise during market bubbles and crashes. Agent based modeling and the non-parametric models developed in statistical learning are better suited to capture non-linear dependence structures in asset returns. Second, the robust statistical tests developed in the finance literature to reject the EMH are not or insufficiently used in the ABM and statistical learning literature. The consequence is that a large number of studies reporting violations of the EMH when using ABMs or statistical learning lack robustness and need further investigation with better statistical tests. Third and last, the methodology developed in statistical learning to control for overfitting presents opportunities in agent based modeling that have not yet been fully exploited.

1.1 Predictive limitations of regression models

Wold's decomposition theorem (Mills and Markellos, 2008, Hamilton, 1994) states that every weakly stationary, purely non-deterministic stochastic process $\{X_t\}$ can be written as a linear filter

$$X_t = \sum_{i=0}^{\infty} \psi_i a_{t-i} = \psi\left(B\right) a_t, \tag{1.1}$$

where the $\{a_t : t = 0, \pm 1, \pm 2, \ldots\}$ are i.i.d. innovations drawn from a distribution with zero mean and constant variance, ψ_i are the coefficients of the infinite polynomial $\psi\left(B\right)$, and B is the back-shift operator. All theoretical time series models defined as a finite-order stochastic difference equation derive from this general concept of a linear filter with infinite lag. Common models are the Auto-Regressive model AR(ϱ), the Moving Average MA(ϱ), and the Auto-Regressive Moving Average model ARMA(ϱ, q). Models with non-stationarity in mean or variance can be built by assuming some stationary model for the mean or variance. Common models for the non-stationary mean or variance are the Auto-

Regressive Integrated Moving Average model ARIMA(ϱ, q), the Auto-Regressive Conditional Heteroskedastic model ARCH(ϱ, q), and the Generalized Auto-Regressive Conditional Heteroskedastic model GARCH(ϱ, q).

Non-linear stochastic processes arise when their representation is obtained by some non-linear function, for example polynomial dependencies in the innovation, asymmetric innovations, or correlated innovations. As discussed by Mills and Markellos (2008, Chap. 6), testing for non-linearity is a challenging task and is not possible in general when the functional form of the non-linearity is unknown. As well, deterministic patterns intertwined with a stochastic process can go undetected using linear filter models. In particular, this thesis will examine a range of stock return sign predictability that goes undetected using regression models.

The shortcomings of regressive models can be illustrated with the two argument exclusive-OR function $XOR\,(a,\, b)$ that returns true ($= 1$) when exactly one of the arguments is true and false ($= -1$) otherwise. An example of XOR like data is

$$\mathbf{X} = \left\{ \left(\left(1,\, 1\right),\, -1 \right),\, \left(\left(1,\, -1\right),\, 1 \right),\, \left(\left(-1,\, 1\right),\, 1 \right),\, \left(\left(-1,\, -1\right),\, -1 \right) \right\}, \tag{1.2}$$

where the four samples are assumed to be independent. The notation of Equation (1.2) is taken from the statistical learning literature (James et al., 2014), where $\mathbf{X}$ is the training data available, and $\left(\left(X_{t-2},\, X_{t-1}\right),\, X_t \right)$ denotes the sample at time t with input $x_t = \left(X_{t-2},\, X_{t-1}\right)$ and output (=response) $y_t = X_t$. Calibrating the autoregressive model AR(2) of order two, given by

$$y_t = \beta_1 x_{t,\,1} + \beta_2 x_{t,\,2} + \alpha, \tag{1.3}$$

yields $\beta_1 = \beta_2 = \alpha = 0$ for the data of Equation (1.2), which fails at capturing the deterministic XOR function. The XOR function is not linearly separable.

Non-linearly separable patterns can be modeled using a partition $R = \{R_1,\, \ldots,\, R_n\}$ of the input (or feature) space into n regions, and assigning the constant values $\{c_1,\, \ldots,\, c_n\}$ to these regions. The resulting evolution of the time series X_t can then be written as

$$y_t = X_t = \sum_{i=1}^{n} c_i \cdot I\left\{x_t \in R_i\right\} + a_t, \tag{1.4}$$

with i.i.d. innovation a_t. This modeling approach allows for an arbitrary flexibility, as any function can be approximated to any precision with a sufficient number of regions. For example, the XOR data from Equation 1.2 can be modeled exactly by the two regions $R_1 = \{x \in \mathbb{R}^2 | x_1 x_2 \geq 0\}$ with $c_1 = -1$, and $R_2 = \{x \in \mathbb{R}^2 | x_1 x_2 < 0\}$ with $c_2 = 1$. The downside of this modeling approach is that the number of parameters increases arbitrarily with the number of regions, and a procedure to control for overfitting is required.

Nonetheless, such an approach is important to explore because the heuristic biases humans are subject too (Tversky and Kahneman, 1974) likely include linearly inseparable biases like the XOR function. A simplistic interpretation of the XOR function on stock markets goes as follows. When the stock went up two days in a row, traders start to expect a reversal and sell the stock. When the stock went up one out of two days, traders expect the stock to keep rising and buy more. When the stock went down two days in a row, the traders panic an keep selling instead of betting on a reversal. Such a bias between the upside and downside can induce linearly inseparable deterministic patterns in stock returns.

1.2 A critique of statistical learning in finance

Extensive research effort has been dedicated to forecasting stock market returns using conventional regressive models and unconventional models. Atsalakis and Valavanis (2013) and Atsalakis and Valavanis (2009) provide a partial overview of conventional, respectively soft computing methods. A large array of statistical learning methods have found applications in the stock markets forecasting research, the most popular being neural networks, support vector machine, random forest, k-nearest neighbors, logistic regression, decision trees, and random forest. Most of the studies find positive results of predictability in stock returns. However, these studies typically do not reject the EMH for one of the following reasons.

The analysis is ex-post, meaning the input variables of the period to forecast are already known and used to predict the output variable. An example being the study by Chiang et al. (1996) that uses a set of economic variables such as GDP, consumer index, unemployment rates, and inflation rates to predict the same year performance of 101 US funds. The neural network prediction outperforms the linear regression prediction in terms of fund performance residuals as a function of economic indicators. While this shows a relation between economic indicators and fund performance it does not determine if fund performance can be predicted out-of-sample.

The out-of-sample test data is relatively small and statistical significance of the performance metrics are not provided. For example, the studies by Xue-shen et al. (2007) and Huang et al. (2008) perform an out-of-sample directional prediction of a few hundred returns and find a directional accuracy better then random. Unfortunately, in the hypothetical case were the predicted returns had a directional asymmetry during this time period, with for example more up moves, a simple strategy always predicting up (buy & hold) would have above random directional accuracy as well. Without a bootstrap computation of the p-value for the observed directional accuracy the result could origin from a bias in the data.

A plethora of performance metrics are provided but no clear comparison benchmark is available. As in the studies by Tay and Cao (2001) and Chen et al. (2006), the typical performance metrics are the Mean Squared Error (MSE), Normalized MSE, Mean Absolute Error (MAE), directional symmetry (DS), and Weighted Direction Symmetry (WDS). These performance metrics can be useful to compare two methods, but do not give the reader any intuition of the actual trading performance, which could even be negative. As well, in many cases the different metrics exhibit an almost identical performance difference between the compared methods and a single metric would be sufficient.

The predictive performance is significant but does not imply market inefficiency. The study by Villa and Stella (2014) finds impressive directional accuracy on intraday returns on the foreign exchange market of multiple currencies. The prediction is performed out-of-sample on millions of ticks during a one year time period and leaves no doubt to its statistical significance. However, directional accuracy does not necessarily imply profitability as predicting correctly many small returns could easily be offset by transaction costs or predicting incorrectly a few large returns.

The abnormal returns of the strategy, including transaction costs, are compared against the buy & hold strategy but the difference could be insignificant. The statistical learning portfolio from stocks of the S&P 500 created by Creamer and Freund (2010) has significant abnormal returns with respect to the buy & hold strategy or other common portfolio strategies. The abnormal returns even hold after accounting for realistic transaction costs available at a typical broker. However, the test period is short with only 400 trading days and the excess performance is concentrated at the beginning of the period. Without a bootstrap analysis of the returns there is no guarantee that the abnormal returns are significant. A possible bootstrapping approach would be to compute the statistical significance of the observed returns with respect to random portfolios.

This large variety of benchmarks for forecasting, and trading, performance makes it difficult to compare the results of different studies. This line of research needs a standardized framework to assess forecasting and trading performance, and its implication for the EMH. Such a framework should always include a test statistic with respect to a null distribution, and the break-even transaction cost.

1.3 Calibration issues in agent based models

Currently, the use of ABMs remains confined to some academic research fields and has not yet reached the acceptance of the economists working in finance related domains. As discussed in Sornette (2014), the problem in gaining a better acceptance is twofold.

The first problem is that ABMs are highly complex, usually non-linear at the agent level, which makes it difficult to relate the macro behavior to the ingredients at the micro level. To circumvent these issues, the choices made to build a given agent-based model often represent the personal preferences or biases of the modeler, which would not be agreeable to another modeler. An ABM is often constructed with the goal of illustrating a given behavior that is already encoded more or less explicitly in the chosen rules (De Grauwe, 2010, Galla and Farmer, 2013). Therefore, the correctness of the model relies mostly on the relevance of the used rules, and the predictive power is often constrained to a particular domain, so that generalization is not obvious. This makes it difficult to compare the different ABMs found in the literature and gives an impression of lack of robustness in the results that are often sensitive to details of the modeler's choices.

The second problem concerns the calibration of ABMs that due to their inherent non-linear, sometimes chaotic nature, can not be performed by any standardized method. The likelihood function of an ABM generating a given empirical data is very difficult to determine and can often not be sampled entirely due to the large number of parameters in the ABM. This curse of dimensionality forces the modeler to use heuristic optimization methods such as genetic algorithms to fit an ABM to a given dataset. This lack of standardized and analytically well understood calibration method is a major obstacle for economists to trust an ABM in real life applications.

To avoid such biases and minimize the calibration complexity, the class of agent-based models considered in this paper are the ones with a discrete set of possible outcomes, making it possible to list all the available strategies an agent can use. A first such model has been presented in Arthur (1994) to study the "Bar Problem" and subsequently gave rise to an extensive research on the minority game and variants thereof by Challet and Zhang (1997), Challet et al. (1999), and Andersen and Sornette (2002). The study of these models, as performed by Jefferies et al. (2001), showed that they can reproduce a great variety of stylized facts, making them directly relevant to describing real markets. As illustration, for the minority game it has been shown by Challet et al. (2001) that a phase transition exists between an information efficient and absorption phase. Stylized facts such as fat tails and clustered volatility occur at this phase transition, suggesting that within the minority game modeling framework real markets operate in a marginally efficient regime.

Given this knowledge, the next step is to understand how to calibrate these models to a given time series. The logical approach already explored in Johnson et al. (2001), Andersen and Sornette (2005) and Wiesinger et al. (2012) is to turn the strategies of the agents into variables that are then optimized to best reproduce the calibration data. As the space of possible strategies is too large to be sampled entirely, it requires the use of genetic algorithms to reverse engineer the behavior of the involved agents. In contrast to other models, e.g., De Grauwe (2010), Galla and Farmer (2013), the behavior of the individual agents is not biased by the modeler with a specific parametrized functional form, but is determined from the data within all possible behaviors. The calibrated ABM is then used to predict the future return signs, so its predictive power can be tested out-of-sample, which addresses the criticism from Elsenbroich (2011) that “Prediction is a thorn in the side of ABM”. The studies performed by Andersen and Sornette (2005), Wiesinger et al. (2012), and Satinover and Sornette (2012a,b) show that such calibrated ABM’s do possess predictive power.

1.4 Mitigating the replication crisis

In recent years, the discussion of a replication crisis in sciences has intensified (Baker, 2016, Loken and Gelman, 2017). More then half of all scientists agree that there is a significant reproducibility crisis in existing publications, and another third of all scientists agree that there is a slight crisis. While the reproducibility is of particular concern in fields such as psychology and sociology, the phenomenon is prevalent in all disciplines. The strongest driver of the replication crisis is the pressure to publish positive results in top journals, a pressure that can easily lead researcher to perform selective reporting in order to have better results.

I was fortunate enough to write this thesis without pressure to publish and can say in good conscience that the results are presented as truthfully as possible. Besides reporting the results completely and without modifications, I have applied a number of best practices to minimize the risk of errors that could falsify the results.

The most frequent occurrence of mistakes I experienced is during the implementation of a model or method into executable code. The implementation phase bares a major confirmation bias because as long as the results are poor one keeps verifying the code, but when the results are good one tends to consider the implementation as done. Assuming that implementation mistakes are just as likely to improve than to deteriorate the results, the behavioral bias of the implementer asymmetrically boosts the probability that the final implementation contains a mistake improving results. A notorious example is a software package used in neuroscience Eklund et al. (2016) that was found to find significant neural activity in a dead salmon, a result that casts doubt on thousands of previous studies relying on that software.

To avoid implementation mistakes, a best practice is to use well established existing libraries whenever possible. For example, all the used statistical learning methods in this thesis rely on the implementation of the Scikit-learn package from Pedregosa et al. (2011). Scikit-learn is a reliable statistical learning package in Python, which enjoys a good reputation. Mistakes in the statistical learning methods are therefore unlikely, as a large number of people have already approved the good functioning of the Scikit-learn package.

When writing customized code, the best practice is to verify the results for at least two cases where the analytical result can be computed. For example, the random walk model can be used to ensure that the directional accuracy and the average return of a strategy converge to zero as the length N of the time series goes to infinity. Other time series, with controlled predictability, based on repeating patterns, can be used to verify that the expected directional accuracy and average return are obtained in a variety of scenarios.

Typically, when analytical solutions or approximations of special cases match with the simulated results and the overall intuition, the odds are good that the results are correct. However, another class of mistakes can still arise from faulty assumptions. For example, benchmarking directional accuracy with a wrong assumption of the number of up and down moves can lead to positive performance when actually the null hypothesis holds true. It is of great importance to always list the necessary assumptions for a computations and verify them before each application. Some methods, such as bootstrap computations of statistical significance, need much fewer assumptions as they reuse the data to be benchmark and are therefore less error prone.

Finally, when positive results are obtained, the confidence in the results can only be increased by an independent experiment. For example, at CERN, a positive observation must always been seen by the two independently built detectors ATLAS and CMS.

1.5 Chapter contributions

The recurring event of stock market bubbles and crashes in the past decades have cast doubts on the EMH. A stock market crash, after a long phase of irrational exuberance leading to grossly overvalued stocks, does not fit well into a theory of rational and efficient markets. To explain this phenomena, the statistical physics community has proposed models in which bubbles and crashes occur naturally as a consequence of imitation and herding among investors (Sornette, 2003). The herding among traders creates positive feedback loops that first drives the stock market into an overvalued regime, the analogue of a critical point in statistical physics. This critical regime is inherently unstable, like the a ruler held up vertically on your finger, and only requires a small shock to trigger a crash. Due to the massive overvaluation build up during a bull market, and the subsequent propagation of panic through the network of herding traders, the markets often crash dramatically below their fundamental value (Harras and Sornette, 2011).

The herding behavior during the growth phase of a bubble, combined with the need of increasingly large returns to compensate for the risk of a crash, was shown by Sornette et al. (1995), Sornette and Johansen (1997), and Johansen et al. (2000) to produce super-exponential and log periodic behavior in stock market prices. The bubbles following these original publications, namely the dotcom bubbles, financial crisis, and Chinese stock market turbulence all exhibited the log-periodic power law behavior (Johansen and Sornette, 2000, Jiang et al., 2010). In the context of ABM, the reverse engineering experiments of Zhang et al. (2013) and Zhang (2013) provide strong support that the herding among traders is as well detectable in daily return patterns.

This cumulative thesis studies the calibration on nonlinear ensemble models to forecast future oil production and daily stock market returns. A particular focus is put on validation through back-testing, and applying state of the art methodology to determine the statistical significance of results. The forecast of a country's oil production could be improved using an ensemble of nonlinear extrapolations of the individual oil fields. The predictability in daily stock market returns found with previous agent based models is analyzed in detail, and the precise mechanism is presented. Subsequent work connects the fields of agent based modeling and statistical learning, formalizing the agent based models of interest as ensembles of interdependent and nonlinear predictors.

Chapter two presents the generalized methodology developed to forecast the oil production of the UK and Norway from 2014 to 2030, based on the publication Fievet et al. (2015). The methodology extrapolates the production of individual oil fields using a stretched exponential, and estimates the extrapolation error from back-tests. The production resulting from future discoveries of giant and dwarf oil fields is estimated using Monte-Carlo simulations. The validation, using the time period 2008-2014 as out-of-sample, showed that the ensemble model aggregating individual fields forecast's was more robust then the Hubbert model applied to aggregate production. Nonetheless, the estimated error of the forecast remains large despite the detailed modeling. When extrapolating to the complexity of global oil production and demand, the results show that the inherent uncertainties allow rational agents to hold a heterogeneous array of beliefs about future oil prices, without conflicting on fundamental factors.

Chapter three introduces the forecasting and trading performance benchmark subsequently used to assess the evaluated statistical learning and agent based strategies. A robust test for the statistical significance of directional accuracy is presented, as well as an approximation to the relationship between directional accuracy and trading performance. The break-even transaction cost for unit and compounded investment is put into relationship with trading frequency and directional accuracy, and compared to typical transaction costs in stock markets. The computation of multiple testing adjusted p-values of trading strategies is presented in detail. Typical portfolio strategies are presented to be used in the assessment of the ex-ante performance of a universe of available strategies. Last, the four-factor regression model is introduces to regress newly found anomalies on known anomalous factors.

Chapter four presents the particular type of agent based models studied in this thesis. The four different types of games played by the agents, namely the majority game, delayed majority game, minority game, and delayed minority game are analyzed separately in a first step. Each type of game is reverse engineered on major equity indices for a single agent model, to evaluate the predictive performance without multi-agent interactions. In a second step, the reverse-engineering results of the mixed game ABM with many agents are presented. The multi-agent model was found to have no predictive superiority to the single agent model. Therefore, the analysis of the predictive power of single agent models, and the stepwise construction of two agent models, was the natural path for the further chapters in this thesis.

Chapter five formalizes the relationship between the binary strategies of the agents in the agent based model and the decision trees. The abnormal predictive performance of a single decision tree motivated the analysis of further statistical learning methods such as logistic regression, support vector machine, gradient boosting, nearest neighbors, linear discriminant analysis, and quadratic discriminant analysis. Theoretical data generating processes are used to illustrate the limitations of regressive models and advantages of statistical learning models. Some of the statistical learning methods were found to have statistically significant predictive performance on the S&P 500, FTSE and CSI 300. In particular, the predictability was found to be particularly high during the crash phase of the dotcom bubble, the financial crisis, the European debt crisis, and the recent Chinese stock market turbulence. The break-even transaction costs of up to 20 bps per round trip, on a 20 year time period, question the weak efficient market hypothesis. Regressing the anomalous returns for the S&P 500 on the three- and four-factor models shows significant monthly intercept and some market correlation. The observed returns are therefore unexplained by current factors, and constitute a new anomalous 3-day crisis persistency factor (or stylized fact) for the S&P 500. Our time trend analysis uncovered that the most profitable periods happened during the volatile crashes of the dot-com bubble, the financial crisis, and the European debt crisis.

Finally, chapter six introduces an extension of statistical learning towards agent based modeling. In particular, it is shown how multiple instances of an arbitrary predictor (e.g. a decision tree or binary strategy) can be combined to predict a given time series. The experiments are limited to two agent models, which can still meaningfully be calibrated. Models with more agents would require more data, which would be available in the context of intraday trading or order book prediction. However, the later cases are out-of-scope for this thesis.

Chapter seven gathers ideas for future research and concludes the thesis.

Chapter 2

General methodology to forecast oil production

2.1 Introduction

The financialization of commodities has massively increased speculative trading in the oil market. Nowadays, anyone with access to an online broker can speculate on oil prices, without ever physically seeing a barrel of oil. Commodities are attractive for many investors because their pricing should reliably derive from fundamental supply and demand. Consequently, an accurate forecast of future production and demand should provide an equally accurate estimate of future equilibrium expected prices. In this chapter, we quantify the inherent uncertainty in forecasting future oil production for Norway and the UK, two major oil producers with public monthly production data.

The work presented in this chapter is based on Fievet et al. (2015). The publication derives from the master thesis by Del Degan (2012) under the supervision of the authors. The methodology itself derives from the research on pricing social network companies such as Zynga, which has been developed by three of the authors Forró et al. (2012).

Forecasting future oil production has been a topic of active interest since the beginning of the past century because of oil's central role in our economy. Its importance ranges from energy production, manufacturing to the pharmaceutical industry. As petroleum is a non-renewable and finite resource, it is primordial to be able to forecast future oil production. The fear of a global oil peak, beyond which production will inevitably decline, has been growing due to stagnating supplies and high oil prices since the crisis in 2008/2009 (Murray and Hansen, 2013). As any industrialized country, Europe is strongly dependent on oil supply to maintain its economic power. In the nowadays difficult geopolitical environment, it is important to know how much of the oil needed in Europe will come from reliable sources. In the past, a big share has been coming from Norway and the U.K., two of Europe's biggest exporters. However, the U.K. already became a net importer in 2005 and Norway's production has been declining rapidly as well (Höök and Aleklett, 2008). This motivates the novel Monte-Carlo methodology we developed to forecast the crude oil production of Norway and the U.K.

The Monte-Carlo methodology to forecast the crude oil production of Norway and the U.K. is based on a two-step process, (i) the nonlinear extrapolation of the current/past performances of individual oil fields and (ii) a stochastic model of the frequency of future oil field discoveries. Compared with the standard methodology that tends to underestimate remaining oil reserves, our method gives a better description of future oil production, as validated by our back-tests starting in 2008. Specifically, we predict remaining reserves extractable until 2030 to be 5.7 ± 0.3 billion barrels for Norway and 3.0 ± 0.3 billion barrels for the UK, which are respectively 45% and 66% above the predictions using an extrapolation of aggregate production.

2.2 State of the art

2.2.1 Hubbert's Model for peak oil forecasting

The methodology behind forecasting future oil production has not evolved much since M. King Hubbert, who in 1956 famously predicted that the U.S. oil production would peak around 1965-1970 (Hubbert, 1956). That prediction has proven itself to be correct. His main argument was based on a geological estimate of the finite oil reserves and to what amounts as the use of the logistic differential equation for the total quantity $P(t)$ of oil extracted up to time t

$$\frac{dP}{dt} = rP\left(1 - \frac{P}{K}\right) \ . \tag{2.1}$$

The logistic differential equation is characterized by an initial exponential growth, which then decreases to zero as the total oil extracted reaches saturation (no more oil is to be found). If $P(t)$ is the amount of oil extracted up to time t, then $f(t) := \frac{dP}{dt}$ is the oil production rate. The parameter r is commonly referred to as the growth rate, and K as the carrying capacity. The parameter K is the total quantity of oil that can be ultimately extracted, also known as URR (Ultimately Recoverable Resources). When fitting the logistic curve to the data, the carrying capacity K can be given as a constraint from geological information (exogenous URR) or be estimated from the fit (endogenous URR). In M. King Hubbert surprisingly accurate prediction of the production peak, he used an exogenous URR based on his geological knowledge. However, as we are not geologists and given that URR information is sometimes unavailable or differs vastly between different sources, we use endogenously determined URR. Estimating the carrying capacity from the production data only is in fact the main subject of this paper.

From a methodological point of view, the Hubbert model has enjoyed a longstanding popularity in modeling future oil production given its simplicity. Various extensions have been studied by Brandt (Brandt, 2007) to account for multi-cycled or asymmetric production curves.

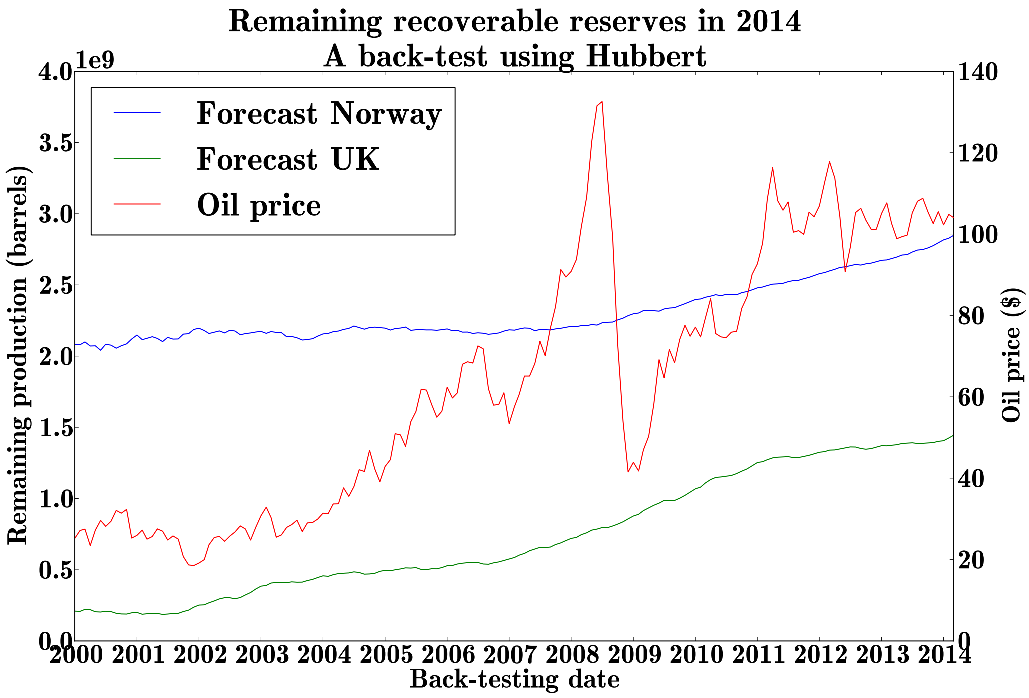


Figure 2.1: A back-test of the forecast stability of the Hubbert model.

2.2.2 Back-testing Hubbert's Model

Despite the popularity of the Hubbert model, the literature says little about its out-of-sample performance, apart from the one case that made the prediction famous, namely the prediction by Hubbert of the conventional oil peak that occurred in the US in 1970. This prediction relied on geological estimates of the URR, and is therefore not only a success of the model but also of the geological measurements.

There are many ways to address this question and we choose to study the remaining recoverable reserves predicted by the Hubbert Model when going back in time. This approach differs notably from Hubbert's approach because the URR is now computed from the model and no longer an exogenous constraint. As the problem of fitting the logistic curve (equation 2.1) is under-constrained before the peak of $f(t)$, the back-test had to be started when the peak is already well visible in the data. For each time T, a month between the start of the back-test and the date of the latest available date, we fitted the Hubbert model to the data truncated at time T. Based on each fit, we computed the predicted remaining recoverable reserves in 2014.

This back-test allowed us to study how the forecast of Hubbert's model evolves over time. A good model, which captures well the underlying dynamics of the system, should yield a stable forecast over time. An unstable forecast, where the model constantly has too adapt to new data, can be considered as an indicator that the model is missing some aspects (technically, it is called "miss-specified"). This is of course assuming that the dynamics of the system is deterministic to a higher degree than the error the model is making in its forecast.

From a mathematical perspective, the Hubbert model is symmetric with respect to its peak and will necessarily make inaccurate predictions for asymmetric production data. It will overestimate the decline of historical production data with a negative skewness and underestimate data with a positive skewness. An extended Hubbert model with an asymmetry parameter may yield an improvement, but must not necessarily capture the dynamics of the asymmetry.

For the studied cases, namely Norway and the UK, a good starting date was chosen to be 01/2000. Fitting Norway's production could be made using the simple logistic equation 2.1. In contrast, for the UK, the production exhibits two clear peaks, which required the use of a double-logistic function that is the sum of two independent logistic functions with time shifted peaks. The peak of each of the two logistic functions is matched up with one of the two peaks in the UK production data.

The results of this back-test can be seen in figure 2.1 and clearly show that the forecast is not stable over time. In 2000, for Norway, the Hubbert model would have predicted 2.08Gb of remaining recoverable reserves by 2014 while, up to the time of writing (July 2014), the predicted amount grew to 2.83Gb (a 36% error) and the growth is not showing sign of slowing down, yet. For the UK, the prediction would have been 0.21Gb but grew all the way to 1.43Gb (an error of 581%), with a recent slow down in growth. These results clearly show that the Hubbert model can be highly unstable and unreliable in its forecast of recoverable reserves, which is not surprising as the historical data is asymmetric with a positive skewness. For completeness, let us stress that the relative error between the year 2000 and July 2014 in total production since the beginning is only 17% for Norway and 6.5% for the UK. Nonetheless, the Hubbert model greatly underestimates the fat tail of the production curve.

Our aim is first to understand why the Hubbert model underestimates the tail of the production curve and then to improve upon it. An argument could be the strong increase in the oil price, which widened the amount of economically recoverable reserves. To discuss this hypothesis, figure 2.1 shows the oil price as well. The forecast for Norway remains fairly stable from 2000 to 2007 during a period when the oil price went from slightly above \$20 per barrel to more than \$80 per barrel. The shift to a larger predicted remaining production does not correlate with the peak, bottom and then rebound of the oil prices since 2007. For the UK, the behavior is different, with increases in the forecast showing very little delay with respect to the increases in oil prices. One could speculate that Norway had little spare capacities and therefore the increase in remaining reserves was always delayed with respect to oil prices as new capacities had first to be build. On the other hand, UK could have had available or upcoming spare capacities that allowed to respond rapidly to the increase in the oil price. While it is certain that increases in oil prices will augment the recoverable reserves, it seems difficult to establish a well-defined mathematical link between oil price and the increase in forecast reserves. Consequently, an extension of the Hubbert model integrating the oil price has not been further explored.

2.2.3 Beyond Hubbert's Model

The existing forecasts of future oil production all use some version of the Hubbert model (Brecha, 2012, Laherrère, 2002, Lynch, 2002) or some economical model applied to aggregate production (Greiner et al., 2011). However, as just shown in the back-test, the Hubbert model does not provide a forecast that is stable in time, underestimating the significant tail in the production curve of the UK and Norway. Considering that none of these Hubbert-based methodologies go into the details of studying the underlying dynamics, it is not surprising that important features of the dynamics bearing on the total recoverable reserves have been missed. In defense of the existing studies, the main reason for the lack of details is certainly the lack of available data. For this reason, the Hubbert model will likely remain a tool of choice to estimate the total reserves when only the aggregated production of a region (typically a country) is available. Nonetheless, in this article, a new methodology is introduced to forecast future oil production. Instead of taking the aggregate oil production profile and fitting it with the Hubbert curve or its variants (such as the multi-cyclic Hubbert curve), the production profile of each individual oil field is used. By extending their production into the future and extrapolating the future rate of discovery of new fields, the future oil production is forecast by means of a Monte Carlo simulation. To demonstrate the generality of the methodology presented here, it is applied to two major oil producing countries with publicly available data at the field level: Norway Norwegian Petroleum Directorate (2014) and the U.K. (GOV.UK, 2014).

2.3 Methodology

The idea behind the methodology presented here is to model the future aggregate oil production of a country by studying the production dynamics of its individual constituents, the oil fields. In order to implement our approach, one must be able 1) to **extend the oil production of each individual field** into the future and 2) to **extrapolate the rate of discoveries** of new oil fields.

2.3.1 Extending the oil production of individual fields

The first step to predicting the future oil production of a country is to extrapolate the future production of existing fields and to estimate the error on this extrapolation. The data of the fields developed in the past shows a repeating asymmetric pattern. A good example is the Oseberg field shown in figure 2.2, with a quick ramp up once the field is being developed, and then a peak or plateau before the oil field production starts decaying. The decay can take many different shapes and is governed by a variety of geological and economical factors. The goal of the fitting procedure is to capture as much as possible the impact of these different factors.

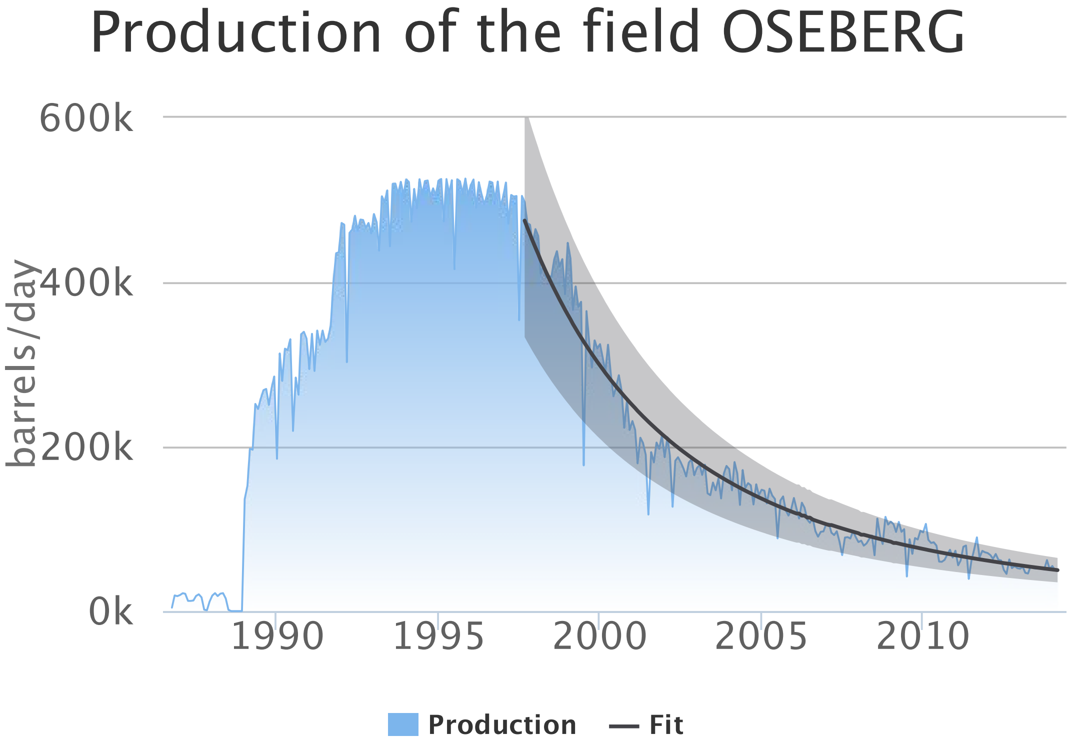


Figure 2.2: Example of the time evolution of the oil production per day of a regular field, parametrized beyond the peak in the decay regime by the stretched exponential function (2.2) with $\beta = 0.66 \pm 0.01$ and $\tau = -55 \pm 1$ months, shown with the black line. The one standard deviation given by expression (2.7) is represented by the gray band.

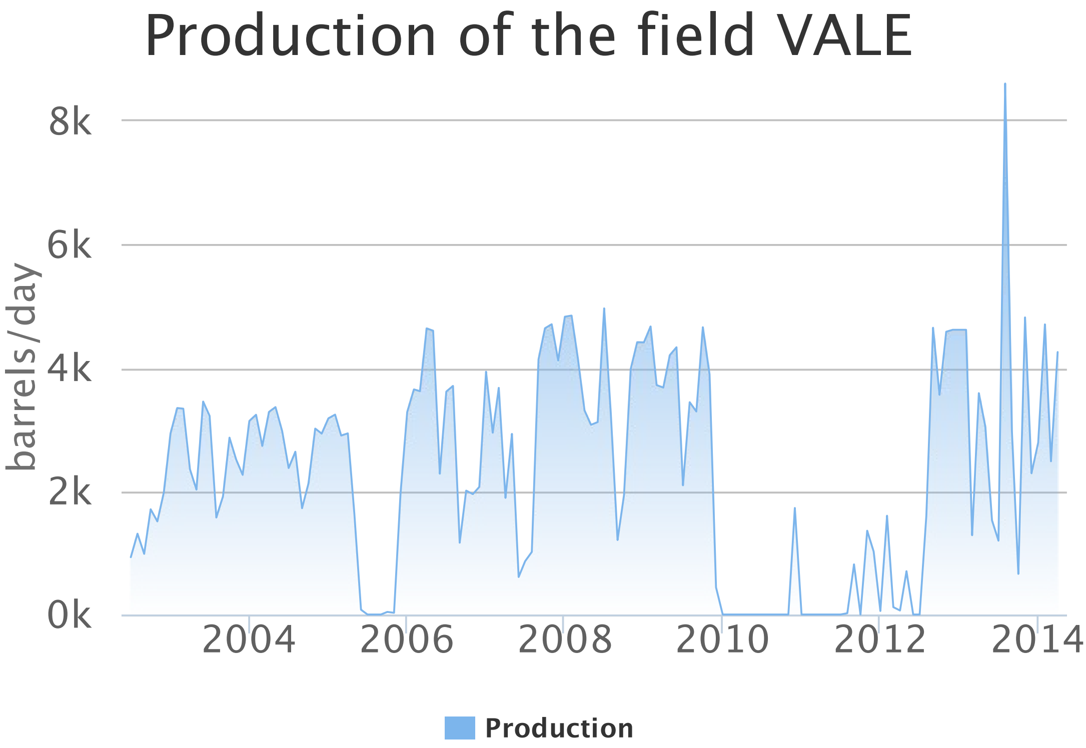


Figure 2.3: Example of an irregular field.

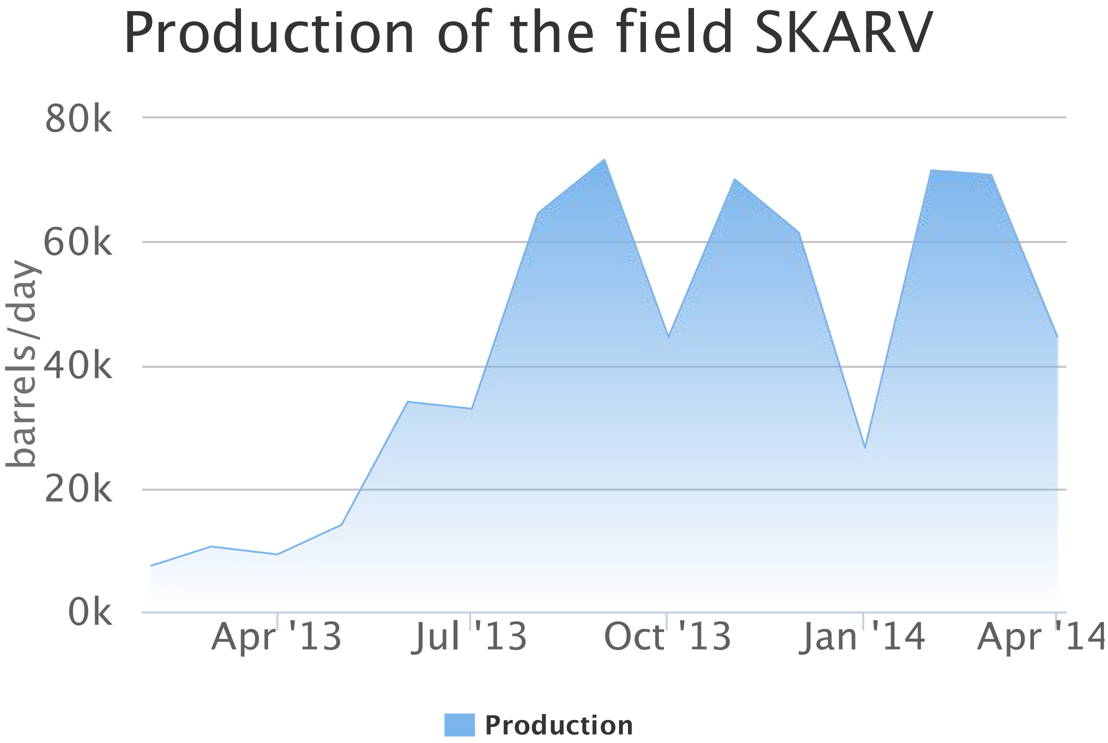


Figure 2.4: Example of a new field.

2.3.1.1 Regular, irregular and new fields

To forecast the oil production of each individual field, regularity had to be found in the production's dynamics. Unfortunately, the whole production profile can take a large variety of shapes with often very asymmetric shapes. Therefore, modeling the whole production profile from the beginning of extraction with the Hubbert model seems elusive. It would require the use of various extensions of the Hubbert model and for each field the most appropriate extension would need to be determined. This is impractical when dealing with hundreds of oil field.

Fortunately, modeling the decay process is sufficient in order to extrapolate future oil production. A preliminary classification is necessary to achieve that goal. Figures 2.2, 2.3 and 2.4 show that, independent of the country, oil fields can be classified into three main categories:

- **Regular fields** - Their decays show some regularity (see figure 2.2);
- **Irregular fields** - The ones that do not decay in a regular fashion (see figure 2.3);
- **New fields** - The ones that do not decay yet. As such, there is no easy way to forecast their future oil production based on past data (see figure 2.4).

All the fields have been fitted using an automated algorithm, but the results have been subsequently checked visually to sort out the irregular fields which could not be fitted. As of January 2014, regular fields make up 85% and 87% of the number of fields and 94% and 71% of the total produced oil volume in Norway and the U.K. respectively. As such, being able to model them is crucial. To capture as many different decay dynamics as possible, the decay part of the oil production rate $f(t) := \frac{dP}{dt}$ has been fitted by the stretched exponential

$$f(t) = f_0 \; e^{-\left(\frac{t}{\tau}\right)^{\beta}} \; . \tag{2.2}$$

The stretched exponential function has many advantages as it generalizes the power law and can therefore capture a broad variety of distributions as shown by Laherrère and Sornette (1998). Moreover, (Malevergne et al., 2005) showed that the power law function can be obtained as an asymptotic case of the stretched exponential family, allowing for asymptotically nested statistical tests. As can be seen in figure 2.2, the stretched exponential (equation 2.2) is a good functional form to fit the decay process of regular fields.

For the minority of irregular fields, we assume no difference in the decay of production between giants and dwarfs, which has been modeled as follows: the decay time scale τ has been picked to be the average τ over the regular fields. Then β has been fixed so that the sum of the field production over its lifetime be equal to the official ultimate recovery estimates, when such an estimate is available.

The minority of new fields, which did not yet enter their decay phase, cannot be extrapolated and will therefore be treated as new discoveries. The technical details of how to treat them as new discoveries are discussed in section 2.3.2.3.

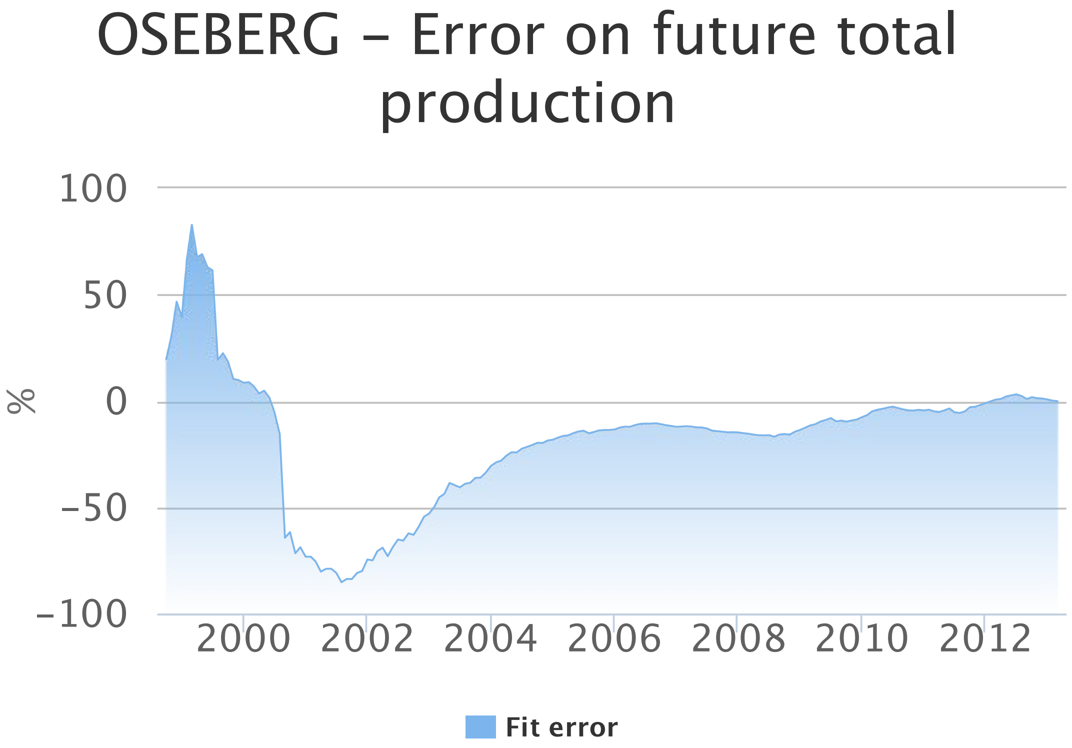


Figure 2.5: Oesberg field - Relative error defined by expression (2.4) of the predicted total production from time t indicated in the abscissa until 2014. One can observe that the predicted future total production is over-estimated by as much as 70% in 1999, then under-estimated by the same amount in 2002, while the forecast errors remain smaller than 20% since 2004.

2.3.1.2 Back-testing & Error

For every extrapolation, be it the Hubbert model applied to aggregate production at the country level or the stretched exponential applied to a field's decay, we want to determine how good and stable the predictive power is. To do so, we perform a complete back-testing in every case. This includes a monthly back-test for the extrapolation, using a stretched-exponential, of the decay process of every oil field.

A single back-test is made as follows:

- The production data $\{p_0, \ldots, p_N\}$ of the field or country is truncated at a certain date in the past $T \in \{0, \ldots, N\}$, where T is the time counted in months since the production start of the field.
- The extrapolation of the oil production rate $f(t) := \frac{dP}{dt}$ is made based on the truncated data $\{p_0, \ldots, p_T\}$.
- The future production predicted by the extrapolation function $f(t)$ can be compared to the actual production from the date T in the past up to the present $T_f = N$. The

extrapolated total production can be computed as

$$P_e(T) = \int_T^{T_f} f(t)dt \tag{2.3}$$

and the relative error is given by

$$e(T) = \frac{P_e(T) - \sum_{i=T}^{T_f} p_i}{P_e(T)}, \tag{2.4}$$

where both $P_e(T)$ and $e(T)$ are functions of the truncation time T.

Computing this back-test, for every month in the past since the production started decaying, yields a plot showing the evolution of the relative error over time defined by expression (2.5). By construction, the relative error will tend to zero as the truncation time T approaches the present. Nonetheless, it is a useful indicator for the stability of the extrapolation. As can been seen for the Oseberg field in figure 2.5, the relative error on future production remained fairly stable during the past decade.

From the complete back-test, we compute the average relative error

$$\bar{e} = \frac{1}{N}\sum_{i=0}^{N} e(i) \tag{2.5}$$

of the extrapolation made on the future production. Assuming that the relative errors are normally distributed around the average relative error, the standard deviation on the average relative error is given by

$$\sigma_e = \sqrt{\frac{1}{N}\sum_{i=0}^{N} (e(i) - \bar{e})^2}. \tag{2.6}$$

As the average relative error is often fairly constant, the extrapolation was corrected by the average relative error, that is, if the extrapolation consistently over-estimated the production by 10% during the back-test, the extrapolation was reduced by 10%. This results in an extrapolated production $p(t)$, including a 1σ confidence interval, given by

$$p(t) = (1 - \bar{e})\, f(t) \pm \sigma_e. \tag{2.7}$$

An example of such an extrapolation including a one standard deviation range is shown in figure 2.2 for the Oseberg field.

2.3.1.3 Aggregate error

Once the individual fields have been extrapolated using formula 2.7, we compute the extrapolation of the oil production for the whole country. While it is straightforward to sum the extrapolations of the individual fields to obtain the expected production, some care has to be taken with respect to the confidence interval of the production at the country level.

As shown later in section 2.4, the same extrapolation including a complete monthly back-test of total future production has been performed at the country level and the resulting relative error is much smaller than the average error observed on the individual fields. To account for this observation, the assumption made is that the relative error between individual fields is uncorrelated. While one could imagine that some inter-dependence could result from a coordinated response of supply to a sharp increase/decrease of demand, we have not observed this to be the case at a significant level. Therefore, the fields can be considered as a portfolio of assets with a return given by their extrapolation $p(t)$ (eq. 2.7) and a risk given by σ_{field} (eq. 2.6). This means that the average standard deviation per field at the country level from the extrapolated production can be computed as

$$\sigma^2_{country} = \frac{1}{\#fields} \sum_{field \in fields} \sigma^2_{field}. \tag{2.8}$$

Intuitively, this models well the fact that the uncorrelated errors among fields will mostly cancel out.

2.3.2 Discovery rate of new fields

Knowing the future production rate of existing fields is not enough as new fields will be discovered in the future. The model describing the discovery rate of new fields should satisfy two fundamental observations.

1. The rate of new discoveries should tend to zero as time goes to infinity. This is a consequence of the finite number of oil fields.

2. The rate of new discoveries should depend on the size of the oil fields. As of today, giant oil fields are discovered much less frequently than dwarf oil fields.

2.3.2.1 Discoveries modeled as logistic growth

A natural choice for such a model is a non-homogeneous Poisson process. The Poisson process is a process that generates independent events at a rate λ. It is non-homogeneous if the rate is time-dependent, $\lambda \to \lambda(t)$. The standard way to measure $\lambda(t)$ is to find a functional form for $N(t)$, the statistical average of the cumulative number of events (discoveries) up to time t. Then, $\lambda(t)$ is simply a smoothed estimation of the observed rate $\frac{dN(t)}{dt}$. Figure 2.6 shows $N(t)$ for Norwegian fields classified according to their size in two classes, dwarfs and giant fields. The logistic curve is a good fit to the data (integral form of equation 2.1). This implies that after an initial increase, the rate of new discoveries reaches a peak followed by a decrease until no more oil fields are to be found, consistent with our fundamental observations. This same approach has already been successfully applied by Forró et al. (2012) to estimate the number of daily active users on Zynga.

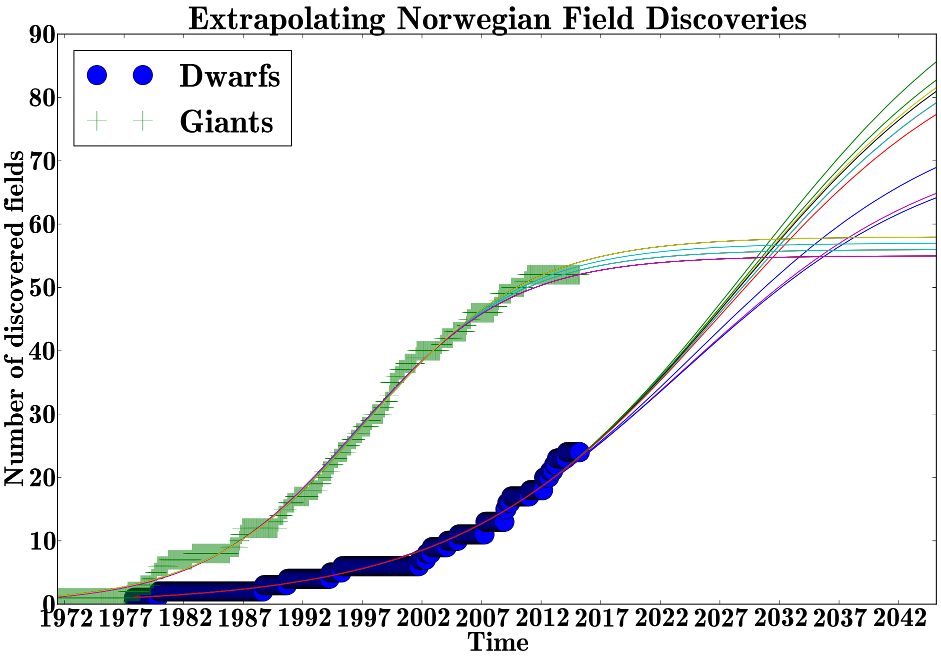


Figure 2.6: Logistic fit of the function solution of expression (2.1) to the number of discoveries for Norway. The discovery rate of new oil fields is dependent on their size as explained in the main text.

As the discovery and production dynamics are not independent of the field size, the fields have been split into two groups: dwarfs and giants. Unfortunately, the two logistic curves thus obtained are highly sensitive to the splitting size. This results from the major issue, when fitting a logistic curve to data, that the carrying capacity can not be determined if the data does not already exhibit the slowdown in growth towards the carrying capacity. However, it is mentioned in the literature that often dwarf fields have already been discovered a long time ago, but their production has been postponed for economical reasons (Lynch, 2002, p. 378). Therefore, it is expected that the large oil fields have mostly been found and produced, and that future discoveries will mostly be made up of dwarf fields. Consequently, the splitting size has been picked as small as possible in order to maximize the number of giant fields but large enough to avoid recent discoveries. Our definition is thus:

- **Dwarfs**: Fields which produced less than $50 \cdot 10^6$ barrels.
- **Giants**: Fields which produced more than $50 \cdot 10^6$ barrels.

We note that this definition differs by a factor 10 from the more standard one, for which oil fields with an ultimate recoverable resource of 0.5 billion barrels (Gb) or higher are classified as giants, while oil fields with smaller URR are considered to be dwarfs (Höök and Aleklett, 2008).

The resulting plot shown in figure 2.6 pictures the dynamics: giant oil fields have

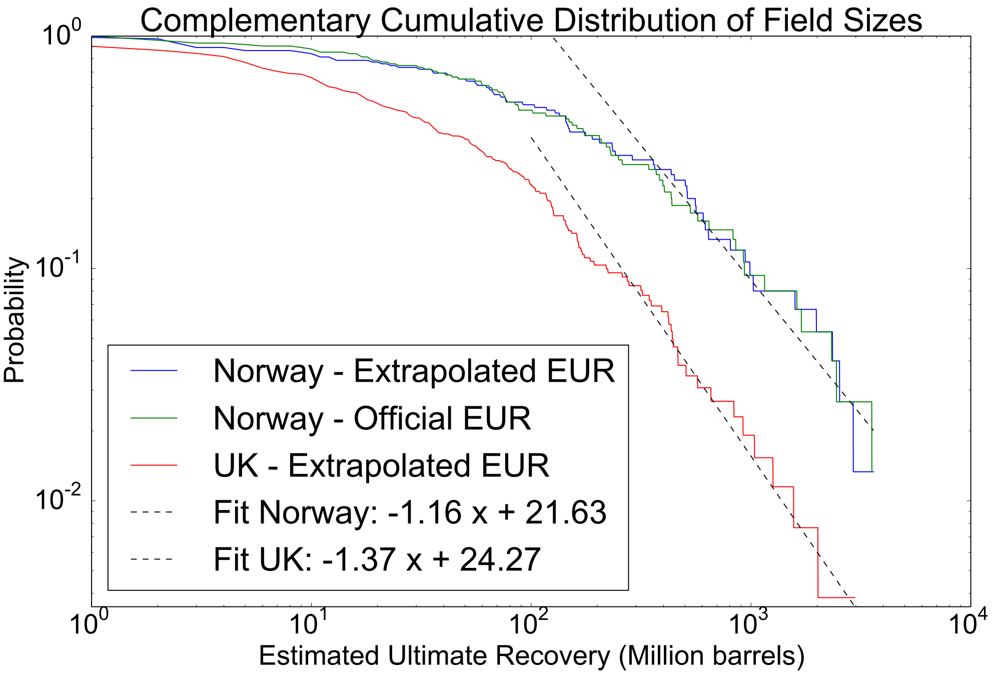


Figure 2.7: Complementary cumulative distribution function (CCDF) of known oil field sizes S from Norway and the UK. The two dashes lines visualize the power law behaviour (2.9) of the tail of the distributions.

mostly been found while the discovery process for dwarf fields is still ongoing. The logistic growth curve fit to the giant discoveries is well constrained, however the fit for the dwarfs is poorly constrained. As can be seen in figure 2.6, the carrying capacity K of the logistic growth model is not well constrained by the available data. A large spectrum of values for K can lead to an equally good fit of the data. We have not taken into account the possible effect that newly discovered dwarf fields could become smaller and smaller until the new fields have a size that is too small to be economical to drill.

There are in fact two competing effects that are likely to compensate each other. On the one hand, figure 2.7 shows the complementary cumulative distribution function (CCDF) of known oil field sizes S from Norway and the UK. Two salient properties can be observed. First, the tails of the distributions are well described by power laws

$$\mathrm{CCDF}(S) \sim \frac{1}{S^{1+\alpha}} , \tag{2.9}$$

with exponent $\alpha = 1.2 \pm 0.1$ for Norway and $\alpha = 1.4 \pm 0.1$ for the UK. The fact that the estimations of the exponents α are larger than 1 implies that the cumulative oil reserves are asymptotically controlled by the largest fields, and not the small ones (Sornette, 2004). However, the fact that the exponents α are rather close to 1 (which is called “Zipf law”) would make the many small oil field contributing significantly in total. This brings us to

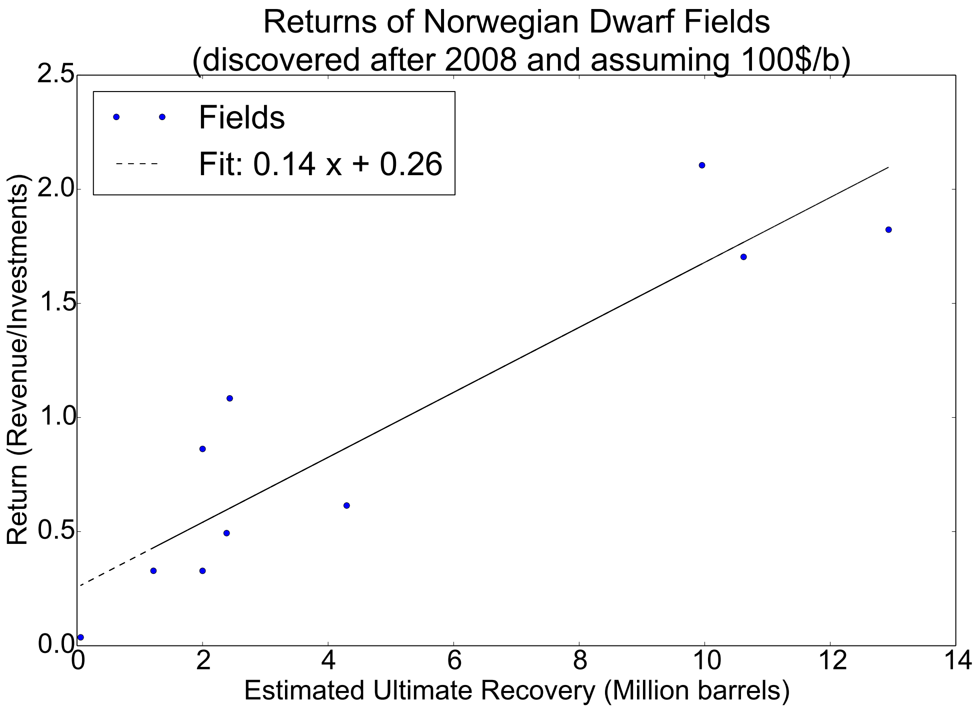


Figure 2.8: Total estimated revenue divided by investment as a function of the estimated ultimate recovery for a number of small oil fields.

the second important feature exhibited by figure 2.7, namely the roll-overs of the CCDFs for small oil fields, likely due to an under-sampling of the data. Indeed, as for most data sets involving broad distributions of sizes such as oil fields, the distributions are in general incomplete for the small events due to the non exhaustive sampling. This incompleteness raises the specter that our extrapolations might be grossly underestimating the large potential contributions to the total reserves of the many small yet undiscovered small oil fields. Assuming that the power law (2.9) would hold for smaller fields down to size of 1 million barrels leads to a number of such fields 10 to 100 times larger than presently known.

But there is another key factor that needs to be considered, namely the fact that small oil fields are not economically viable for exploitation, which leads to an effective truncation in the distribution (2.9) relevant for the estimation of recoverable oil. In 1980, the threshold for economic viability was as high as 70 million barrels (Smith, 1980, p. 591), but with today's oil price we expect a much lower threshold. Taking the data from the Norwegian Petroleum Directorate (2014) providing the yearly investments broken by fields in Norway, let us consider the illustrative case of the field GAUPE. The investment spent to develop its exploitation was \$380M, while its estimated size is $\simeq 1.2Mb$, which corresponds approximately to a total market value of $\$120M$ at \$100 per barrel. It is thus not a surprise that investment to exploit this field was interrupted in 2013. Figure 2.8 shows the total estimated revenue divided by investment as a function of the estimated ultimate recovery for a number of small oil fields. It can be inferred that oil fields of sizes smaller than about 10 Mb are not economically viable as long as the market oil price does not grow much higher than \$100 per barrel. This implies that, notwithstanding the large number of unknown small oil fields, economic considerations oblige us to neglect the small oil fields, therefore providing a justification of our procedure. In fact, economic constraints may lead to cap the carrying capacity at a value smaller than the one shown in figure 2.6.

To address these issues from a more solid angle, the method described in the next section 2.3.2.2 has been used to compute the probability of different carrying capacities.

2.3.2.2 Likelihood function for the number of discoveries

To overcome the poor constraint on the carrying capacity K obtained from the fitting procedure for dwarf fields, a method already used by Smith (1980) has been implemented. This method makes the following two postulates:

1. "The discovery of reservoirs in a petroleum play can be modeled statistically as sampling without replacement from the underlying population of reservoirs."
2. "The discovery of a particular reservoir from among the existing population is random, with a probability of discovery being dependent on (proportional to) reservoir size."

The fields are split into J size bins denoted $S_1, \ldots, S_J$ occurring with frequency $n_1, \ldots, n_J$. Each discovery is considered as a step i at which a field of size $I(i) \in \{S_1, \ldots, S_J\}$ is found and m_{ij} denotes the number of fields of size $j \in \{1, \ldots, J\}$ discovered before the i^{th} step. Then, the probability that the discovery at step i is of size j can be expressed as

$$P\left(I(i) = S_j\right) = \frac{\left(n_j - m_{ij}\right) \cdot S_j}{\sum_{k=1}^{J} \left(n_k - m_{ik}\right) \cdot S_k}. \tag{2.10}$$

The likelihood L for a complete sequence of N discoveries $\{I(1), \ldots, I(N)\}$ can then be expressed as

$$L = \prod_{i=1}^{N} \frac{\left(n_{I(i)} - m_{iI(i)}\right) \cdot S_{I(i)}}{\sum_{j=1}^{J} \left(n_j - m_{ij}\right) \cdot S_j}. \tag{2.11}$$

The unknown parameters are the number of fields $n_1, \ldots, n_J$, whose likelihood can now be estimated based on the existing discoveries. Using a brute force approach, the entire space of plausible values for the variables $n_1, \ldots, n_J$ has been sampled. The values n_j have been sampled between the number of existing fields m_{Nj} in the bin j and up to a value n_j^{upper}, such that the scenario with the largest likelihood according to equation (2.11) has $n_j^{max} = n_j^{upper}/2$ fields in the bin j. Subsequently, the likelihood of each scenario (value of the tuple $n_1, \ldots, n_J$) has been normalized such that the total likelihood of all generated scenarios equals one.

For the analysis of the discoveries of the North Sea oil fields, the number of size bins has been fixed to $J = 2$ splitting between dwarfs (1) and giants (2) as described in section 2.3.2.1.

Table 2.1: Likelihood estimation for the number of dwarf (1) and giant (2) fields. For Norway, our logistic fit suggests that up to two new giant fields could be ultimately discovered. For the U.K., the prediction is more bleak, suggesting that the most probable scenario is that one giant field will be found (which is most likely already discovered but classified as a new field).

	m_{N1}	$n_1 \pm \sigma_1$	m_{N2}	$n_2 \pm \sigma_2$
Norway	24	88.4 ± 10.0	52	56.4 ± 1.6
U.K.	162	208 ± 11	99	100 ± 0.4

The results shown in table 2.1 are coherent with the intuitive expectation that discovering a new giant field is unlikely and that future discoveries will mostly be made up of dwarf fields. The likelihoods obtained for the carrying capacities of dwarfs and giants have been used to constrain the logistic curve fitted to the discoveries. Figure 2.6 shows a sample of fitted logistic curves, each curve being weighted by the likelihood of its carrying capacity given by equation (2.11).

A comparison with the results in table 1 of Smith (1980) show that our results are coherent. Back in 1980, Smith (1980) used $J = 7$, which fortunately maps easily to our splitting using $J = 2$ as his first reservoir size is 0 to 50 Mb as ours, and the sum of the 6 others yields our second reservoir size 50 Mb and above. His estimated total reserves were 43.2 billion barrels with 90% confidence bounds at 38.4 and 65.0, across 203 dwarfs and 117 giants in our definition. Currently, we know that total estimated reserves are reaching 59 Bb, which is still within the 90% confidence interval of Smith. This strongly speaks in favor of his likelihood technique. Given that, in the past 34 years, there have been major technological improvements and a strong increase in the oil price, it is not surprising that the actual reserves are tending towards his upper bound.

A linear scaling by 59/43.2 (actual reserves over Smith expected reserves) of the number of fields Smith predicted, yields 277 dwarfs and 160 giants. This is close to the 296 ± 21 dwarfs and 156 ± 2 giants of our 60% likelihood estimation found for Norway and the UK combined. Such a strong coherence between the two fully independent implementations using different parameters (e.g. $J = 7$ versus $J = 2$) and different data sets (1980 versus 2014) provides a credible validation of the approach.

2.3.2.3 Future production from discoveries

We now compute an expected oil production coming from future discoveries, which requires to combine the steps described in sections 2.3.2.1 and 2.3.2.2.

The method described in section 2.3.2.2 yields probabilities for the total number of fields (including the not yet discovered fields) in each size bins (called a scenario). However, this likelihood method does not give the time distribution of future discoveries. We propose to use the likelihood function to generate scenarios with their respective occurrence probability.

For a given scenario, the carrying capacity K (= total number of fields) is given for each size class. This is useful to resolve the instability in fitting the logistic curve to the number of discovered fields. The time distribution of the discoveries is then given for each size class by the fitted logistic curve.

The actual size of a newly discovered field is generated according to the size distribution of the existing fields in its size class. The probability distribution function of field sizes in a given size bin has been fitted by a stretched exponential function.

The production curve is computed based on the average production curve of all existing fields in the same size category. The production curves of the existing fields have all been normalized to a total production of one and then have been averaged. This yields the typical production profile including a one sigma confidence interval. For a new field, this typical production curve is than multiplied by the size of the field.

Superposing the production curves results in the expected production curve from future oil fields for a given scenario.

As the total parameter space is too large to be sampled entirely, a Monte Carlo technique is applied to compute the expected production with confidence interval from future discoveries. In a nutshell, the algorithm works as follows:

1. Draw a scenario (total number of fields in each size bin) based on its probability according to the likelihood function (2.11). This is done by generating a random number r between 0 and 1, and computing the scenario that is mapped to r by the cumulative distribution function of all scenarios.

2. Compute the time distribution of new discoveries by fitting a logistic curve for each size class.

3. For each discovery, generate a size and the resulting production curve based on the size distribution and production curves of existing fields.

4. Superpose all the production curves.

5. Repeat and average over all drawn scenarios.

The result is the expected production curve of future oil field discoveries. The distribution of generated scenarios yields the confidence interval.

Last but not least, it has to be defined how the expected production from these future oil field discoveries is added to the extrapolated production from existing fields. The start of the simulation of new discoveries does not match up with the date of the latest production data, the reason being that the new fields (defined in section 2.3.1.1), which are already discovered but did not yet enter the decay phase, are not taken into account in the simulation as their final size is not known. The only meaningful way of treating the new fields is to consider them as a discovery. Therefore, the starting point in time of the simulated production resulting from new discoveries has been chosen as the date in the past where it matches the current production from new fields. In order words, the extrapolated production from regular and irregular fields added to the production from simulated future discoveries (which is shifted into the past as to match the production from new fields) must be equal to the current (latest available data) total production from all fields.

2.3.3 Contributions and applicability of the methodology

The methodology described in this section combines tools from different areas to analyses individual oil fields and forecast their combined oil production. In short, the main benefit of this approach, compared to working directly with aggregate production data, is the possibility to forecast non-trivial oil production profiles arising from the combination of all the individual field dynamics. This ability reflects directly the fact that the total production is the sum of the contributions of each individual oil field in production. Thus modeling at the level of each field reflects more closely the reality and is likely to be over-performing and more reliable, as we show below.

2.3.3.1 Contributions

To describe the decay dynamics of individual fields we made use of the stretched exponential, which is extensively used in physics to model a variety of distributions Laherrère and Sornette (1998). This is an improvement which has never been done in the existing literature on oil production and allows us to capture the large differences in the decay dynamics of different fields.

When fitting the decay process of an individual field, we did not only fit the complete data, but performed as well a complete back-test. This is a standard procedure in the analysis of financial times series for instance, but is not or rarely done in the literature on forecasting oil production. This allowed us to compare the stability over time of the different models.

Modeling the discovery dynamics of oil fields is almost untouched in the literature and the algorithm presented in this paper constitutes a novel development. The logistic curve is typically applied to modeling population growth and it is therefore logical to use it to model the growth of discovered oil fields. The Hubbert model, describing total aggregate production, is a logistic curve too and it is consequently logical to assume that the underlying oil field population follows a logistic growth curve over time.

The first challenge we had to tackle concerns the difference between different oil field sizes. As explained in section 2.3.2.1, we used an economical argument to choose the optimal splitting size between dwarfs and giant fields, which indeed yielded stable results.

A second challenge arose from the sample of discovered dwarf fields, which is too small to fit the logistic curve because the carrying capacity is not sufficiently constrained. To overcome this issue, the method uses the likelihood estimation developed by Smith (1980) to estimate the carrying capacity. Using the carrying capacity estimates from

Table 2.2: Remaining oil reserves in billion barrels (Gb) until 2030 predicted by the extrapolation of the **Fit** of the past country production, and predicted by the Monte-Carlo Model. The relative difference between these two predictions is defined by $\Delta = \frac{\text{Model}-\text{Fit}}{\text{Fit}}$.

	Hubbert (Gb)	Fit (Gb)	Model (Gb)	Δ
Norway	2.83	3.95	5.72	45%
U.K.	1.43	1.79	2.98	66%

this likelihood method allowed us to constraint the logistic curve and obtain a stable fit resulting in a prediction of the discovery dynamics over time.

2.3.3.2 Applicability

The Hubbert model is well established to obtain the big picture of aggregate production. However, as discussed, it will fail to capture any asymmetry before and after the production peak and does not allow for non trivial decay dynamics arising from the combination of the heterogeneous dynamics of the individual oil fields. This results in a failure to capture the fat tail in the production curve we assume arises from recently high oil prices.

The method developed in this paper is specifically aimed at modeling the decay dynamics and describing the fat tail, which needs two major requirements:

1. A large fraction (preferably above 80%) of the existing fields need to have entered their decay phase for at least a few years, because the fitting procedure for existing fields is only applied to the decay phase after the peak.
2. A sufficiently large fraction of all fields is already discovered, otherwise the range of likely scenarios generated by the likelihood method is so large as to render the prediction useless.

As mentioned, these two requirements are well met for the UK and Norway. Additionally, our probabilistic method does not require any other input than historical production data. Consequently, the predicted ultimately recoverable reserves can be used as an unbiased cross check for geological URR estimates.

2.4 Results

Based on the methodology described in section 2.3.2, simulating future oil production was straightforward. For each country, the existing oil field productions were extrapolated and the future discoveries were simulated. Figure 2.9 shows the predicted production resulting from the extrapolation of existing fields and the average of 1000 simulations of future discoveries. For each country, the non-symmetric shape of the production dynamics,

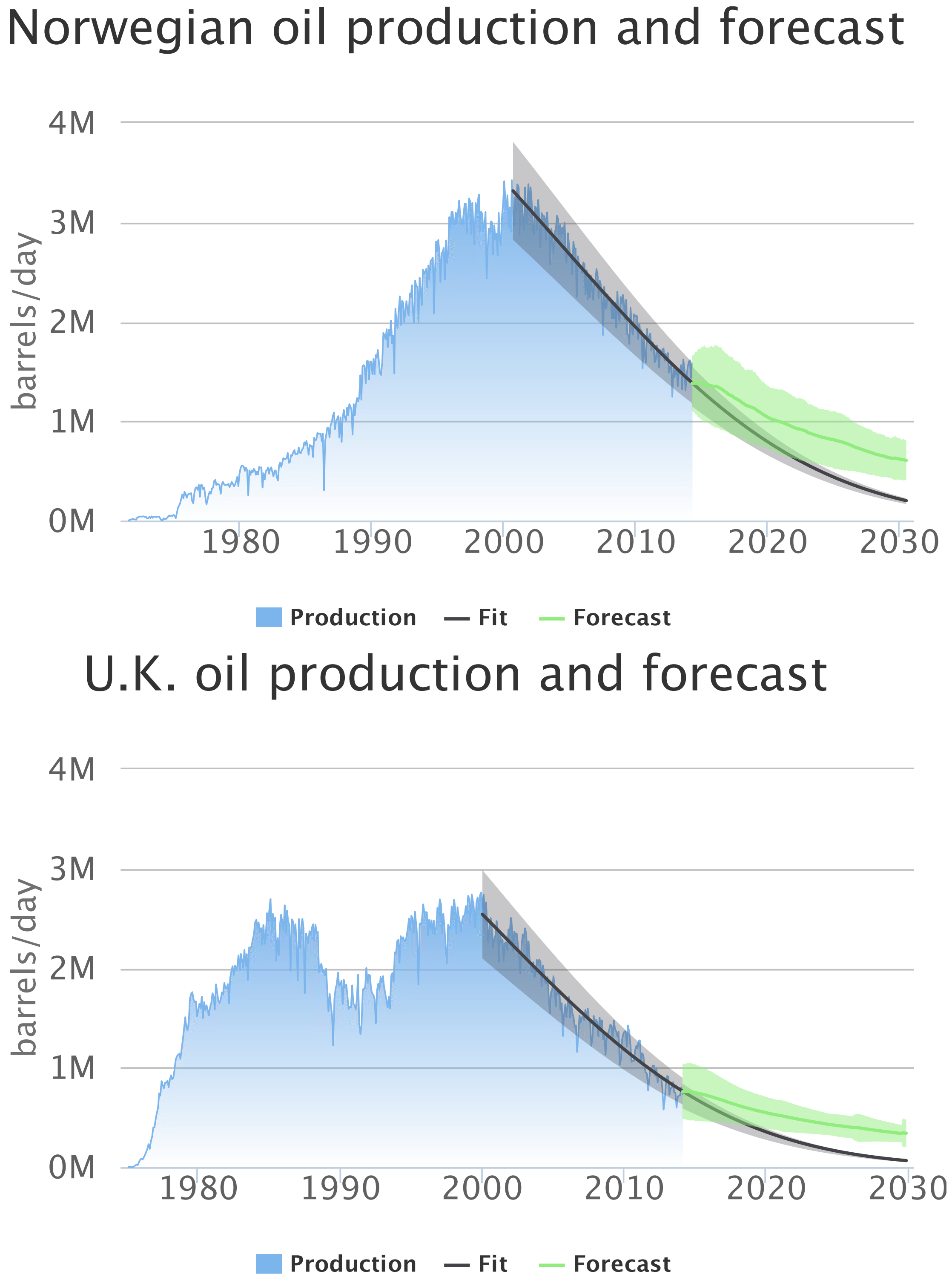


Figure 2.9: Monte-Carlo (green upper continuous line with standard deviation band starting in 2014 onward) and fit forecast based on past production data (lower line and gray one standard deviation band) for Norway (top) and the U.K. (bottom). In both cases, the Monte-Carlo model forecasts a significantly slower decay than the fit by taking into account that new fields will come in production.

which contradicts the prediction based on Hubbert's standard approach, is immediately noticeable.

The results in table 2.2 show a striking difference between the Hubbert model, the extrapolation of the fit and the Monte-Carlo model forecast. The preliminary study discussed in the introduction already showed that the Hubbert model forecast was not stable over time and therefore unlikely to be accurate. Indeed the results confirm that the Hubbert forecast is smaller then fitting the decaying process and only half of what the Monte-Carlo model predicts. This is not surprising, because the Hubbert curve is symmetric and therefore the fitting procedure will be strongly biased by the rapid increase in production before the peak, therefore predicting a rapid decrease. Only fitting the decay process using a stretched exponential resolves this problem, because the decay is entirely decoupled from the increase before the peak. Consequently, we expect the simple fit of the decay to perform better.

According to the fit (extrapolation of aggregate production), Norway's future oil production would still decay much faster than in the Monte-Carlo case. The remaining reserves estimated with the Monte-Carlo methodology are 45% larger than the estimate from the fit. This difference originates from two different effects:

- The sum of the forecast of the individual existing fields is larger than the extrapolation of the aggregate production.
- The extrapolation of the aggregate production does not capture well the discovery process of dwarf fields.

In the U.K., oil production faced a change of regime during the early nineties due to technological innovation, giving rise to the inverted "w shape" of the oil production profile. However, this has not been an issue as this inverted "w shape" can be fitted well by a double logistic (Hubbert model). Alternatively, we can extrapolate the decay of the aggregate production starting at the second peak. The differences between the three methodologies are very similar to the Norwegian case. The Hubbert model again predicts the smallest remaining reserves with about half the amount of the Monte-Carlo model. The fit of the aggregate production underestimates the remaining oil reserves by about 66% compared to the Monte-Carlo model.

Which of the models is more trustworthy? Clearly, the implications in adopting one methodology over the other are significant. The only way to answer this question is to back-test them. In other words: "What would each of the models have predicted, had they been used in the past?" The next section addresses this question and presents the validation step of our approach.

2.5 Validation

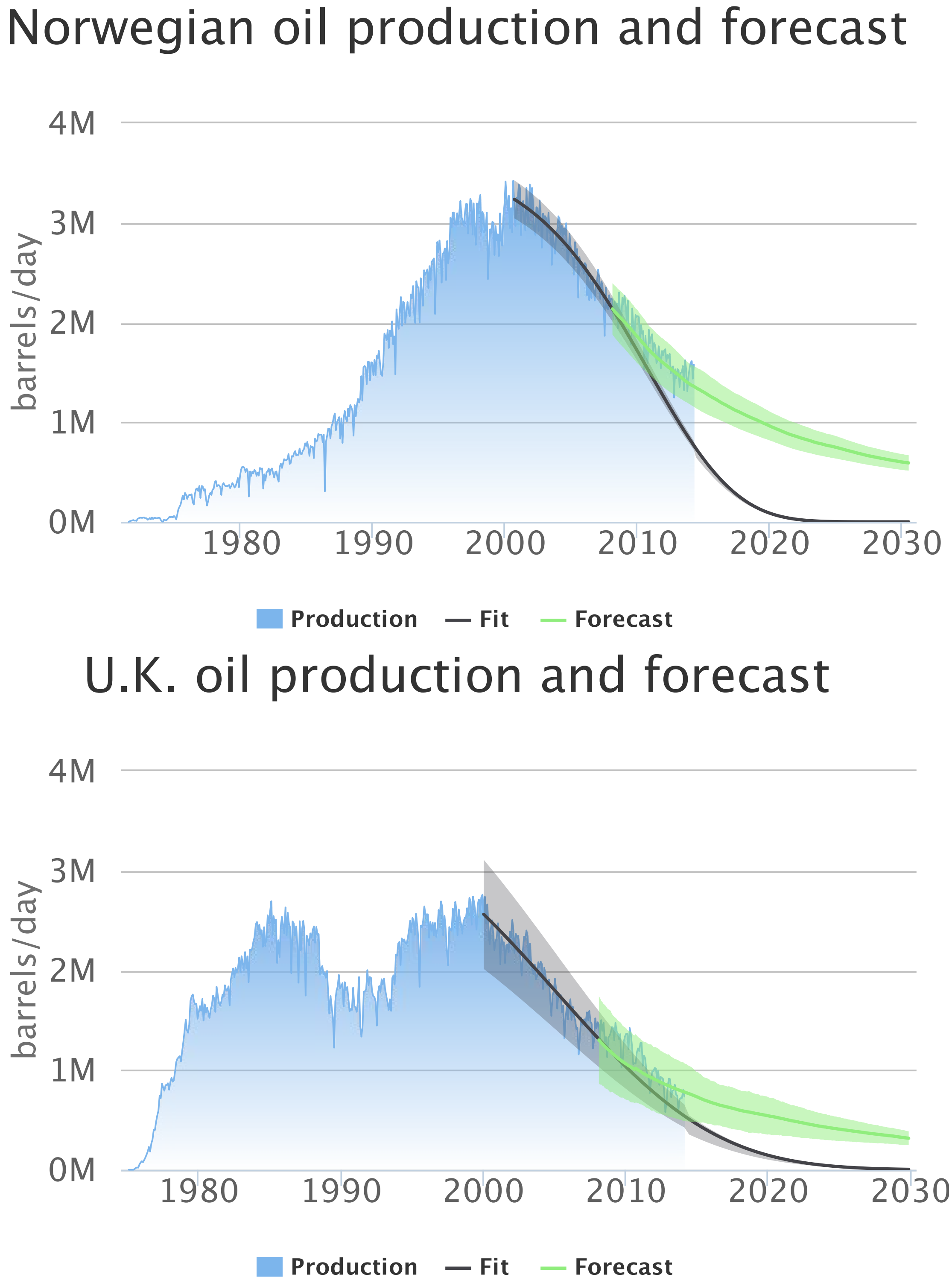


Figure 2.10: Monte-Carlo (green upper continuous line with standard deviation band starting in 2014 onward) and Hubbert forecast based on past production data up to 2008 (lower line and gray one standard deviation band) for Norway (top) and the U.K. (bottom). The results can be compared with the subsequent oil production (blue area). In both cases, the Monte-Carlo methodology is more precise.

Table 2.3: Extrapolation of past oil production ("fit") and prediction using the Monte-Carlo model are made on the data set truncated in 2008. Their forecast for the period 2008-2014 is compared to the actual realized production. All units in billion barrels (Gb).

	Actual (Gb)	Hubbert (Gb)	Fit (Gb)	Model (Gb)
Norway	4.05	3.29	3.28	3.95
U.K.	2.40	1.63	2.00	2.28

Table 2.4: Remaining oil reserves in billion barrels (Gb) forecast for the period 2014-2030 when using the data truncated in 2008, according to the extrapolation of past oil production ("fit") and the Monte-Carlo model. The relative difference between these two predictions is defined by $\Delta = \frac{\text{Model}-\text{Fit}}{\text{Fit}}$.

	Hubbert (Gb)	Fit (Gb)	Model (Gb)	Δ
Norway 08	2.21	0.76	5.20	-584%
U.K. 08	0.73	0.76	2.77	-264%

For both countries, namely Norway and the U.K., a back-test using the data truncated in 2008 has been made. Before that date, too many of the giant fields have not entered their decay phase for a sufficiently long time to apply the extrapolation algorithm. Figure 2.10 shows the resulting production curves for the fit of the aggregate production and of the Monte-Carlo model.

Comparing the forecast of the models with the oil production of the subsequent 6 years shows that the predictive power of the Monte-Carlo model is significantly better than the Hubbert model or than a simple extrapolation of the aggregate past production. Table 2.3 summarizes the difference between the approaches for the back-testing period. The Hubbert model has the poorest performance, underestimating production by 19% (Norway) and 47% (UK). The Monte-Carlo model is found essentially on target with less then 5% error compared with the actual production, while the extrapolation of past production ("fit") is under-estimating the realized production by 19% and 16% respectively for Norway and the U.K. These results confirm that modeling the production dynamics of the individual fields and the discovery of new fields yields a more accurate prediction then methodologies that just use aggregate production.

The errors of the Hubbert model and of the extrapolation of the past oil production method ("fit") are not dramatic over the six years from 2008 to 2014. However, the difference between them and the Monte-Carlo approach becomes huge for the time period from 2014 to 2030, as shown in table 2.4. The Hubbert model and the simple extrapolation decay too fast and entirely miss the fat tails in the decay process of individual fields and the new discoveries. It is interesting to note that the Hubbert model performs similarly to the fit (simple extrapolation) of aggregate production for the UK, but performs much

Table 2.5: Oil import in million barrels per day (Mb/day) at a constant consumption of 1.5 Mb/day for the U.K. and 0.22 Mb/day for Norway. The import for the E.U. and Norway is a lower bound based on the changes in the U.K and Norway. Negative numbers for Norway represent exports.

	2014	2020	2025	2030
Norway (Mb/day)	-1.23	-0.88	-0.58	-0.43
U.K. (Mb/day)	0.7	0.9	1.0	1.1
E.U. + Norway (Mb/day)	9.8	10.45	10.85	11.1

better for Norway. This is due to the fact that the Hubbert model calibrates the decay slope based on the increase before the peak, which in the case of Norway turns out to be beneficial.

The simple extrapolation changed massively between the back-test in table 2.4 and the current fit in table 2.2 (520% for Norway and 236% for the U.K.). In other words, the simple extrapolation is very unstable in its forecast, while the Monte-Carlo forecasts remains very consistent (less than 10% change).

As can be seen in figure 2.10, the actual production of Norway during the back-testing period remained entirely within the quite narrow 1σ interval of the Monte-Carlo methodology, while totally breaking out of the 1σ interval of the simple extrapolation. For the U.K. the Monte-Carlo methodology only performs slightly better when considering the confidence interval, and the confidence interval is much larger due to the uncertainty on future discoveries and their production profile.

2.6 Conclusion

The detailed discussion of a full back-test of the Hubbert model for Norway and the UK showed that its prediction were unstable in time and therefore unlikely to be reliable. We argue that the Hubbert model is likely to be unable to correctly capture the fat tail in the decay process as it does not model the fat tails in the decay of individual fields nor the discovery of new fields. An easily quantifiable link with the oil price, which is the driving factor of oil production, could not be found and we therefore did not pursue the exploration of extensions of the Hubbert model to capture properly the tails of the decay production process.

We preferred a two step Monte-Carlo based methodology to forecast future oil production which: (i) extends the oil production of current fields into the future and (ii) models the discovery rate of new fields. This methodology was then applied to forecast the future oil production of Norway and the U.K. These forecasts are significantly different from the ones obtained with an extrapolation of aggregate production. Indeed, our model forecasts 45% to 66% more remaining oil reserves than the standard extrapolation. The back-test performed on the time period between 2008 and 2014 confirmed that the Monte-Carlo based model better captured the production dynamics.

The results suggest that it is highly likely that the decay of Norwegian and U.K. oil production will be much slower then one would expect from a standard extrapolation. Nonetheless, to maintain current levels of oil consumption in the European Union, more of it will have to be imported from outside Europe, as the imports from Norway will vanish (currently accounting for 11% of E.U. oil imports European Commission, 2014) and the U.K. will need to import more oil.

As shown in table 2.5, at constant consumption, the Monte-Carlo model predicts that in 2030 the E.U. and Norway combined will need to increase their daily oil imports by 1.3 million barrels. These imports will most likely have to come from outside Europe, except for non-standard oil sources yet to be developed.

The present methodology can be applied to many other countries and geological areas, as well as updated at the level of the global oil production. Extensions to include the new wave of shale oil and non-standard oil can in principle be considered and constitute an interesting domain of application of our methodology for the future.

2.7 Three years later

Studying the inherent uncertainty in forecasting the future production of a fundamental resources such as oil is an interesting exercise from the view point of the efficient market hypothesis. For example, the short term forecast (months to a few years) of future oil production for Norway and the UK has a one standard deviation uncertainty of roughly 25%. This implies a 66% difference between an optimistic scenario ($+1\sigma$) and a pessimistic scenario (-1σ). While the short term production fluctuations are buffered by the large oil storage facilities installed all around the world, the medium term forecast leaves ample room for speculation. Consequently, sudden prices drops of 50%, as observed in late 2014 where oil dropped from over 100$ per barrel to below 50$ per barrel, could be well explained by the inherent uncertainties in the fundamental production forecast. As soon as the forecast became precise enough, fundamentals adjusted the price to its correct value. Irrational speculation and herding among traders is not necessary for an explanation. Especially, when considering the complexity of forecasting future oil production at the global level, including geopolitical risks.

The three year delay between the research performed in this chapter and the writing of this thesis makes for a great true out-of-sample test of the methodology. In February 2017, the oil production of Norway, excluding condensates and natural gas liquids not considered in the forecast, stands at $\approx$ 1.6Mb/day, which is near the upper limit of the forecast one standard deviation band. The current oil production of the UK is slightly harder to determine because the previously used source has not been updated since 2015. Nevertheless, other sources report monthly production that fluctuated between $\approx$ 0.77Mb/day and $\approx$ 0.97Mb/day during the second half of 2016. While it is unclear if these production numbers for the UK include or exclude production from condensates and natural gas liquids, they are as well at the upper limit of the forecast one standard deviation band (eventually slightly above). Given the restrictive one standard deviation interval used as a forecast, we can conclude that the methodology is so far robust in a true out-of-sample test. However, the recent surge in production of condensates and natural gas liquids in Norway have not been foreseen.

Chapter 3

Benchmarks for the Efficient Market Hypothesis

3.1 Defining the efficient market hypothesis

The EMH definition popularized by Fama (1970) was based on the concept of a fair game, where based on an information set Ω it is impossible to achieve returns in excess of the equilibrium expected return. This follows from the idea that market prices fully and instantaneously reflect all available information. Mathematically, the returns r_t of a fair game must satisfy

$$\mathrm{E}\left[r_t - \mathrm{E}_{Equilibrium}\left[r_t|\Omega\right]\right] = 0, \tag{3.1}$$

where the inner expectation is taken against the equilibrium model and the outer expectation is taken against the fair game. When the data generating process of the fair game is known, the model for the equilibrium expected return is uniquely defined by the fair game itself. Hence, the EMH definition of Equation (3.1) becomes tautological, as the inner and outer expectation are identical and the difference is necessarily zero. However the data generation process (i.e. game) of real stock market returns is unknown, and the model for the equilibrium expected return is merely a sensible choice with respect to the pricing strategies in use. Therefore, it is unknown if the stock market game is fair and the existence of a trading system that generates profits in excess of the equilibrium expected return is possible.

The dependency of the EMH on the information set Ω, used to build a profitable trading system, has divided the hypothesis into three levels: (i) the weak EMH that restricts the information to only past returns; (ii) the semi-strong form where all publicly available information is considered; and (iii) the strong form that includes private information potentially available to insiders. The literature (Subrahmanyam, 2008) provides evidence challenging the semi-strong and strong form of the EMH. This seems intuitive, as the events that impact markets are first discussed and decided beyond closed doors, and insiders with knowledge of such private information have better then chance odds at predicting future returns. Therefore, the EMH may not be a fair game with respect to private information, but remain a fair game for most investors that are limited to public data.

The martingale formulation of the fair game hypothesis excludes return predictability in excess of the equilibrium returns at any time scale. As a consequence, such a theoretical market bares no arbitrage opportunities that investors can exploit to generate profits. This stands in fundamental opposition with the existence of professional investors that analyze costly information and perform costly transactions in exchange for profits to cover their costs. Without these active market participants their would be no mechanism by which new information is reflected in asset prices. To reconcile this aspect with the EMH, Grossman and Stiglitz (1980) propose the more accurate description of stock markets being in an equilibrium degree of disequilibrium. The fastest and best informed traders obtain a return for their arbitrage activity, which then partially reveals their information to the uninformed traders.

The lack of incentives for traders in a fair game to pursue their arbitrage activity has as well been proven by Milgrom and Stokey (1982) as the no trade theorem. Beyond the no trade issue, the fair game hypothesis as well falls short in describing real markets as it excludes all forms of predictability. However, predictability that is unprofitable after transaction costs of real markets would not be arbitraged. Indeed, the review by Fama (1991) found extensive evidence of predictability and therefore started a shift of the EMH away from a fair game and towards the impossibility of profits in excess of the equilibrium benchmark.

A common benchmark for an equity index is the buy-and-hold strategy. The EMH implies that strategies trading the index cannot generate profits in excess of the buy-and-hold strategy of the market. However, the expected return of portfolio strategies combining the individual assets, which constitute the equity index of interest, is not immediately clear. A simple model for the equilibrium expected return r_t^i of an asset i at time t is the Capital Asset Pricing Model (CAPM) (Sharpe, 1964) given by

$$\mathrm{E}\left[r_t^i\right] = r_t^f + \beta^i \left(\mathrm{E}\left[r_t^m\right] - r_t^f\right), \tag{3.2}$$

where r_i^f is the risk free rate at time t (often constant), the market return r_t^m, and the market sensitivity β^i of the asset i. The market sensitivity is defined by its covariance with the market as

$$\beta^i = \frac{Cov\left(r^i, r^m\right)}{Var\left(r^m\right)} = \rho^{i,m}\frac{\sigma^i}{\sigma^m}, \tag{3.3}$$

where $\rho^{i,m}$ is the correlation with the market, σ^i is the volatility of the asset, and σ^m is the volatility of the market. In the CAPM, an asset with $\beta = 1$ (e.g. the market portfolio itself) has for expected return $\mathrm{E}[r] = \mathrm{E}[r^m]$ as it correlates exactly with the market. An asset has higher expected return then the market only when $\beta > 1$, $\mathrm{E}[r_t^m] > 0$ or $\beta \lesssim -1$, $\mathrm{E}[r_t^m] < 0$. The case $|\beta| > 1$ requires $\sigma > \sigma^m$, which implies that the asset has larger volatility then the market. Hence, the CAPM implicitly defines the market price of risk. An efficient portfolio provides highest return for a given level of risk. Returns in excess of the market return are achievable only in exchange of greater risk.

The CAPM provides an elegant theoretical framework to discuss the EMH in the context of portfolio and remains in use by practitioners due to its simplicity. Nonetheless, the CAPM makes several limiting assumptions. The first being that the utility of return versus risk is entirely expressible through the first two moments on the return distribution. This is certainly a debatable assumption given the skewness and fat tails of typical stock returns. A second assumption is again the zero transaction costs necessary for investors to constantly re-balance their portfolio to be optimal. Besides these theoretical limitations, the CAPM has failed empirically with multiple other factors found to influence an assets return. The Fama-French three factor model (Fama and French, 1993) and the four factor model (Carhart, 1997) have identified: a premium for stocks with high book to market value; a premium for small market capitalization; and a momentum factor that is a premium on past winners.

The four factor model has a good empirical support to accurately describe the equilibrium expected return of an asset. Nonetheless, the recurring event of bubbles has cast doubt on the concept of rational and efficient markets (Malkiel, 2003b). During the growth phase of a bubble, investors exhibit irrational exuberance that drives stock prices far above sustainable price over earning ratios. Nevertheless, in the context of new technologies with a large uncertainty on future profits (i.e. dotcom bubble), these valuations can be compatible with a rational model of expected equilibrium prices. When the true expected earnings become more narrowly defined (uncertainty about future profits decreases), while the prices are already far above their true equilibrium price, the arbitrage of the investors adjusts prices with a severe crash. In my perspective, such a "rational" bubble does not violate the EMH as long as no investor could have foreseen the sudden adjustment in the equilibrium expected returns. However, the irrational herding behavior of investors that decouples prices from fundamentals can be observed before the crash, and should allow rational investors to exit the market in time (Shiller, 1981). A strategy that would systematically exit the market before a crash would generate profits in excess of the buy-and-hold strategy and therefore violate the EMH.

Technical trading rules are potential strategies that could beat the buy-and-hold strategy due to their anchoring in psychological price levels that may influence traders in a systematic manner. An example are resistance levels at which investors using technical trading rules will enter or exit an asset because they expect a rebound or reversal. Assuming that a sufficient number of traders believes in these rules and applies them, the market could exhibit predictable return patterns violating the EMH. Indeed, the study by Sullivan et al. (1999) found technical trading rules that were statistically significantly better then the buy-and-hold strategy for the Dow Jones during a 100 year time period ending in 1986. However, the result did not hold true in the following 10 year out-of-sample test period where the same technical trading rules performed below the benchmark. Two explanations for such findings are possible. First, the finding was spurious despite the statistical significance of risk adjusted returns measured by the Sharpe ratio. In particular this is plausible if the strategy has large down side risks that are not capture by the first two moments entering the Sharpe ratio. A more appropriate measure of risk would have reduced the statistical significance. A second explanation is that the finding was genuine, but the publication attracted a sufficient amount of capital into the arbitrage of this inefficiency, which removed this violation of the EMH in later returns.

The typical spuriousness of significantly performing technical trading rules leaves room to debate if they really challenge the EMH hypothesis. A major argument is the lack of connection to fundamental factors explaining why these rules could continue to outperform the benchmark. In contrast, the factor regression models derive directly from fundamentals, for example it is intuitive that stocks with small capitalization have more potential to grow then stocks with a large capitalization. Besides their lack of a fundamental explanation, technical trading rules with significant performance are typically found in a search across a large universe of technical trading rules. Despite the statistical significance being corrected for the data-snooping effect within many rules, it often remains unclear if traders could have selected an outperforming technical rule ex-ante. Without a sensible method to select the winning technical trading rule, the realized profits of that rule could not have been realized in real time. Conclusively, the case of technical trading rules has not yet been strong enough to reevaluate the model for equilibrium expected returns.

The exploration of thousands of technical trading rules shows the inherent conflict between the goal of rejecting the EMH and data-snooping. To find a potential trading strategy that violates the EMH one has to systematically test all sensible strategies. However, the more strategies are tested, the more likely it is to observe a strategy with spurious out-performance. Especially, when multiple spurious strategies are found, there are good chances that one of them will continue to perform spuriously in later out-of-sample tests. The opposite holds true as well, a strategy with genuine skill could be lost in a sea of randomly performing strategies. Ultimately, the space of all possible trading strategies increases exponentially with the duration of interest, and testing every strategy represents an impossible NP-complete problem (Maymin, 2011). Conclusively, rejecting the EMH entails finding a strategy that has statistically significant past abnormal returns, could have been selected ex-ante, and is reasonably simple with a fundamental explanation that should still hold in the future.

Within this thesis, I want to build upon the following definition of the EMH.

Efficient Market Hypothesis: "A market is efficient with respect to the 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 forecasting model in M_t defined over predictor variables in the information set Ω_t and selected using a search technology in S_t." Timmermann and Granger (2004)

This definition is fully operational, nonetheless I consolidate three aspects. First, it is unclear how to determine the profitable search technology in the set S_t. This would require a single top level search technology S_t^*, selecting a search technology in S_t. To avoid such ambiguity, we limit this study to a single search technology. Second, the economic profits must be significant at the 95% confidence level (or higher) with respect to a benchmark such as the buy-and-hold strategy, after adjusting for data-snooping. Otherwise, any spurious economic profit would suffice to claim inefficiency. Third, the abnormal economic profits with respect to the benchmark must be measured with a risk adjusted metric, and the performance assessment period should be sufficient to account for potential down side risks in higher moments of the return distribution. For example, a test period that only contains a bull or bear market should not be considered sufficient.

A consolidated definition of the EMH would read as follows.

Efficient Market Hypothesis (consolidated): A market is efficient with respect to the information set Ω_t, search technology S_t, and forecasting models M_t, if it is impossible to make economic profits **in excess of equilibrium expected profits** by trading on the basis of signals produced from forecasting model in M_t defined over predictor variables in the information set Ω_t and selected using the search technology in S_t. **The abnormal risk adjusted returns of that strategy have to be statistically significant above the 95% confidence level. The measurement period has to be representative of higher moment down side risks not included in risk adjusted performance metrics such as the Sharpe ratio.**

I remark as well that the above definition implicitly assumes that the set of forecasting models includes all evaluated models. An ex-post selection of models could falsify the results by introducing a bias towards the top performing models. Likewise, the search technology should be determined before evaluating the models to avoid any ex-post biases towards some models.

3.2 The difficulty of distinguishing true skill from luck

When considering a large number of strategies some will necessarily perform well by luck alone. Therefore, a crucial question is to determine if the performance of the top strategies is significantly better then luck and can be attributed to skill. To approach this question, let us consider a set of N_{RW} random walks with zero mean and one skilled strategy following a random walk with positive mean $\bar{r}$. Let us now ask, by how much does the skilled strategy have to outperform in order to stand out significantly among the N_{RW} random strategies?

To answer that question, let us consider a sequence $\{r_1, \ldots, r_T\}$ of T returns drawn in $r_t \sim \mathcal{N}(0, \sigma_r = 0.01)$ and random models predicting up or down with a probability of 50% each. The choice of $\sigma_r = 1\%$ corresponds to the typical daily volatility of equity indices. As the random models are always long or short, it follows that their returns are drawn in the distribution $\mathcal{N}(0, \sigma_r)$. An example of hundred random walks during one

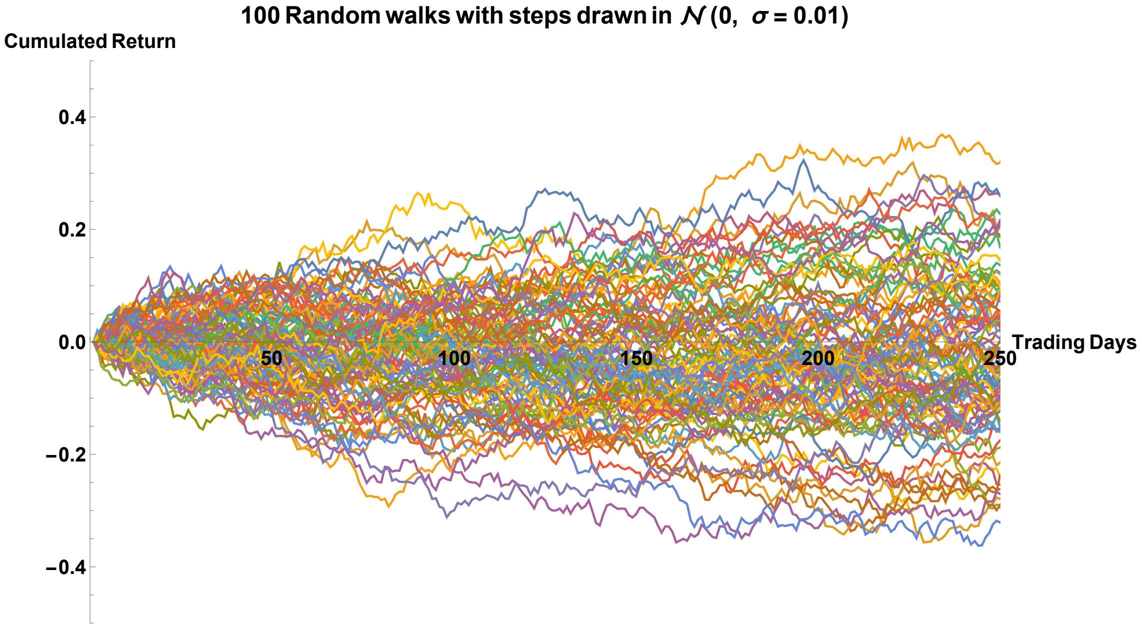


Figure 3.1: **Hundred independent random walks during a one year trading period**. The random walks are generated by drawing returns from the normal distribution $\mathcal{N}(0, \sigma = 0.01)$, where $0.01 = 1\%$ is a typical standard deviation of daily equity index returns.

trading year, which typically has 250 trading days, is shown in Figure 3.1. The simulation shows that the best performing random walks have a cumulative return close to +40%, which could easily be misinterpreted for skilled trading. In real life, thousands of funds are available to investors and distinguishing true skill from luck is even harder.

The expected cumulative return of a single strategy is

$$\mathrm{E}\left[r_1 + \ldots + r_T\right] = 0, \tag{3.4}$$

and the variance is given by

$$\mathrm{Var}\left[r_1 + \ldots + r_T\right] = \sqrt{T}\sigma_r.$$

Hence, the mean returns of the random models after T steps are distributed as

$$\text{random walk mean return} \sim \mathcal{N}\left(0,\, \sigma_{\langle r\rangle} = \frac{\sigma_r}{\sqrt{T}}\right), \tag{3.5}$$

and their daily Sharpe ratios are distributed as

$$\text{random walk Sharpe ratio} \sim \mathcal{N}\left(0,\, \sigma_{SR_d} = \frac{1}{\sqrt{T}}\right). \tag{3.6}$$

The daily Sharpe ratios of the N_{RW} random models are all below the ensemble Sharpe

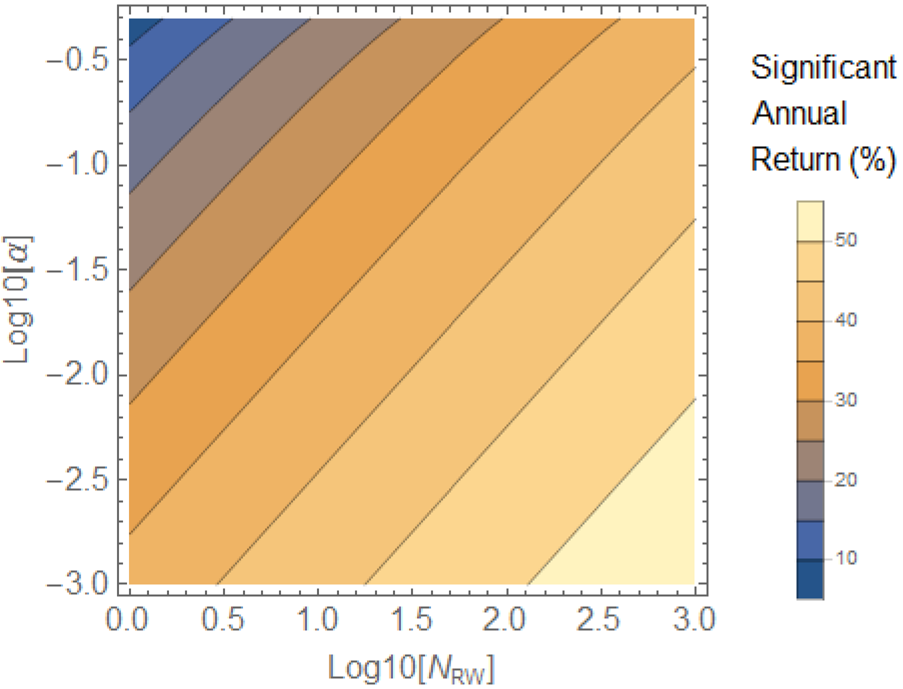


Figure 3.2: **Annual return in percent needed for a strategy to stand out at significance level α among N_{RW} random strategies after one year of trading (i.e. 250 trading days).** The contour plot is generated based on Equation (3.8) with $T = 250$.

ratio of significance level α given by

$$\sigma_{SR_d}^{\alpha,\,N_{RW}} = \Phi^{-1}\left((1-\alpha)^{\frac{1}{N_{RW}}}\right) \cdot \frac{1}{\sqrt{T}}, \tag{3.7}$$

where Φ is the cumulative distribution function of the standard normal distribution, and $\sigma_{SR_d} = \frac{1}{\sqrt{T}}$ has been used. For example, at a significance level of $\alpha = 0.05$, the probability of all $N_{RW} = 100$ Sharpe ratios being below $\sigma_{SR_d}^{0.05,\,100} \approx 0.156$ is equal to 95%. A truly predictive model will stand out as significant at the 95% confidence level when its Sharpe ratio satisfies $SR_d \geq 0.156$. This implies a mean daily return of $\bar{r}_d = SR_d \cdot \sigma_r \geq 0.156\%$, or a yearly cumulative return of $\bar{r}_y \geq 38\%$.

The annual cumulative return needed for a strategy to stand out as significant can be computed as

$$r_{yearly}^{significant} = 250 \cdot \sigma_{SR_d}^{\alpha,\,N_{RW}} \cdot \sigma_r = \frac{250 \cdot \sigma_r}{\sqrt{T}} \cdot \Phi^{-1}\left((1-\alpha)^{\frac{1}{N_{RW}}}\right), \tag{3.8}$$

where $\sigma_{SR_d}^{\alpha,\,N_{RW}}$ is given by Equation (3.7), 250 is the number of trading days in a year, and σ_r is the daily volatility of the market. At fixed significance level α and number of random strategies N_{RW}, the required annual cumulative return required for significance decreases as $T^{-1/2}$. A contour plot of Equation (3.8) for $T = 250$ and $T = 2500$ is shown in Figure 3.3, respectively Figure 3.3. Isometric lines of constant significant return are

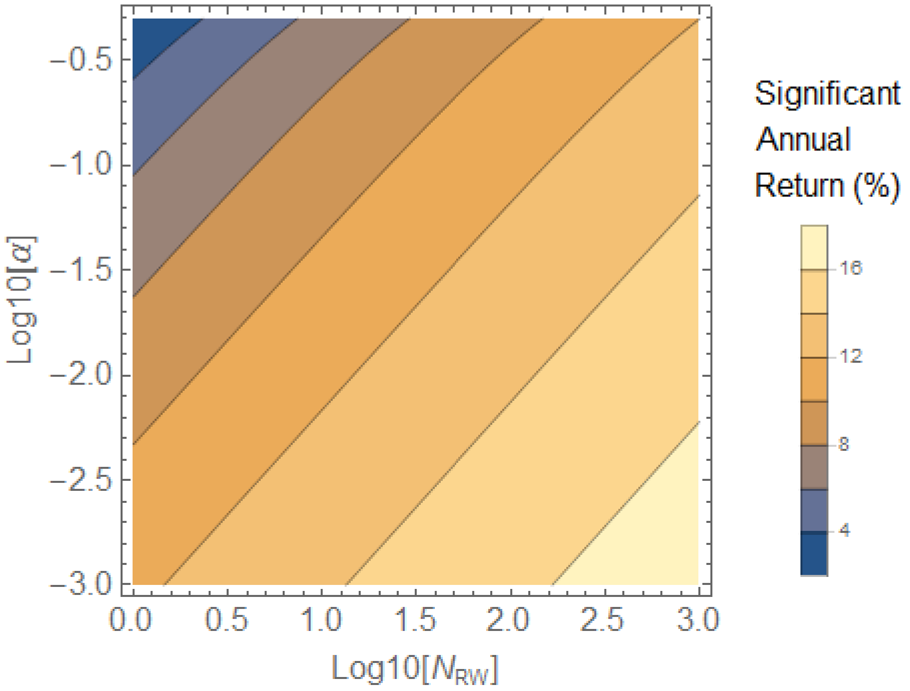


Figure 3.3: **Annual return in percent needed for a strategy to stand out at significance level α among N_{RW} random strategies after ten year of trading (i.e. 2500 trading days).** The contour plot is generated based on Equation (3.8) with $T = 2500$.

roughly linear in a log-log scale of α and N_{RW}.

The random walk model discussed in this section allows us to build a good intuition for the typical time spans and return levels required for a strategy to emerge as significant among many random strategies. However, in real life applications a number of factors complicate the computation of the significance level of an observed strategy. Most importantly, the market returns are not normally distributed and independent. Sharpe ratio estimations have to be corrected for autocorrelations and heteroskedasticity. As well, the strategies are typically not independent either, and bootstrap simulations are required to computed significance levels controlling for the Familywise Error Rate (FWE).

3.3 Performance metrics

3.3.1 Directional accuracy

When calibrating a forecasting model, a simple cost function to maximize is the number of correct predictions. In the context of trading, a model should correctly forecast up and down moves. This directional accuracy is occasionally used in the finance literature (White, 2000), but is of limited interest by itself as it is not a good proxy of risk adjusted returns. In contrast, directional accuracy is widely used in statistical learning (machine learning) for classification problems. For a binary classification problem, the predictions (= forecasting) results fall into four types as presented in Table 3.1: true positives; false

		Predicted Class	
		+	-
Class	+	True positive	False negative
	-	False positive	True negative

Table 3.1: **Confusion matrix of a binary classification problem.**

negatives; false positives; and true negatives. When the cost (=loss) of false positives and false negatives is equal, the directional accuracy is a good metric to optimize. However, in cases like diagnosing diseases, the cost of false negatives is much higher then the cost of a second examination for a false positive. In such asymmetric scenarios, the directional accuracy is unsuited as a performance metric and a problem specific utility function derived from the directional accuracy should be used as a cost function.

Depending on the problem context, one may want to maximize the sensitivity, specificity or Brier score. The sensitivity expresses the fraction of correctly predicted positives over the total number of positives in the population. The specificity expresses the fraction of correctly predicted negatives over the total number of negatives in the population. The Brier score can be used when the prediction probability is available, where a probability of zero means total certainty of class zero and probability one means total certainty of class one. The Brier score is than defined as the mean squared error between the predicted probability and the true label.

The widespread use of the directional accuracy in the statistical (machine) learning research has lead to its use in finance related applications of statistical learning. For example, the forecasting performance of daily returns of the S&P 500 by James et al. (2014) is given as an accuracy of 56%. Unfortunately, a better then random directional accuracy does not necessarily imply that a trading strategy derived from the predictions would have been profitable. Let us now discuss the necessary conditions for directional accuracy to equate with trading performance by considering the directional accuracy $p_d(r)$ of correctly predicting a return with amplitude r. The expected return (i.e. mean return) can then be given as

$$E[r] = \int_{-\infty}^{+\infty} |r| \cdot p(r) \cdot \left(p_d(r) - \frac{1}{2} \right), \tag{3.9}$$

where $p(r)$ is the probability of a return of amplitude r in the predicted return sequence. When predicting up or down moves, the expected return is expressed in terms of the probability of correctly predicting an up move (p_d^+) and a down move (p_d^-) as

$$E[r] = \left(\frac{1}{2} - p_d^- \right) \int_{-\infty}^{0} r \cdot p(r) + \left(p_d^+ - \frac{1}{2} \right) \int_{0}^{+\infty} r \cdot p(r), \tag{3.10}$$

which equals the exact expected return of Equation (3.9) if and only if

$$p_d^- = \frac{\int_{-\infty}^{0} r \cdot p(r) \cdot p_d(r)}{\int_{-\infty}^{0} r \cdot p(r)} \text{ and } p_d^+ = \frac{\int_{0}^{+\infty} r \cdot p(r) \cdot p_d(r)}{\int_{0}^{+\infty} r \cdot p(r)}. \tag{3.11}$$

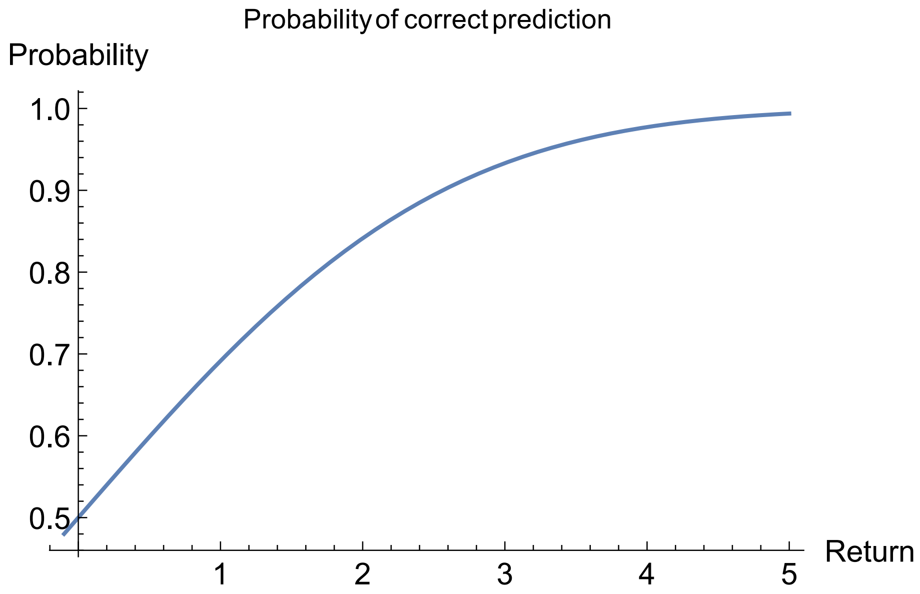


Figure 3.4: **Probability of correctly predicting an up move for an autoregressive process of order one AR**(1) **with** $\phi = 0.5$ **and** $\sigma = 1$**.** The probability is computed based on Equation (3.14). Due to the autocorrelation, large returns are more likely to be correctly predicted then small returns.

These condition are for example satisfied when the directional accuracy is independent of the positive or negative return amplitude, so given by

$$p_d(r) = \begin{cases} p_d^+ & r > 0 \\ p_d^- & r \leq 0 \end{cases}. \tag{3.12}$$

This condition may work as a decent first order approximation, but is already violated for basic data generative processes. Let us for example look at the autoregressive model of order one AR(1) given by

$$r_{t+1} = \phi r_t + a_t, \tag{3.13}$$

with intercept zero, autoregressive parameter ϕ and normally distributed innovation $a_t \sim \mathcal{N}(0, \sigma)$. When $\phi > 0$ and $r_t > 0$ the inequality $P(r_{t+1} r_t > 0) > P(r_{t+1} r_t < 0)$ holds true due to the normally distributed innovation with mean zero, and the prediction is made according to the sign of r_t. Therefore, at fixed r_{t+1}, the probability of correctly predicting the sign is given by

$$\begin{aligned} p_d(r_{t+1} | r_{t+1} > 0) &= \frac{\int_0^{+\infty} \mathcal{N}(r_{t+1} - \phi r_t | 0, \sigma) \mathcal{N}(r_t | 0, \sigma_z) \, dr_t}{\int_{-\infty}^{+\infty} \mathcal{N}(r_{t+1} - \phi r_t | 0, \sigma) \mathcal{N}(r_t | 0, \sigma_z) \, dr_t}, \qquad (3.14) \\ &= \frac{1}{2} \left(1 + \mathrm{Erf}\left(\frac{\phi}{\sigma\sqrt{2}} r_{t+1}\right)\right), \qquad (3.15) \end{aligned}$$

where $\sigma_z^2 = \frac{\sigma^2}{1-\phi^2}$ is the variance of the autoregressive returns. The prediction probability is not constant in r_{t+1}. The prediction probability as a function of the return amplitude is shown in Figure 3.4. Due to the autocorrelation, the probability of correctly predicting a return increases with the return amplitude. Very large returns are almost always predicted correctly.

A performance metric that is a better proxy for the mean return can be constructed from an accuracy metric based on more then two classes (i.e. up and down moves). One picks n_r classes that provide a good discrete approximation of the probability distributions $p\left(r\right)$, and n_c classes that provide a good approximation of $p_d\left(r\right)$. A good discrete approximation of these two distributions ensures that the mean return computed as in Equation (3.9) is well approximated too. The multi-class prediction accuracy can then be computed using the non-parametric test of predictive performance by Pesaran and Timmermann (1992).

This non-parametric test relies on a contingency matrix O with n_r rows for the true returns and n_c columns for the predicted returns. Each sample in the dataset is an observation of a true and a predicted return, which can be allocated to the corresponding cell of the contingency matrix O. Under the null hypothesis of random strategies the true and predicted returns are independent. For sufficiently large samples, Pearson's chi-square test can be used to measure the independence of observations (Pearson, 1900, Plackett, 1983). For contingency matrix O, with observation numbers $O_{i,j}$, the chi-squared is defined as

$$\hat{\chi}_d^2 = \sum_{i=1}^{n_r}\sum_{j=1}^{n_c} \frac{\left(O_{i,j} - \mathrm{E}_{i,j}\right)^2}{\mathrm{E}_{i,j}}, \tag{3.16}$$

where

$$\mathrm{E}_{i,j} = N\rho_{i\cdot}\rho_{\cdot j}, \quad N = \sum_{i,j}^{n_r,n_c} O_{i,j}, \quad \rho_{i\cdot} = \frac{1}{N}\sum_{j=1}^{n_c} O_{i,j}, \quad \text{and} \quad \rho_{\cdot j} = \frac{1}{N}\sum_{i=1}^{n_r} O_{i,j}. \tag{3.17}$$

The number of degrees of freedom of the contingency table is given by $n_f = (n_r-1)(n_c-1)$. The value $\hat{\chi}_d^2$ is used as the performance measure for directional accuracy. Under the null hypothesis of independence, the p-value of the estimated chi-square is

$$P\left(\chi_d^2 \geq \hat{\chi}_d^2,\, n_f\right) = \Gamma\left(\frac{n_f}{2}\right)^{-1} \Gamma\left(\frac{n_f}{2}, \frac{\hat{\chi}_d^2}{2}\right). \tag{3.18}$$

The p-value is the probability of a directional chi-squared value χ_d^2 larger then the observed value $\hat{\chi}_d^2$. This result is useful to approximate efficiently the directional accuracy p-value, when a Monte-Carlo approach is not feasible because of computational limits.

In the case of binary classes $n_r = n_c = 2$, the directional p-value is given by

$$P\left(\chi_d^2 \geq \hat{\chi}_d^2,\, 1\right) = \sqrt{\pi}\Gamma\left(\frac{1}{2}\right)^{-1} \mathrm{Erfc}\left(\sqrt{\frac{\hat{\chi}_d^2}{2}}\right). \tag{3.19}$$

This result shows that in the binary case the $\hat{\chi}_d^2$ value is normally distributed with mean zero.

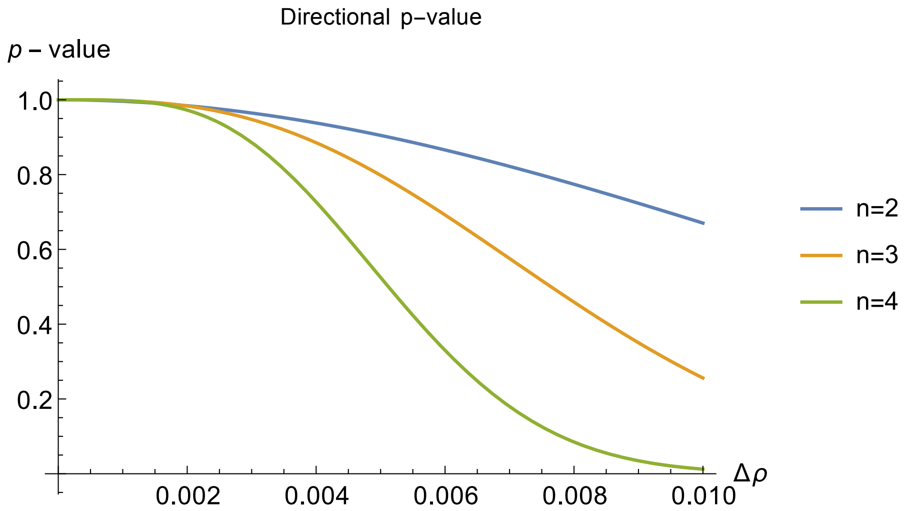


Figure 3.5: **Directional p-value for $n = \{2, 3, 4\}$ and $N = 250$ days as given by Equation (3.18) when using the chi-squared expression of Equation (3.22).** An excess accuracy of $\Delta\rho = 1\%$ in each class is already highly significant in the four class case, but far from significant in the two and three class cases.

The chi-squared value is not particularly intuitive to visualize, in particular in a multi-class context. To gain a better intuition, let us introduce the excess predictability

$$\Delta\rho_{i,j} = \frac{1}{N}\left(O_{i,j} - E_{i,j}\right). \tag{3.20}$$

In the case of an equal number $n = n_r = n_c$ of true and prediction classes, the excess predictability $\Delta\rho_{i,j}$ is a square matrix. Further on, let us assume that the excess predictability takes the homogeneous form

$$\Delta\rho_{i,j} = \left(\frac{n}{n-1}\delta_{i,j} - \frac{1}{n-1}\right)\Delta\rho, \tag{3.21}$$

which satisfies the probability constraints $\sum_i \rho_{i,j} = 1$ and $\sum_j \rho_{i,j} = 1$ following from Equation (3.17). The excess predictability $n\Delta\rho/(n-1)$ in all correct predictions is homogeneously balanced by the reduced predictability $\Delta\rho/(n-1)$ in all wrong predictions. It has to be noted that in the binary case $n = 2$ this is the only possible expression for $\Delta\rho_{i,j}$, while for $n > 2$ in-homogeneous excess predictability can arise. Now the chi-squared can be computed as

$$\hat{\chi}_d^2 = \sum_{i=1}^{n_r}\sum_{j=1}^{n_c} \frac{n^2}{N}\left(N\Delta\rho_{i,j}\right)^2 = n^4 N \Delta\rho^2, \tag{3.22}$$

where $E_{i,j} = \frac{N}{n^2}$ has been used. Some examples of Equation (3.18) when using the chi-squared expression of Equation (3.22) are shown in Figure (3.5).

3.3.2 Directional accuracy & compounded returns

Directional accuracy provides a good indicator to determine if a model possesses some predictive power, and allows us to efficiently compute the associated p-value as given in Equation (3.18). However, as discussed in Section (3.3.1), better than luck directional accuracy does not necessarily imply profitability of a trading strategy. Typically, an investor reinvests his returns and is therefore primarily interested in the compounded return.

Hereafter, the compounded return at time t is designated as wealth W_t, which is normalized to unit initial wealth ($W_1 = 1$). At each time step t the wealth W_t is invested based on a model's signal $s_t \in \mathbb{R}$, compounding with the asset return r_t. When the absolute signal is different from one ($|s_t| \neq 1$), the strategy's remaining cash or leveraged position is invested respectively borrowed at the risk free rate r_t^f. The transaction cost $\Delta\varsigma$ is in percentage points of the change in position. For ease of computation, the transaction cost is assumed to be constant in time. For example, the round trip cost of taking a long position and then selling it is $2\Delta\varsigma$ percentage points. The wealth at time t can so be expressed as

$$W_t = \prod_{i=1}^{t-1} \underbrace{\left(1 - |s_i - s_{i-1}|\,\Delta\varsigma\right)}_{\text{transaction cost}} \underbrace{\left(1 + s_i \cdot r_i + (1 - |s_i|) \cdot r_{\mathrm{i}}^f\right)}_{\text{model return}}, \tag{3.23}$$

where $s_0 = 0$ is an out of the market signal before the first predicted signal s_1.

To understand how directional accuracy impacts compounded wealth, we compute an analytic approximation of the expected wealth as a function of the trading pattern, the contingency matrix, and the transaction cost. The number of observations of a true return of class i and a predicted return of class j is denoted by $O_{i,j}$. The total number of observations is $N = \sum_{i,j} O_{i,j}$. The directional accuracy is assumed constant in time and independent of the return amplitude. In a first step, we approximate the logarithm of the compounded wealth as

$$\begin{aligned}
\log\left(W_N\right) &= \sum_{t=1}^{N-1} \log\left(1 - \Delta s_t \Delta\varsigma\right) + \log\left(1 + s_t \cdot r_t + (1 - |s_t|) \cdot r_t^f\right) && (3.24)\\
&\sim \sum_{i,j} n_{\Delta\mathfrak{s}_{i,j}} \cdot \log\left(1 - \Delta\mathfrak{s}_{i,j}\Delta\varsigma\right) && (3.25)\\
&+ \sum_{i,j} O_{i,j} \cdot \left(\log\left(1 + \mathfrak{s}_j \cdot \overline{r}_i\right) + \log\left(1 + \frac{(1 - |\mathfrak{s}_j|)}{1 + \mathfrak{s}_j \cdot \overline{r}_i} \cdot r^f\right)\right). && (3.26)
\end{aligned}$$

We denote by $\mathfrak{s}_i$ the signal of class i in the contingency matrix, by $\Delta\mathfrak{s}_{i,j} = |\mathfrak{s}_j - \mathfrak{s}_i|$ the absolute difference between two signal classes, and by $n_{\Delta\mathfrak{s}_{i,j}}$ the number of transitions between a signal of class i and a signal of class j. The average asset return of category i is denoted by $\overline{r}_i$, and the average risk free rate by r^f. Using the excess predictability of

Equation (3.20) the expression simplifies to

$$\log\left(W_T\right) \quad \sim \quad W_{\Delta\varsigma} + W_f + I^p_\Sigma \cdot r_m + N\sum_{i,j} \Delta\rho_{i,j} \log\left(1 + \mathfrak{s}_j \overline{r}_i\right), \tag{3.27}$$

where

$$W_{\Delta\varsigma} = \sum_{i,j} n_{\Delta\mathfrak{s}_{i,j}} \cdot \log\left(1 - \Delta\mathfrak{s}_{i,j}\Delta\varsigma\right) \tag{3.28}$$

is the transaction cost, and

$$W_f = \sum_{i,j} O_{i,j} \log\left(1 + \frac{(1 - |\mathfrak{s}_j|)}{1 + \mathfrak{s}_j \cdot \overline{r}_i} \cdot r^f\right) \tag{3.29}$$

is the risk free wealth and/or interest payed on leveraged positions. The third term is the product of the average position and the average market return defined as

$$I^p_\Sigma = N\sum_j \rho_{\cdot j}\mathfrak{s}_j, \quad \text{respectively} \quad r_m = \frac{1}{n_c}\sum_{i,j} \rho_{i\cdot} \log\left(1 + \mathfrak{s}_j \cdot \overline{r}_i\right), \tag{3.30}$$

where the values $\rho_{i.}$ and $\rho_{.j}$ have been defined in equation 3.17 as $\mathrm{E}_{i,j} = N\rho_{i.}\rho_{.j}$. Having more, or stronger, long positions then short positions ($I^p_\Sigma > 0$) in a bull market ($r_m > 0$) will always result in positive returns. The opposite holds true for more, or stronger, short positions in a bearish market.

The fourth term $\sum_{i,j} \Delta\rho_{i,j} \log\left(1 + \mathfrak{s}_j\overline{r}_i\right)$ expresses the return stemming from the excess predictability. This term is constrained by the equalities $\sum_i \Delta\rho_{i,j} = 0$ and $\sum_j \Delta\rho_{i,j} = 0$. These equalities follow from the definitions of $\rho_{i.}$ and $\rho_{.j}$ given in equation 3.17, which imply $\sum_i \rho_{i.} = 1$ and $\sum_j \rho_{.j} = 1$. To further simplify this term in the case of an equal number $n = n_r = n_c$ of true and predicted returns, the excess predictability is assumed to take the homogeneous form defined in Equation (3.21). The assumption of homogeneous excess predictability transforms the last term into

$$N\frac{n}{n-1}\Delta\rho\sum_i \log\left(1 + \mathfrak{s}_i\overline{r}_i\right) - \frac{N}{n-1}\Delta\rho\sum_{i,j} \log\left(1 + \mathfrak{s}_j\overline{r}_i\right). \tag{3.31}$$

In the case of balanced signals, satisfying $\forall\mathfrak{s}_j \in \mathfrak{s}$, $-\mathfrak{s}_j \in \mathfrak{s}$, the second term can be neglected up to second order in $\overline{r}_i$.

Finally, introducing the average return $r_c = \frac{1}{n}\sum_i \log\left(1 + \mathfrak{s}_i\overline{r}_i\right)$ of a correct prediction, the compounded wealth can be expressed as

$$\log\left(W_N\right) \sim W_{\Delta\varsigma} + W_f + I^p_\Sigma \cdot r_m + N\frac{n^2}{n-1}\Delta\rho \cdot r_c. \tag{3.32}$$

The last term in equation 3.32 is the expected profit made from the excess predictability, which is proportional to the average profit r_c of a correct prediction and the excess predictability $\Delta\rho$. This result is rather intuitive, as making $\Delta\rho$ percent more correct than wrong predictions, with a gain or loss of r_c on each prediction, must yield an overall gain proportional to $\Delta\rho \cdot r_c$. If the compounded wealth differs significantly from the value obtained in equation 3.32, the directional accuracy is not independent of the return amplitude, as assumed in this computation.

To finally compare the transaction costs and excess profits, let us consider a random trading strategy going only long or short ($n = 2$). A random sequence of binary signals, with identical probability, trades on average every second day. Consequently, the first term evaluates to $\frac{N}{2}\log(1 - 2\cdot\Delta\varsigma)$, as $\Delta\mathfrak{s}_{i,j} = 2\delta_{i,j}$ and $n_{\Delta\mathfrak{s}_{i,j}} = \frac{N}{4}$. The risk free wealth is zero ($W_f = 0$), as the random strategy is always long or short. The average return of a correct prediction is equal to the average absolute market return ($r_c = \langle\log(1 + |r|)\rangle$). Therefore, in the case of a flat market ($r_m = 0$), the profits from excess directional accuracy exceed the trading costs if

$$-\frac{1}{2}\log(1 - 2\cdot\Delta\varsigma) + 4\Delta\rho\cdot r_c > 0 \;\Rightarrow\; \Delta\rho \gtrsim \frac{\Delta\varsigma}{4r_c}. \tag{3.33}$$

A typical equity index, such as the S&P 500, has an average absolute daily return $\langle|r|\rangle \sim 1\%$. This implies that for every basis point in spread between buy and sell (spread $= 2\cdot\Delta\varsigma$) at least $100 \times \frac{0.005\%}{4\times 1\%} = 0.125\%$ in excess predictability $\Delta\rho$ is needed. The excess predictability $\Delta\rho$ is linked to the Chi-squared by Equation (3.22) as $\chi^2_d = 16N\Delta\rho^2$, assuming $n = 2$.

3.3.3 Directional accuracy & mean return

In certain circumstances, one may not be interested in the compounded returns, but instead in the cumulative return or mean return. This occurs when a fixed sum is invested at every time step. Profits are not reinvested and losses are compensated at every step. The relation between directional accuracy and cumulative return is analogous to the compounded wealth derived in Section 3.3.2, replacing log returns by the returns. The cumulative return r_t at time t is given by

$$r_t = \sum_{i=1}^{t-1} \underbrace{s_i \cdot r_i + (1 - |s_i|)\cdot r^f_{\mathrm{i}}}_{\text{model return}} - \underbrace{|s_i - s_{i-1}|\,\Delta\varsigma}_{\text{transaction cost}}, \tag{3.34}$$

where $s_0 = 0$ is an out of the market signal before the first predicted signal s_1.

The number of observations of a true return of class i and a predicted return of class j is denoted by $O_{i,j}$. The total number of observations is $N = \sum_{i,j} O_{i,j}$. The directional accuracy is assumed constant in time and independent of the return amplitude. The cumulative return is given by

$$r_N \sim r_{\Delta\varsigma} + r_f + I^p_\Sigma \cdot r_m + N\frac{n^2}{n-1}\Delta\rho\cdot r_c, \tag{3.35}$$

which is an identical form to the log wealth of Equation (3.32). The term

$$r_{\Delta\varsigma} = -\Delta\varsigma \sum_{i,\,j} n_{\Delta\mathfrak{s}_{i,\,j}} \Delta\mathfrak{s}_{i,\,j} \tag{3.36}$$

is the transaction cost, and

$$r_f = r^f \sum_{i,\,j} O_{i,\,j} \left(1 - |\mathfrak{s}_j|\right) \tag{3.37}$$

is the risk free return and interest payed on leveraged positions. The third term is the product of the average position and the average market return now defined as

$$I_{\Sigma}^{p} = N \sum_{j} \rho_{\cdot j} \cdot \mathfrak{s}_j, \quad \text{respectively} \quad r_m = \frac{1}{n_c} \sum_{i,\,j} \rho_{i\cdot} \cdot \mathfrak{s}_j \cdot \overline{r}_i, \tag{3.38}$$

where the values $\rho_{i.}$ and $\rho_{.j}$ have been defined in equation 3.17 as $\mathrm{E}_{i,\,j} = N\rho_{i.}\rho_{.j}$. The last term is the profitability resulting from the excess predictability, which depends on the average return $r_c = \frac{1}{n}\sum_i \mathfrak{s}_i \cdot \overline{r}_i$ of a correct prediction.

3.3.4 Sharpe ratio

A trading strategy's profit is given by its wealth or cumulative return after transaction costs, depending on the investment style. While the profitability is an important aspect, it does not provide any information about the inherent risk of a strategy. As discussed in Section 3.1, the EMH states that a strategy cannot make risk adjusted profits in excess of the benchmark. Asset pricing models such as the CAPM implicitly define the price of risk. Consequently, a better metric measuring risk adjusted returns is needed.

Most commonly, risk adjusted returns are measured by the Sharpe ratio of a strategy. The Sharpe ratio for a strategy with returns $\{r_t\}_T$ is given by

$$SR_t = \frac{\mathrm{E}\left(r_t\right) - \mathrm{E}\left(r_t^f\right)}{\sqrt{\mathrm{E}\left(r_t^2\right) - \mathrm{E}\left(r_t\right)^2}}. \tag{3.39}$$

The Sharpe ratio provides risk adjusted returns assuming that the return distribution is stationary and entirely characterized by the first two moments. In the context of financial time series none of these assumptions hold strictly speaking, but they provide a good approximation. In any case, the true generative process of returns is unknown and moments have to be estimated from the finite sample available through back-testing. Finite sample properties of the Sharpe ratio are discussed in detail by Lo (2002). To maximize the sample size, and so minimize the estimation error, this thesis uses the daily Sharpe ratio unless stated otherwise. Monthly or yearly Sharpe ratios have a large estimation error and are not as reliable.

3.3.4.1 Studentized Sharpe ratio test statistic

The Sharpe ratio is used to determine the risk adjusted performance of the analyzed strategies. Hence, to test the EMH, one has to show that the Sharpe ratio of a strategy is significantly better then the Sharpe ratio of the benchmark, and therefore the test statistic of interest is the Sharpe ratio difference with respect to the benchmark. The following computation of the test statistic is based on my interpretation and implementation of the method developed by Ledoit and Wolf (2008).

Assuming two return sequences $(X_{1,1}, \ldots, X_{T,1})$ and $(X_{1,2}, \ldots, X_{T,2})$ of length t, with estimated means $(\hat{\mu}_1, \hat{\mu}_2)$ and variances $(\hat{\sigma}_1^2, \hat{\sigma}_2^2)$, the estimated test statistic reads

$$\hat{\Delta}\left(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1, \hat{\sigma}_2\right) = \hat{\mathrm{Sh}}_1 - \hat{\mathrm{Sh}}_2 = \frac{\hat{\mu}_1}{\hat{\sigma}_1} - \frac{\hat{\mu}_2}{\hat{\sigma}_2}. \tag{3.40}$$

This estimated test statistic has to be taken cautiously because it can be biased in the case of correlated or heteroskedastic returns, and has a significant variance in finite samples. To mitigate these biases, it is better to use the studentized test statistic

$$\hat{\Delta}_S = \frac{\hat{\Delta}}{s\left(\hat{\Delta}\right)}, \tag{3.41}$$

corrected for the estimation error $s(\hat{\Delta})$. Considering the vector of estimated moments $\hat{\nu} = (\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1, \hat{\sigma}_2)$, with covariance matrix Ψ, the estimation error is computed as

$$s\left(\hat{\Delta}\right) = \sqrt{\frac{\nabla_\nu \hat{\Delta}\left(\hat{\nu}\right)^T \Psi \nabla_\nu \hat{\Delta}\left(\hat{\nu}\right)}{T}}, \tag{3.42}$$

where $\nabla_\nu \hat{\Delta}\left(\hat{\nu}\right)$ is the gradient with respect to the moments ν of the test statistic defined in Equation (3.40), and the covariance matrix of the moments is given by $\Psi = \lim_{T\to+\infty} E\left[\hat{\nu}^T \hat{\nu}\right]$.

3.3.4.2 Robust covariance estimation

In finite samples, the covariance matrix Ψ needed to estimate the studentized test statistic defined in Equation (3.41) may be subject to biases as well. A HAC robust estimation of Ψ is obtained with a kernel k as

$$\hat{\Psi} = \frac{t}{t-4} \sum_{j=-t+1}^{t-1} k\left(\frac{j}{S_T}\right) \hat{\Gamma}_t\left(j\right), \tag{3.43}$$

where

$$\hat{\Gamma}_T\left(j\right) = \begin{cases} \frac{1}{t} \sum_{t=j+1}^{t} \hat{\nu}_t^T \hat{\nu}_{t-j} & j \geq 0 \\ \frac{1}{t} \sum_{t=-j+1}^{t} \hat{\nu}_{t+j}^T \hat{\nu}_t & j < 0 \end{cases}. \tag{3.44}$$

Within this paper we use the Quadratic-Spectral (QS) kernel, for which the optimal bandwidth S_T^* can be computed using automatic methods derived by Andrews (1991) and Newey and West (1994). The optimal bandwidth for the QS-kernel reads

$$S_T^* \approx 1.32\,(a \cdot T)^{2/5}\,, \tag{3.45}$$

where a is a constant dependent on the DGP (e.g. AR(ϱ), ARMA(ϱ, q), MA(ϱ)). When using real financial data, we first regress an appropriate model on the data, and then compute the constant a based on simulations with the regressed model.

3.4 Statistical significance & multiple testing

3.4.1 Methodology setup

A major requirement to reject the EMH as defined in this thesis (Definition 3.1) is not only to find a strategy profitable in excess of the benchmark, but that the observed excess profitability is statistically significant given all tested strategies. To compute the statistical significance of a strategy within a universe of strategies, this thesis follows the methodology of Sullivan et al. (1999), White (2000), and Romano and Wolf (2005a). This methodology uses bootstrap simulations (i.e. Monte-Carlo for non statisticians) to compute the statistical significance of a performance metric for a given set of strategies. In particular, it allows to preserve the exact dependency structure across all tested strategies. The algorithm used to adjusted the p-values for multiple testing is defined in Romano and Wolf (2016). This algorithm maximizes the number of rejected null hypotheses without violating the familywise error rate.

A central element of the multiple-testing methodology is the observed data matrix $X_{T,\mathcal{S}+1}$ (i.e. returns), with $1 \leq t \leq T$ time steps of the $1 \leq s \leq \mathcal{S}$ different strategies. The last column $\mathcal{S}+1$ is reserved for the benchmark. A strategy is generically described by its sequence of signals $\{s_t\}_1^T \in \mathbb{R}^T$ determining the position at each time step on the returns $\{X_t\}_1^T$. The entries of the observed data matrix are obtained as $r_{t,s} \equiv X_{t,s} = \{X_t\}_1^T \circ \{s_{t,s}\}_1^T$, where $s_{t,s}$ is the trading signal of strategy s at time t, and $\circ$ denotes the Hadamard product. In matrix form, the observed data matrix reads as

$$X_{T,\mathcal{S}+1} = \begin{bmatrix} r_{1,1} & \cdots & r_{1,\mathcal{S}} & r_{1,b} \\ \vdots & \ddots & \vdots & \vdots \\ r_{T,1} & \cdots & r_{T,\mathcal{S}} & r_{T,b} \end{bmatrix},$$

where each row represents a time step, and each column contains the return of a strategy over time, with the last column containing the benchmark. The benchmark is typically given by the buy-and-hold strategy $r_{t,b} = X_t$.

3.4.2 Null bootstrap of returns

The distribution of a chosen performance metric is computed using bootstrapped realizations of $X_{T,\mathcal{S}+1}$. The unknown probability distribution of returns is typically assumed to be stationary, and new realizations are generated using the circular block bootstrap of Politis and Romano (1992). This bootstrap method samples, with replacement, blocks $X_{t_0 \leq t \leq t_0+b,\, S+1}$ of size b from the observed returns, to generate bootstrapped observations $X^*_{T,\, S+1}$. A bootstrap block is written in matrix form as

$$X_{t_0 \leq t \leq t_0+b,\, S+1} = \begin{bmatrix} r_{t_0,\,1} & \cdots & r_{t_0,\,\mathcal{S}} & r_{1,\,b} \\ \vdots & \ddots & \vdots & \vdots \\ r_{t_0+b,\,1} & \cdots & r_{t_0+b,\,\mathcal{S}} & r_{T,\,b} \end{bmatrix},$$

which is a sub-matrix of $X_{T,\mathcal{S}+1}$ with b rows starting at row t_0. A bootstrapped data matrix $X^*_{T,\,S+1}$ is obtained by stacking row-wise $i \in \left\{1,\, \ldots,\, \left\lceil \frac{T}{b} \right\rceil \right\}$ blocks with uniformly sampled $t_{0,\,i} \sim \text{unif}\left\{1,\, \ldots,\, \left\lceil \frac{T}{b} \right\rceil \right\}$, and removing the last $\left\lceil \frac{T}{b} \right\rceil - T$ rows in case T is not a multiple of b.

Bootstrapping the data breaks the correlation structure between the blocks of size b, which limits the estimation of the covariance matrix to the blocks of size b and could introduce major biases in the estimated test statistic. Nonetheless, this issue can be overcome by finding the optimal block size for a given DGP. The block size selection is performed by computing the estimated test statistic at confidence level α in a situation were the true test statistic is known. For example, let us consider two independent realizations of length t of an AR(ϱ) with given parameter ϕ, the first being the strategy and the second the benchmark. The two realizations have identical expected studentized Sharpe ratio. Therefore, at significance level α, the test statistic should reject for a fraction $1-\alpha$ of the bootstrap realizations the hypothesis that the strategy has superior test statistic then the benchmark. The optimal block size $b\left(\alpha\right)$ is hence a function of the significance level α. The pseudo-code for the block size selection can be found in Ledoit and Wolf (2008).

Given $m = 1,\, \ldots,\, M$ bootstrap resamples $X^{*,\, m}_{T,\, S+1}$ obtained with the optimal block size, the centered studentized test statistic of strategy s for bootstrap m is

$$\hat{\Delta}^{*,\, m}_{S,\, s} = \frac{\left| \hat{\Delta}^{*,\, m}_{s} - \hat{\Delta}_s \right|}{s\left(\hat{\Delta}^{*,\, m}_{s}\right)}, \tag{3.46}$$

where the test statistic is always computed with respect to the benchmark in column $S+1$. The individual p-values are computed as

$$p^{*,\, m}_{s} = \frac{\left\{ \hat{\Delta}^{*,\, m}_{S,\, s} \geq \hat{\Delta}_{S,\, s} \right\}}{M+1}, \tag{3.47}$$

which is the fraction of centered test statistics that exceed the original test statistic. The precision of the computed p-values depends on the number of null resamples M.

For practical purposes $M = 1000$ is sufficient to estimate the p-values at three decimal places, and determine if typical confidence levels of 95% or 99% are reached. Making an assumption on the functional form of the test statistic distribution, one can extrapolate the tails when higher accuracy is needed for extremely significant results. However, the true tail behavior is usually unknown, and the tail estimation obtained by extrapolation of a specific distribution is at risk of large biases.

3.4.3 Adjusting p-values for multiple testing

The individual test statistics computed in Equation (3.47) do not correct for the multiple testing of several strategies simultaneously. To adjust for multiple testing, the stepdown procedure of Romano and Wolf (2016) has to be applied. In a first step, the individual strategies are ordered by increasing p-value. The indices $\{r_1, \ldots, r_S\}$ denote the permutation of $\{1, \ldots, S\}$ that fulfills $p_{r_1} \leq p_{r_2} \leq \ldots \leq p_{r_S}$. In a second step, the smallest p-value in the mth resample of the $S - j$ worst strategies is denoted by

$$\min_{p,j}^{*,m} = \min\left\{p_{r_j}^{*,m}, \ldots, p_{r_S}^{*,m}\right\}.$$

Introducing the p-value $p_0^{adj} = 0$, the adjusted p-values can be computed recursively as

$$p_j^{adj} = \max\left\{\frac{\#\left\{\min_{p,j}^{*,m} \leq p_j\right\} + 1}{M + 1}, p_{j-1}^{adj}\right\}$$

for $j = 1, \ldots, S$.

3.4.4 Null bootstrap of trading signals

A drawback of bootstrapping the returns of the strategies and benchmark is the stationarity assumption. For this assumption to hold, the predicted stock market has to be in the same regime during the entire test period. Unfortunately, real stock markets undergo regime changes, the most common being bull and bear markets. Bootstraps over the past two decades would include realizations that do not contain the dotcom bubble and financial crisis. This consequence of the stationarity condition defeats the purpose of computing p-values with respect to realizations similar to the real market.

To avoid the issue of non-stationarity of the returns, one can instead bootstrap the signals (long or short) of the strategies and keep the returns of the benchmark unchanged in every bootstrap. For each bootstrap, the returns of the strategies are computed based on the benchmark returns and the bootstrapped signals. In this setting, the bootstrapped test statistic is computed as in Equation (3.41), and is not centered. The obtained p-values describe the probability that the randomized strategies with same number of long and short positions, and same cross-strategy correlation structure of the signals, achieve the observed performance of the actual strategies.

The null bootstrap of the trading signals acts on the signal matrix

$$\mathcal{S}_{T,\,\mathcal{S}+1} = \begin{bmatrix} s_{1,\,1} & \cdots & s_{1,\,\mathcal{S}} & s_{1,\,b} \\ \vdots & \ddots & \vdots & \vdots \\ s_{T,\,1} & \cdots & s_{T,\,\mathcal{S}} & s_{T,\,b} \end{bmatrix},$$

which generates the observed data matrix

$$X_{T,\,\mathcal{S}+1} = \mathcal{S}_{T,\,\mathcal{S}+1} \circ \{X_t\}_1^T = \begin{bmatrix} s_{1,\,1} \cdot r_1 & \cdots & s_{1,\,\mathcal{S}} \cdot r_1 & s_{1,\,b} \cdot r_1 \\ \vdots & \ddots & \vdots & \vdots \\ s_{T,\,1} \cdot r_T & \cdots & s_{T,\,\mathcal{S}} \cdot r_T & s_{T,\,b} \cdot r_T \end{bmatrix},$$

where $\circ$ is the Hadamard product and $\{X_t\}_1^T = \{r_t\}_1^T$ are the returns of the predicted asset. The buy-and-hold strategy as benchmark is defined by $s_{i,\,b} = 1,\, \forall i$.

Analogously to the null bootstrap of returns in Section (3.4.2), new realization of the observed data matrix are obtained by generating bootstrapped realization $\mathcal{S}^*_{T,\,\mathcal{S}+1}$ of the trading signals. The bootstrap realizations $\mathcal{S}^*_{T,\,\mathcal{S}+1}$ are obtained by stacking uniformly sampled blocks of size b from the signal matrix $\mathcal{S}_{T,\,\mathcal{S}+1}$. The bootstrap data matrix realizations are obtained as $X^*_{T,\,\mathcal{S}+1} = \mathcal{S}^*_{T,\,\mathcal{S}+1} \circ \{X_t\}_1^T$.

In the context of bootstrapping the strategy signals, instead of the strategy returns, the optimal bootstrap block size becomes more intuitive when considered with respect to the mean position duration. For example, for a strategy with a mean position duration of ten days, the bootstrapped realization should as well have a mean position duration of ten days. Otherwise, the performance of the true strategy would be benchmarked with respect to bootstrapped strategies having a different investment style. In order not to compare apples and pears, the block size used for bootstrapping should preserve the investment style. However, when comparing a set of strategies with different investment style, no block size will preserve the characteristic position distribution of all strategies. Nevertheless, a block size may be selected based on a simulation of a corresponding null hypothesis at given significance level. In any case, the block size should be selected so as to never violate the familywise error rate, the downside being that the power of the test statistic may be sub-optimal for some strategies.

To gain some intuition of the relation between investment style and trading performance, let us study the impact of the mean position duration t_s on the wealth of random strategies. Figure 3.6 shows the wealth quantiles as a function of the mean position duration for the S&P 500. The lower quantiles, below the 95% confidence level, are almost insensitive to the mean position duration. However, larger quantiles decrease significantly with mean position duration. For example, the 99.9% quantile decreases by almost 40% between mean position duration two ($W_{random}\,(t_s = 2) \sim 11$) and mean position duration 20 ($W_{random}\,(t_s = 20) \sim 7$). This decrease implies that the randomization has to be performed with respect to the lowest mean position duration in order to not violate the familywise error rate.

In conclusion, selecting the bootstrap block size with respect to the longest position

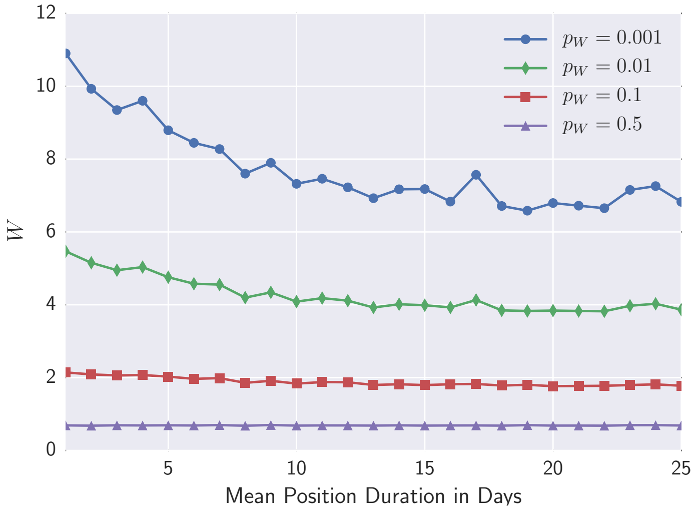


Figure 3.6: **Wealth quantiles of random trading strategies as a function of the mean position duration in days.** The simulation is based on daily returns of the S&P 500 between Jan. 1, 1997 and Dec. 31, 2015. At each duration in [1, 2, . . . , 25], the wealth distribution was estimated on a set of 10000 samples. The tail values were evaluated from a generalized normal distribution fitted to the samples. The roughness of the curves is a result of the particular return structure of the S&P 500, and is not a result of an insufficient sample size.

duration would insufficiently randomize the strategies with shorter position duration. On the other hand, selecting the block size with respect to the shortest position duration would overly randomize strategies with longer position duration. To maximize the statistical power of the randomization test, without violating the familywise error rate, the block size should always be chosen with respect to the strategies with shortest mean position duration.

3.4.5 P-value bias of the mean return

The p-values of certain metrics, for simple prediction strategies, can exhibit non-intuitive biases. For example, let us consider the strategy with signal

$$s_{t+1} = \mathrm{sign}\left(r_t\right), \tag{3.48}$$

which predicts for each time step the sign of the previous return. We benchmark this strategy against the buy-and-hold strategy $s_{t+1}^{BH} = 1$ using the mean return as a performance metric. When predicting two returns (r_0, r_1, r_2), where r_0 is a dummy return needed for the prediction, the mean return difference between the two strategies is

$$\Delta\bar{r} = \underbrace{\frac{1}{2}\left(\mathrm{sign}\left(r_0\right) r_1 + \mathrm{sign}\left(r_1\right) r_2\right)}_{\text{strategy}} - \underbrace{\frac{1}{2}\left(r_1 + r_2\right)}_{\text{buy-and-hold}}. \tag{3.49}$$

$\text{sign}\,(r_0,\, r_1,\, r_2)$	$\Delta\bar{r}$
$(+,\,+,\,+)$	$r_1 + r_2 - r_1 - r_2 = 0$
$(+,\,+,\,-)$	$r_1 + r_2 - r_1 - r_2 = 0$
$(+,\,-,\,+)$	$\frac{1}{2}\left(r_1 - r_2 - r_1 - r_2\right) = \boldsymbol{-r_2 \leq 0}$
$(+,\,-,\,-)$	$\frac{1}{2}\left(r_1 - r_2 - r_1 - r_2\right) = -r_2 \geq 0$
$(-,\,-,\,-)$	$\frac{1}{2}\left(-r_1 - r_2 - r_1 - r_2\right) = -r_1 - r_2 \geq 0$
$(-,\,-,\,+)$	$\frac{1}{2}\left(-r_1 - r_2 - r_1 - r_2\right) = -r_1 - r_2 \sim 0$
$(-,\,+,\,-)$	$\frac{1}{2}\left(-r_1 + r_2 - r_1 - r_2\right) = \boldsymbol{-r_1 \leq 0}$
$(-,\,+,\,+)$	$\frac{1}{2}\left(-r_1 + r_2 - r_1 - r_2\right) = \boldsymbol{-r_1 \leq 0}$

Table 3.2: **Mean return difference between the buy-and-hold and previous sign prediction strategies**. The table lists all cases predicting two returns r_1 and r_2.

While the expected mean return difference is zero, $E\left[\Delta\bar{r}\right] = 0$, the probabilities of it being positive or negative are not symmetric. To show that $p\left(\Delta\bar{r} > 0\right) \neq p\left(\Delta\bar{r} < 0\right)$, we compute the value for all possible scenarios in Table 3.2, and can determine the probabilities to almost surely be

$$p\left(\Delta\bar{r} > 0\right) = \frac{5}{16},\, p\left(\Delta\bar{r} = 0\right) = \frac{4}{16} \text{ and } p\left(\Delta\bar{r} < 0\right) = \frac{7}{16}. \tag{3.50}$$

The p-value associated to the strategy outperforming the benchmark would be $p\left(\Delta\bar{r} < 0\right) + \frac{1}{2}p\left(\Delta\bar{r} = 0\right) = \frac{7}{16} + \frac{1}{2} \cdot \frac{4}{16} = 0.5625$. This bias decreases with the number of predicted returns as shown in Figure 3.7. The bias decays with the sample size T, but remains significant for the sample sizes of interest.

3.5 Portfolio of strategies

Another major requirement to reject the EMH as defined in this thesis (Definition 3.1) is not only to find a strategy with statistically significant profits in excess of the benchmark, but as well a search technology that would have selected this strategy ex-ante among all available strategies. For a large universe of models, a mean variance portfolio as search technology is technically challenging to implement (Senneret et al., 2016). To reasonably invert the covariance matrix of the returns of all models, the number of past returns must be greater than or equal to the number of models. Therefore, the ex-ante performance is evaluated using the simpler best Sharpe ratio search technology, which invests at every time step into the model with the best past Sharpe ratio.

Let us denote by $\vec{r}^M(t)$ the set of returns at time t of the models in the universe M of models. Then a search technology is a function $\vec{S}\left(M,\, t\right)$ computing the weights of a portfolio of models in M at time t, resulting in portfolio returns $r^S(t) = \vec{S}\left(M,\, t-1\right) \cdot \vec{r}^M(t)$. Denoting the Sharpe ratio of model $M_i \in M$ at time t by $SR_i\left(t\right)$, the best Sharpe

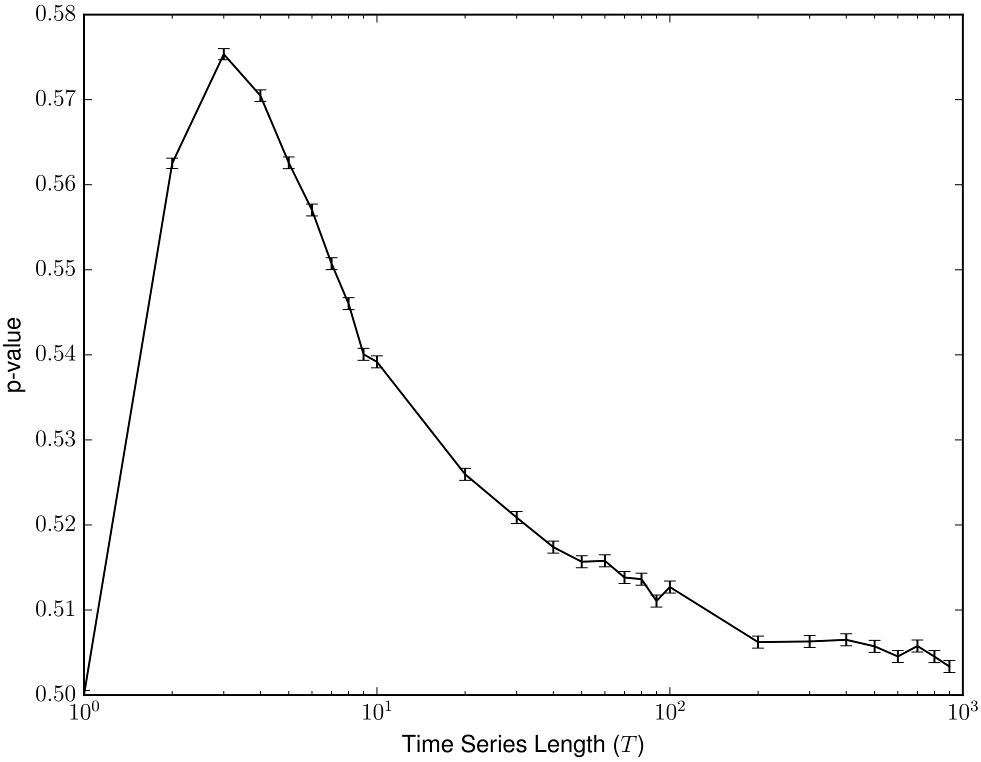


Figure 3.7: **P-value of the mean return metric for the previous sign prediction strategy as a function of the sample size t.** As computed in Equation 3.50, the p-value is exactly $p = 0.5625$ for $t = 2$. The bias decays with the sample length, but remains significant above 0.5% at sample length t.

ratio search technology reads

$$S_i\left(M,\,t\right)=\begin{cases}0 & \text{if } i\neq\underset{M_i\in M}{\operatorname{argmax}}\,SR_i\left(t\right)\\ 1 & \text{if } i=\underset{M_i\in M}{\operatorname{argmax}}\,SR_i\left(t\right)\end{cases}. \tag{3.51}$$

An initial window (e.g. one year) before the first portfolio return is necessary to have an initial estimate for the Sharpe ratio of each strategy. Subsequently, the Sharpe ratio of each strategy is computed in an expanding window of all past returns.

This search technology can be refined with the Reality Check method (White, 2000), as done by Hsu et al. (2016). The refined method computes a portfolio of statistically significant models, weighted by the p-value of their Sharpe ratio after adjusting for multiple testing.

3.6 Factor regression tests

The buy-and-hold strategy provides a solid model of equilibrium expected returns and is used through out the thesis as benchmark. However, a range of factors are known to generate abnormal risk adjusted returns. Therefore, when finding a strategy with abnormal returns, the obtained returns have to be tested for correlation with known anomalous factors potentially explaining the observation.

To determine if abnormal returns can be explained by known factors, the returns are regressed with the CAPM, the three-factor model of Fama and French (1993), and the four-factor model of Carhart (1997). The full four-factor model measures performance as a regression of

$$r-r^f[t]=\alpha+\beta_{MKT}\left(r_m[t]-r^f[t]\right)+\beta_{SMB}SMB_t+\beta_{HML}HML_t+\beta_{MOM}MOM_t+e_t. \tag{3.52}$$

In this regression, r is the strategy return at month t, $r^f[t]$ is the risk-free rate (the 1-month U.S. Treasury bill rate), $r_m[t]$ is the market return, SMB_t and HML_t are the size and value-growth returns of Fama and French (1993), MOM_t is the momentum return, α is the average return not explained by the benchmark model, and e_t the residual error term. The values for $r^f[t]$, $r_m[t]$, SMB_t, HML_t and MOM_t are taken from Ken French's data library (French, 2012), and derive from underlying stock returns data from the Center for Research in Security Prices (CRSP). The three-factor model is obtained by leaving out the momentum term, and the CAPM is obtained by further leaving out the SMB and HML factors.

3.7 Testing the efficient market hypothesis: an overview

The existing studies assessing the EMH using a multiple-testing framework provide a heterogeneous picture. The study by White (2000) finds no significant technical trading rule during a three year time span on the S&P 500. This result is confirmed by Sullivan et al. (1999) on a 10 year time period on the S&P 500, however they find significant trading rules on the Dow Jones for a 100 year time period ending in 1996. Subsequently, Hsu and Kuan (2005) analyze an extended universe of strategies and find no significantly performing strategy on the Dow Jones and S&P 500 during the period 1990-2000. However, they find significant trading rules for the more recent NASDAQ and Russell 2000 during the same period. Later, Hsu et al. (2010) compare the pre-ETF and post-ETF period for the U.S. and emerging markets. The post-ETF period is found to be fairly efficient, and in both periods emerging markets are found to be less efficient then the more mature U.S. markets. The efficiency check of foreign exchange markets by Hsu et al. (2016) reveals that technical trading has been profitable in the past, but returns have been declining. As for stock markets, the foreign exchanges of emerging countries are less efficient then in developed countries.

The picture is equally mixed when analyzing funds. The analysis by Fama and French (2009) shows that during the period of 1984 to 2006 mutual funds underperform in aggregate the three-factor, and four-factor benchmarks by about the transaction costs. When looking at the individual performance, even the best performers are not statistically significant. The study of Barras et al. (2005) mitigates this picture, finding significantly performing funds prior to 1996, and almost none afterwards. This result supports the stock market studies discussed above that showed increasing efficiency over time. In contrast, the portfolio approach by Yen et al. (2015) to measure fund performance does find significant performance during a 10 year period ending in 2007.

Chapter 4

Agent Based Models with Binary Strategies

4.1 Motivation & review

Econometric modeling has allowed researchers to develop a detailed understanding of the dynamics in economic and financial time series. The methods can describe the autocorrelation structure of returns, the autocorrelation of absolute returns, conditional fat tails, and many other stylized facts. Nevertheless, the forecasts from these models only work short term, until the next regime change renders a forecast obsolete. Econometric models are fundamentally unsuited to described non-equilibrium dynamics that arise in economics and finance. To move past Dynamic Stochastic General Equilibrium models (DSGEs), agent based modeling is a promising tool as it naturally allows for bounded rationality and out-of-equilibrium dynamics (Lebaron, 2006). These features are crucial to capture the complexity of the real economy, and only such a holistic approach can attempt to forecast regime changes from first principles (Farmer and Foley, 2009). The applications of agent based models to describe order book dynamics, wealth distribution among individuals, and financial markets (Chakraborti et al., 2011) have opened up the path to a promising set of tools beyond standard econometrics.

Stock markets returns exhibit a significant number of stylized facts (Cont, 2001), most prominent are the absence of linear autocorrelation, but autocorrelation of absolute returns, skewness, and fat tails. The field of econometrics had developed a large number of models (ARCH, GARCH, etc) to describe this stylized facts. In particular, the autocorrelation of absolute returns, or volatility clustering, can be well forecast using econometric methodology. However, these models do not provide a fundamental mechanism explaining why stock markets exhibit these stylized facts. This lack of an explanatory component in econometrics has motivated the construction of agent based model to reproduce these stylized facts from first principles. Surprisingly, relatively simple models with a noise trader component and fundamental trader component turned out to be sufficient to generate volatility clustering and mean reversal (Lux, 1995). Subsequent extensions of the model proposed a detailed mechanism explaining bull and bear markets (Lux and Marchesi, 1999), and could generate return distributions with significant leptokurtosis and kurtosis (Lux and Marchesi, 2000, Raberto et al., 2001, Cont, 2005, Alfi et al., 2009a,b). In these models, the periodic dominance of herding among speculative investors over the impact of external news is a crucial component to observe the stylized facts (Sornette and Zhou, 2006). A complete historical overview is provided by Samadiou et al. (2007).

Agent based models not only provide an endogenous mechanism for multiple stylized facts, but explain as well why stock markets are highly stochastic. The adaptive rational equilibrium pursued by the agents naturally gives rise to local instabilities and complicated global dynamics (Brock and Hommes, 1997). The interplay between fundamentalist and chartist traders at the micro level causes the emergence of deterministic chaos at the macro level of the stock market (Hommes, 2006). This emergent property satisfies the hypotheses of marginally efficient and rational markets when measured using econometric tests, despite the agents bounded rationality that allows the market do go through transient out-of-equilibrium phases Chen and Yeh (2002).

Despite the powerful explanatory power of agent based models for the common stylized facts, their use among practitioners is almost inexisting due to the lack of standardized calibration procedure. First attempts to quantitatively match the stylized facts were successful for simple stationary models, but inconclusive for more realistic non-stationary adaptive belief systems (Hommes, 2002). Subsequent approaches used aggregation and linearization of the agent based model to estimate the parameters using standard statistical methodology. Despite the inherent difficulties in calibrating ABMs, a number of achievements have been made in linking them to stochastic models that can be calibrated to data Feng et al. (2012). The advantage of deriving a general linear model from an ABM is that the parameters can be interpreted in terms of the influence of fundamentalist and noise traders on the market (Alfarano et al., 2005). The results support the hypothesis of strong social interaction in short-run sentiment (Lux, 2012), but the model could not beat the random walk in an out-of-sample prediction benchmark. Recent efforts leveraging the generalized method of moments achieved volatility forecasting performance equal to the seminal GARCH model, and added additional information about the trend following component (Ghonghadze and Lux, 2016).

The class of agent based models using the adaptive belief system of fundamentalist and chartist traders has proven difficult to understand analytically. Likely for this reason, a simpler class of agent based models was developed in parallel, deriving from the El Farol Bar problem (Arthur, 1994). In this problem, the agents only have to make a binary choice: to go or not to go to the bar. The payoff function the agents try to optimize is the minority game, namely avoiding the bar when it is overcrowded. This model was later adapted to stock markets with agents that only go long or short (Challet and Zhang, 1997). It turned out that such minority game stock markets posses phase transitions of marginal efficiency with stylized facts matching those of real stock markets (Challet et al., 2001). While such artificial stock market never reaches full efficiency, the continuous transfer of information from fundamentalists to speculators ensures marginal efficiency (Challet et al., 1999, Zhang, 1999), exactly as discussed by Grossman and Stiglitz (1980) two decades earlier. The advantage of the minority game model is that the stationary state of heterogeneous interacting agents can be described as the ground state of a disordered spin model (Challet et al., 2000), which is well understood in statistical physics.

The rationally optimizing agents in minority games were shown to underperform the average return of all agents (Satinover and Sornette, 2009), and in particular underperform the non-optimizing agents (Satinover and Sornette, 2007). Nonetheless, in such games, better informed agents with longer memory can always beat simpler agents with shorter memory (Challet and Zhang, 1998). Besides this undesirable predictability, the pure minority game stock markets lack as well bubbles matching those observed on real stock markets. The crucial ingredient to generate bubbles has already been well understood in models with fundamentalist and chartist traders, where large amounts of noise traders can drive the price far above or below the fundamental level as a consequence of overreaction to the action of a smaller group of fundamental traders (Bak et al., 1997). Extensions of this model adding a risk free asset are known to generate common stylized facts and super-exponential bubbles (Kaizoji et al., 2015). Indeed, generalizing the minority game by adding trend following agents does generate trends and bubbles (De Martino et al., 2004). These new agents play the majority game, always trying to predict what the majority will do at the next step. Adding a delayed majority game with one step delayed payoff function, leading to the \$-game, exhibits regime transitions from herding where the majority wins, to reversals where the minority wins. A closely related model has analyzed the role of imitation resulting from public and friend's information (Harras and Sornette, 2011), and found that a lucky strike of random events can trigger a herding phase among agents that pushes the price far above its fundamental value. Hence, simple rationally optimizing agents can drive stock markets into states of extreme susceptibility to small exogenous shocks (Patzelt and Pawelzik, 2013).

Following the qualitative success of minority, majority and mixed game models, a number of efforts have been made to calibrate these models to real stock markets using genetic algorithms. Assuming that these models capture the market dynamics emerging from the individual agents, their forecast should beat the buy-and-hold benchmark. Initial studies reverse engineering the agent strategies and using them for out-of-sample prediction on hourly Forex data achieved impressive performance (Jefferies et al., 2001). Attaching to this success, Lamper et al. (2002) showed that predictability increases prior to large changes in agent based models with agents competing for limited resources. Further pockets of predictability were identified by Andersen and Sornette (2005), arising when the agents action decouples from the past returns.

Without doubt, these models have allowed researchers to make significant progress with respect to the two major issues plaguing agent based models (Sornette, 2014): model complexity and biases in the chosen rules; and lack of standard calibration method due to the difficulty of determining the likelihood function. These models minimize complexity by using only binary histories and actions, and is bias free as agents can play any of the finite number of binary strategies. The genetic algorithms used to optimizes the agent strategies, so as to reproduce the observed time series, are well established.

Noticeable directional accuracy in daily returns of the Shanghai stock market was as well achieved by reverse engineering mixed-game models (Chen et al., 2008). The exact nature of the predictability in stock markets found with mixed agent-based is not yet fully understood. The work by Satinover and Sornette (2012a,b) showed that the individual agents playing a single game can as well achieve impressive directional accuracy. Consequently, a more detailed study is necessary to determine which fraction of the predictability can be attributed to the collective intelligence of the agents, and which is explained by the micro level optimization of the agents. In mixed game models, the predictability is typically highest during trending periods (Wiesinger et al., 2012). In this chapter, we compared the trading performance on single agent models and mixed-game models, to determine if the collective intelligence of multiple agents increases predictability.

4.2 Agent behavior

4.2.1 Strategy space

The rational agents have for only information the Z past returns $\vec{r}_t = \{r_{t-Z+1}, \ldots, r_t\}$ and want to determine the strategy f that will predict the return r_{t+1} at time $t+1$. The predicting function f, often designated as **strategy**, is a function $f : \mathbb{R}^Z \to \mathbb{R}$ mapping the vector $\vec{r}_t$ of past returns to the predicted return

$$\tilde{r}_{t+1} = f(\vec{r}_t). \tag{4.1}$$

The space $\mathbb{S}^Z := \{f | \forall f : \mathbb{R}^Z \to \mathbb{R}\}$ of all possible strategies is infinite, while the sequence of known returns is finite, which makes it impossible to constrain the function f in a meaningful way. To obtain a finite space $\mathbb{S}^Z$, the returns have to be **discretized** by a projection onto a finite set of return categories $\mathcal{R}$. The projection is defined by the choice of a function $\mu : \mathbb{R} \to \mathcal{R}$. This discretization reduces the space of all possible strategies to

$$\mathbb{S}^Z_\mu := \{f | \forall f : \mathcal{R}^Z \to \mathcal{R}\}. \tag{4.2}$$

Using $|\mathcal{R}| = n$, the cardinal of $\mathcal{R}$, the space $\mathbb{S}^Z_\mu$ of discretized strategies contains $\left|\mathbb{S}^Z_\mu\right| = n^{n^Z}$ elements. Past research has always used the simplest projection $\mu : r_t \to sign(r_t)$, which projects onto the binary set $\{-, +\}$. The small set of binary strategies is sufficient to give rise to all the stylized facts of interest, and is therefore predominantly used within this thesis. Nonetheless, in theory, one can construct strategies based on arbitrarily complex discretization, for example to create multi-scale strategies that predict based on past returns at different time scales.

4.2.2 Minority & majority payoff

To maximize profits, are rational agent always tries to correctly predict the next return. To mimic the heterogeneity of beliefs among real traders, two major types of games are considered: the minority game where agents believe that the recent trend will reverse (anti-persistency); and the majority game where agents believe the recent trend to continue (persistency). The games are modeled by a payoff function $\pi : \mathcal{R} \times \mathcal{R} \to \{-1,\, 1\}$ that associates to a true return r_t and predicted return $\tilde{r}_t$ the unitary gain or loss $\pi\left(r_t,\, \tilde{r}_t\right)$.

A payoff defines for each combination of a true and a predicted return in $\mathcal{R}^{\times 2}$ the associate payoff in $\{-1,\, 1\}$, which allows for a total of 2^{n^2} possible payoff functions. Associating the same payoff to a predicted return independently of the true return is meaningless as there would be nothing to optimize. Consequently, in the binary case with $n = 2$, there are two meaningful payoff function. One is the minority payoff function

$$\pi_{MG} : (r,\, \tilde{r}) \to 1 - 2\delta_{r,\, \tilde{r}}, \tag{4.3}$$

where one wins when being in the minority. An example is the decision to go to a bar or not; if there are few people, one is in the minority and the experience is most enjoyable. The other is the majority payoff function

$$\pi_{MAJG} : (r,\, \tilde{r}) \to 2\delta_{r,\, \tilde{r}} - 1, \tag{4.4}$$

where one wins when being in the majority. An example being stock markets during a herding phase.

4.2.3 Strategy performance

Being endowed with the knowledge of a finite set of strategies and a payoff function, the strategies can be ranked by their past performance and the best one can be selected for trading at the next time step. The performance U of a strategy f, for the payoff function π, with a delay d, on the past window of size L, is given by

$$U_{\pi,\, d,\, L}(f,\, t) = \frac{1}{L} \sum_{j=t-1-T}^{t-1} \pi\left(r_{j+d+1},\, f\left(\vec{r}_{j,\, m}\right)\right), \tag{4.5}$$

where $\vec{r}_{t,\, m} = (r_{t-m},\, \ldots,\, r_t)$. The parameter L determines how many past prediction are used for the performance computation. Beyond that time, the history is considered as obsolete for current events.

The parameter d introduces a possible delay between the time of the prediction and the time of its execution on the market. For $d = 0$ the performance measure corresponds to the standard minority (MG) and majority (MAJG) games. A one step delay, with $d = 1$, corresponds to the delayed minority (dMG) and delayed majority (dMAJG) games.

4.2.4 Memory length

The cardinal of the binary strategy space is given by $\left|\mathbb{S}_\mu^Z\right| = 2^{2^Z}$ and growths super-exponentially with the number of past returns Z:

$$\begin{aligned}
Z = 1 &\rightarrow \left|\mathbb{S}_\mu^Z\right| = 4, \\
Z = 2 &\rightarrow \left|\mathbb{S}_\mu^Z\right| = 16, \\
Z = 3 &\rightarrow \left|\mathbb{S}_\mu^Z\right| = 256, \\
Z = 4 &\rightarrow \left|\mathbb{S}_\mu^Z\right| = 65536. \\
&\vdots
\end{aligned} \tag{4.6}$$

In contrast, there are only $L+1$ unique performance values a strategy can take on a window of size L, ranging from always predicting wrong (payoff $= -1$) to always predicting correctly (payoff $= 1$) in increments of $\frac{2}{L}$. By the pigeonhole principle, it is unlikely to distinguish a single best strategy if $\left|\mathbb{S}_\mu^Z\right| \gg L+1$. Hence, the number of past returns taken as input by a strategy is limited to a **memory length** m that satisfies $\left|\mathbb{S}_\mu^m\right| = 2^{2^m} < L+1$.

4.2.5 Predictor summary

In a nutshell, a predictor $P_{\mu,\, m,\, \pi,\, d,\, L}$ (= agent) has the following components:

- A finite strategy space $\mathbb{S} \subseteq \mathbb{S}_\mu^m$ with n^{n^m} distinct strategies for the discretization map μ and memory length m.
- A payoff function π that determines the loss or gain for a true and a predicted return.
- The performance function $U_{\pi,\, d,\, L}(f,\, t)$ that computes the total payoff of a strategy $f \in \mathbb{S}$ at time t, with a delay d on the last L time steps.

At each time step t, the prediction $\tilde{r}_t$ is given by

$$\tilde{r}_t = f\left(\vec{r}_{t-d,\, m}\right), \text{ where } f = \operatorname{argmax}_{f' \in \mathbb{S}} U_{\pi,\, d,\, L}(f',\, t), \tag{4.7}$$

which is the prediction of the strategy with the highest performance on the past L time steps.

This definition of an agent is relatively bias free, as an agent can play any of the strategies, and the manageable number of the binary strategies allows us to determine the optimal strategy at each time step. Hence, these agents are a solid foundation for an ABM that avoids the two typical problems of a modeler's bias and intractable likelihood function (Sornette, 2014).

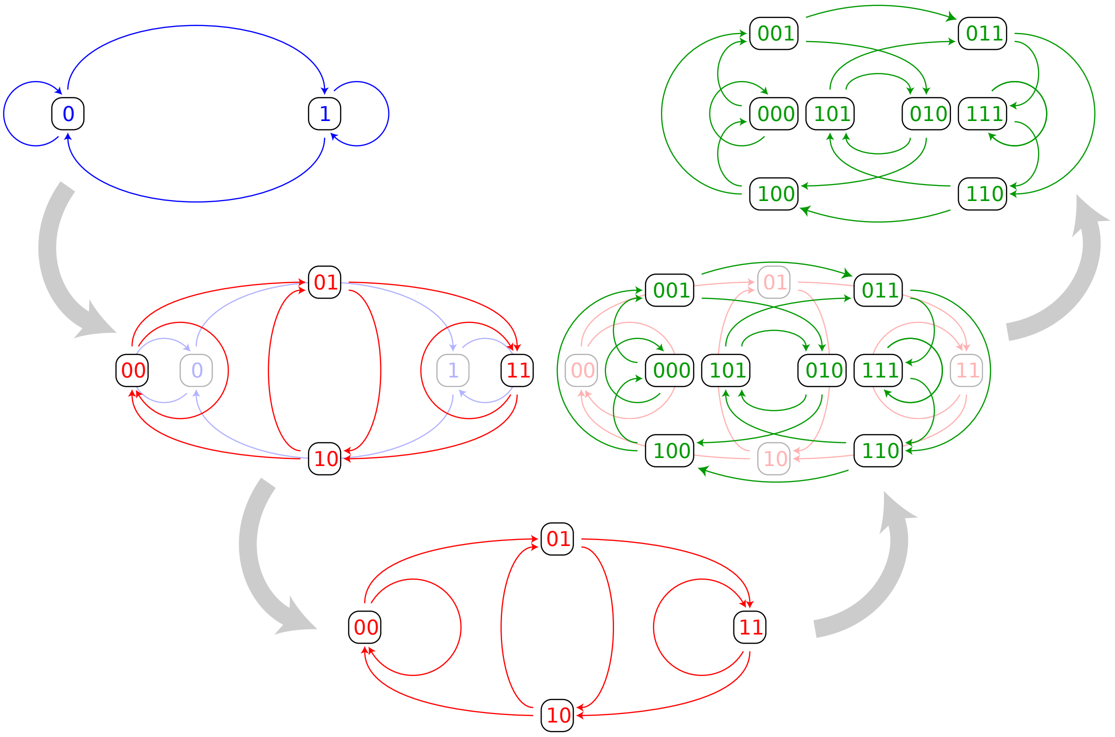


Figure 4.1: De Bruijn graphs for $\mathcal{R} = \{0, 1\}$ and $m \in \{1, 2, 3\}$.

4.2.6 De Bruijn representation

The strategy space $\mathbb{S}_{\mu}^{m} := \{f | \forall f : \mathcal{R}^m \to \mathcal{R}\}$ is best visualized as a De Bruijn graph, which is shown in Figure 4.1 for $\mathcal{R} = \{0, 1\}$ and $m \in \{1, 2, 3\}$. The input space $\mathcal{R}^m$ of the strategies defines the $|\mathcal{R}|^m$ nodes of the graph, and the directed edges are defined by the $|\mathcal{R}|^{m+1}$ possible transitions

$$(r_1, \ldots, r_m) \to (r_2, \ldots, r_{m+1}) \,. \tag{4.8}$$

An individual strategy $f \in \mathbb{S}_{\mu}^{m}$ associates to each node $\vec{r} = (r_1, \ldots, r_m) \in \mathcal{R}^m$ a unique transition defined by

$$\vec{r} = (r_1, \ldots, r_m) \to (r_2, \ldots, r_m, f(\vec{r})) \,, \tag{4.9}$$

and is therefore a sub-graph of the De Bruijn graph. A strategy $f \in \mathbb{S}_{\mu,m}^{Z}$ defines the directed graph

$$G_f = (\mathcal{R}^m, \{(\vec{r}, (r_2, \ldots, r_m, f(\vec{r}))) \,|\, \vec{r} \in \mathcal{R}^m\}) \,. \tag{4.10}$$

By construction of the map f, all nodes in this graph have out-degree one and an in-degree smaller or equal then $n = |\mathcal{R}|$. Therefore the map f^c, composing c times f , is well defined for every node. Consequently, for every input $\vec{r} \in \mathcal{R}^m$, a strategy f can be used to predict an arbitrary number of steps. Given the finite number of nodes in the graph,

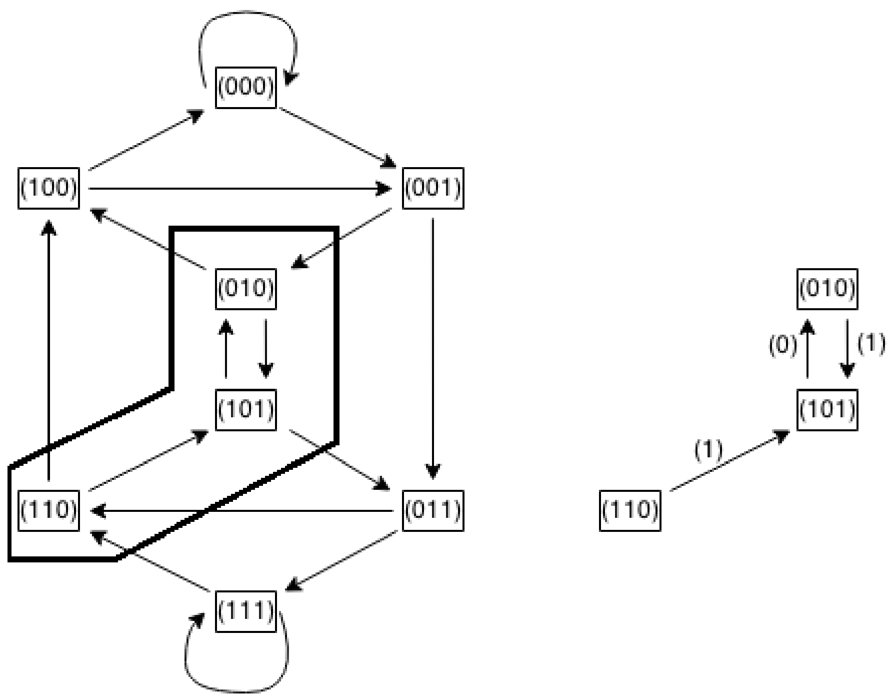


Figure 4.2: To the left the De Bruijn graph for $\mathcal{R}_2 = \{0,\ 1\}$ and $m = 3$. To the right, the sub-graph constrained by the sequence $\{1,\ 1,\ 0,\ 1,\ 0,\ 1,\ 0\}$.

there exists at least one node $\vec{r} \in \mathcal{R}^m$ and $c_r \in \mathbb{N}$, $1 \leq c_r \leq n^m$, such that $f^{c_r}(\vec{r}) = \vec{r}$. In words, this graph has at least one cycle of length at least one. All the nodes that are not in a cycle are attached to one of the cycles by some non-periodic path. It follows that for every input $\vec{r} \in \mathcal{R}^m$ the prediction made by f will necessarily become periodic after a finite number of steps, with a period smaller or equal to n^m. Because of this **periodicity**, the predictor is usually designated as a **cycle predictor**.

To build an intuition, let us consider the De Bruijn graph for $\mathcal{R}_2 = \{0,\ 1\}$ and $m = 3$, and discuss how the sequence

$$\{1,\ 1,\ 0,\ 1,\ 0,\ 1,\ 0\} \equiv \{(1,1,0),\ (1,0,1),\ (0,1,0),\ (1,0,1),\ (0,1,0)\} \tag{4.11}$$

constraints the optimal strategy. Figure 4.2 shows on the left the complete De Bruijn graph, and on the right the nodes and edges constrained by the sequence in equation 4.11. Any strategy $f \in \mathbb{S}_\mu^3$ satisfying the constraints

$$\begin{aligned} &(1,\ 1,\ 0) \to 1, \\ &(1,\ 0,\ 1) \to 0, \\ &(0,\ 1,\ 0) \to 1, \end{aligned} \tag{4.12}$$

will maximize the performance. Therefore, there are $2^{2^3-3} = 2^5 = 32$ strategies in $\mathbb{S}_\mu^3$ with maximal performance. To avoid such degeneracy of optimal strategies, the path traced by the in-sample data has to go through all nodes in the De Bruijn graph at least once.

4.2.7 Calibration optimization

To find the strategy in $\mathbb{S}_\mu^m$ with the highest performance, as defined in Equation (4.5), the brute force approach is to compute the performance of every strategy individually. However, the performance computation in Equation (4.5) can be optimized by rearranging the sums as

$$U_{\pi,\,d,\,L}(f,\,t) \quad = \quad \frac{1}{L} \sum_{\vec{r} \in \mathcal{R}^m} \sum_{j \in \mathcal{L} | \vec{r}_j = \vec{r}} \pi \left(r_{j+d+1},\, f\left(\vec{r}_j \right) \right), \tag{4.13}$$

where $\mathcal{L} = \{t - L,\, \ldots,\, t-1\}$. Now the inner sum, over a fixed $\vec{r}$, can be maximized independently from the outer sum. In the De Bruijn graph, the payoff $\pi\left(r_{j+d+1},\, f\left(\vec{r}_j\right)\right)$ is associated to the edge $\left(\vec{r}_j,\, \left(r_{j-m+2},\, \ldots,\, r_j,\, r_{j+d+1}\right)\right)$. Denoting by $p_{\vec{r},\,r}$ the probability of finding this edge in the in-sample data used for calibration, the performance function is given by

$$U_{\pi,\,d,\,L}(f,\,t) \quad = \quad \frac{1}{n^m} \sum_{\vec{r} \in \mathcal{R}^m} \sum_{r \in \mathcal{R}} p_{\vec{r},\,r} \pi \left(r,\, f\left(\vec{r} \right) \right). \tag{4.14}$$

To maximize the inner sum, the strategy is now entirely determined by

$$f\left(\vec{r}\right) = \underset{r' \in \mathcal{R}}{\operatorname{argmax}} \sum_{r \in \mathcal{R}} p_{\vec{r},\,r} \pi\left(r,\, r'\right). \tag{4.15}$$

In case several values $r_{\vec{r}} \in \mathcal{R}$ maximize the inner sum, a random choice is made. The probabilities $p_{\vec{r},\,r}$ can be computed linearly with respect to the in-sample data size L, and the optimal $r_{\vec{r}}$ can be computed linearly with respect to $n = |\mathcal{R}|$, which makes the entire computation independent from the number of strategies in the space $\mathbb{S}_\mu^m$.

Updating the best performing strategy from one time step to the next can be achieved in constant time. Removing time $t - L$ from $\mathcal{L}$ and adding time t will affect at most two of the probabilities $p_{\vec{r},\,r}$, which can be updated without recomputing any of the other probabilities. This change in the probabilities affects at most two of the inner sums in Equation (4.14), and therefore requires to update at most two of the values of $f\left(\vec{r}\right)$.

4.2.8 Relations between games

The minority game (MG) corresponds to the belief that markets are anti-persistent, and the majority game (MAJG) correspond to the belief that markets are persistent. This diametrical opposition of the two beliefs is mathematically expressed by the payoff function relation $\pi_{MG}\left(r,\, \tilde{r}\right) = -\pi_{MAJG}\left(r,\, \tilde{r}\right)$. Hence, in the case of binary cycle predictors ($\mathcal{R} = \{-,\, +\}$), it follows from Equation (4.15) that the one step ahead forecast of the minority game is exactly opposite to the majority game forecast. When predicting one step ahead, the minority and majority strategies are perfectly anti-correlated, and their mean return is exactly opposite.

For fixed calibration data, the graph of the best minority strategy and the best majority strategy, as defined in equation 4.10, share no edges. Consequently, when predicting several steps ahead, the minority and majority predictors are no longer bound to make opposite predictions from the second step onward, which leads to decreasing anti-correlation as the number of predicted steps increases. When forecasting several steps ahead, the mean return of the minority strategy is no longer given as the opposite of the mean return of the majority strategy.

The anti-correlation between the minority and majority game, when predicting one step ahead, no longer holds for predictors with more then two classes ($|\mathcal{R}| > 2$). Given an odd number of states, the concept of opposite is not clearly defined anymore. A meaningful definition for $\mathcal{R}_3 = \{-, 0, +\}$ could be to define minus and plus as opposite to each other and zero as its own opposite. However, when computing the highest performing strategy according to Equation (4.15), there is no guarantee that the minority and majority prediction will be opposite to each other according to this definition. Therefore, in the three state case, the minority and majority predictors are not necessarily correlated, even when predicting one step ahead.

The delayed games incorporate the possible belief that markets are anti-persistent (dMG) or persistent (dMAJG) in a delayed manner. However, it has to be noted that any d steps delayed strategy of memory m can be mapped to a none delayed strategy of memory $m + d$. The none delayed strategy makes the same prediction irrespective of the last d values in the input. In other terms, the strategy space $\mathbb{S}_{\mu}^{m+d}$ is a super-set of $\mathbb{S}_{\mu}^{m}$, and therefore the performance of the delayed predictors should be explained to a large extend by predictors with longer memory.

4.2.9 Viable cycle predictors

As discussed in Subsection 4.2.6, the best strategy is uniquely defined if and only if the in-sample data goes at least once through every node of the De Bruijn graph of the strategy space. Therefore, the parameters that determine the calibration process are the number of classes $n = |\mathcal{R}|$ of the discretization, the memory length m of the strategies, and the window size L used to compute the strategy performance.

In the context of predicting daily time series, a decade of data corresponds to ≈ 2500 data points. As a long out-of-sample prediction period is needed to obtain statistically significant results, the calibration window L should not take away more then 20% of the data, corresponding to the bound $L \leq 500$. For the in-sample data to eventually cover all nodes in the De Bruijn graph, the constraint $n^m \leq L$ is obtained by the pigeonhole principle. However, in practice, the in-sample data is unlikely to cover uniformly all nodes, and a correct bound should be chosen empirically. In the most optimistic case, the possible combinations are given by

$$\left(n \in \{1, \ldots, L\}, m \in \left\{1, \ldots, \left\lfloor \frac{\log(L)}{\log(n)} \right\rfloor \right\}\right).$$

The discretization $\mu(r) = \operatorname{sign}(r)$ is the simplest possible choice in the binary case. However, the most general discretization, with topologically connected inverse image for each class, introduces $n-1$ new parameters. The general discrete map to $\mathcal{R}_2 = \{-, +\}$ reads

$$\mu_2(r) = \begin{cases} - & \text{if } r \leq r_0 \\ + & \text{if } r > r_0 \end{cases},$$

where $r_0 \in \mathbb{R}$ defines the split threshold between up and down moves. An evident inconvenience of binary cycle predictors is that they treat small and large moves as equal, which does not filter out the noise from small returns. Hence, the trinary cycle predictors are an interesting option as they circumvent this issue by mapping small returns to a separate class. The general discrete map to $\mathcal{R}_3 = \{-, 0, +\}$ reads

$$\mu_3(r) = \begin{cases} - & \text{if } r \leq r_- \\ 0 & \text{if } r_- < r \leq r_+ \\ + & \text{if } r > r_+ \end{cases},$$

where $r_-, r_+ \in \mathbb{R}$ define thresholds between down moves and zero moves, respectively zeros moves and up moves.

4.3 Mixed game agent based model

4.3.1 Agent heterogeneity

The single agent prediction experiments showed that the minority and majority game both had significant predictive power when using the correct memory length m and calibration length L. Therefore, it is a logical next step to determine if the collective intelligence of a multi-agent model can improve upon the forecasting power of the single agent model. Such a model follows a bottom-up approach in which interactions between the individual agents generate improved forecast at the aggregate macro-level.

The complete Markov representation of multi-agent models with a single game (MG, MAJG), single memory length m, and single calibration length L have already been derived by Satinover and Sornette (2012a). However, these models are too rigid to combine the different pockets of predictability found with the single agent model. Excess predictability was found in different games, and a more promising multi-agent model should mix the different games. Hence, a mixed game model with N agents is constructed, where agent $1 \leq i \leq N$ trades according to his private predictor $P^i_{\mu, m, \pi_i, d_i, L}$. Each agent is explicitly assigned an individual payoff function and delay $(\pi_i, d_i) \in \{\pi_{MG}, \pi_{MAJG}\} \times \{0, 1\}$, which gives rise to four groups of agents: minority game; delayed minority game; majority game; and delayed majority game. The assignments are made so as to have the same number of agents for each of the four games. This models well the heterogeneity of believes among traders about the current market regime and the time frame (delay) in which regime changes occur.

In the scenario where each agent is endowed with the full knowledge of all strategies ($\mathbb{S}^i = \mathbb{S}^m_\mu$), the agents playing the same game would all select the same optimal strategy. Consequently, the model would be composed of only four effective agents, one for each of the four games. To introduce further heterogeneity of the agents, they could be assigned different memory lengths m, calibration window L, or strategy set $\mathbb{S}^i$. In nowadays trading environment, all the agents have similar access to stock price histories and computing power. Therefore, the memory length m and calibration window L will be shared by all the agents. The additional heterogeneity between agents playing the same game is introduced via their private strategy set. Each agent is assigned a subset $\mathbb{S}^m_i$ of all possible strategies, of size $|\mathbb{S}^m_i| = n_s < n^{n^m}$, which represents his personal belief about which strategies will have the best performance. The calibration algorithm defined in Subsection 4.2.7 can be adapted to determine the optimal strategy in the reduced set of strategies $\mathbb{S}^m_i$ of agent i.

4.3.2 Agent trading threshold

So far, an agent would always buy or sell according to his private predictor, however in real markets an agent can decide not to trade if none of his strategies has a convincing performance. This concept was introduced by Jefferies et al. (2001) to help model the liquidity in the real stock markets. The liquidity is controlled by the trading threshold τ, which is the minimal performance the best strategy needs to exceed for being used. Below this threshold an agent will not trade.

The performance function defined in Equation (4.5) is bounded by $|U| \leq |\pi| = 1$. The performance of random strategies converges to zero at large L for the minority and majority games, because the payoff function of these games satisfies $\sum_{r \in \mathcal{R}} \pi\left(r,\, \tilde{r}\right) = 0$, $\forall \tilde{r}$. Hence, the performance is a measure of the excess payoff compared to random predictions, and the trading threshold τ has to be chosen in [0, 1]. The action $a^i_{t+1}(\vec{r}_t)$ taken by the agent i at time step $t+1$ is given by

$$a^i_{t+1}\left(\vec{r}_t\right) = \begin{cases} f^i_t\left(\vec{r}_t\right) & \text{if} \quad U_{\pi_i,\, d_i,\, L}(f^i_t,\, t) \geq \tau \\ 0 & \text{if} \quad U_{\pi_i,\, d_i,\, L}(f^i_t,\, t) < \tau \end{cases}, \tag{4.16}$$

where f^i_t is the highest performing strategy of agent i at time step t.

Due to the trading threshold, the relative fraction of active agents in each game (MG, dMG, MAJG, dMAJG) can vary over time within a simulation. This dynamic aligns with the evidence that financial markets are characterized by regime shifts, with changing types of investment styles, driven by the monetary policies of central banks and macroeconomic conditions. In particular Lux and Marchesi (1999) stress the importance of including time varying fractions of investment styles to account for the stylized facts of financial returns.

4.3.3 Aggregate prediction

The behavior of the individual agents has been entirely determined, but the aggregation mechanism of the individual actions still needs to be defined. The aggregation mechanism of stock markets is the topic of an ongoing debate, with proponents of a linear aggregation function (Kyle, 1985, Farmer, 2002), and proponents of nonlinear aggregation functions (Lillo et al., 2003, Almgren, 2003, Bouchaud et al., 2006, Farmer et al., 2012). For the purpose of this agent based model, a simple linear impact function is used to determine the return predicted at the next time step as

$$\tilde{r}_{t+1} = \frac{1}{\lambda} \sum_{i=1}^{N} a_t^i \left(\vec{r}_t\right), \tag{4.17}$$

where λ is a normalization factor called liquidity. If there are more buyers than sellers, the price will go up and vice versa. The prediction $\tilde{r}_{t+1}$ takes a finite number of N values in the set $\{-N/\lambda,\, (-N+2)/\lambda,\, \ldots,\, N/\lambda\}$, which is more fine grained then predictions of individual agents.

4.3.4 Mixed-game model summary

The mixed game multi-agent model is determined by the following parameters:

- The agents $\{1,\, \ldots,\, i,\, \ldots,\, N\}$ trading according to the prediction of their individual cycle predictor $P^i_{\mu,\, m,\, n_s,\, \pi_i,\, d_i,\, L}$ over a reduced set of strategies $\mathbb{S}_i^m$. The payoff functions π_i and delays d_i are chosen so as to have an equal number of agents in each of the four games (MG, dMG, MAJG, dMAJG).
- The number of strategies n_s in the strategy space $\mathbb{S}_i \subseteq \mathbb{S}_\mu^m$ of each agent.
- The performance threshold τ the best strategy needs to exceed for being played.

The strategies used by the agents are taken within the set of all possible binary strategies, hence this agent based model does not suffer any particular bias introduced by the modeler. However, some features are overly simplistic, for example the lack of capital constraints of the agents, which trade the same irrespective of their gains or losses.

4.3.5 Calibration

For a given return history, the prediction of each individual agent is deterministic, hence the aggregate prediction of the agents is deterministic as well. However, the heterogeneity of the strategies tracked by the individual agents generates aggregate dynamics that cannot be forecast by a single agent. Due to the heterogeneity of the strategy sets, the individual agents switch their best strategy at different times, making it impossible for a single

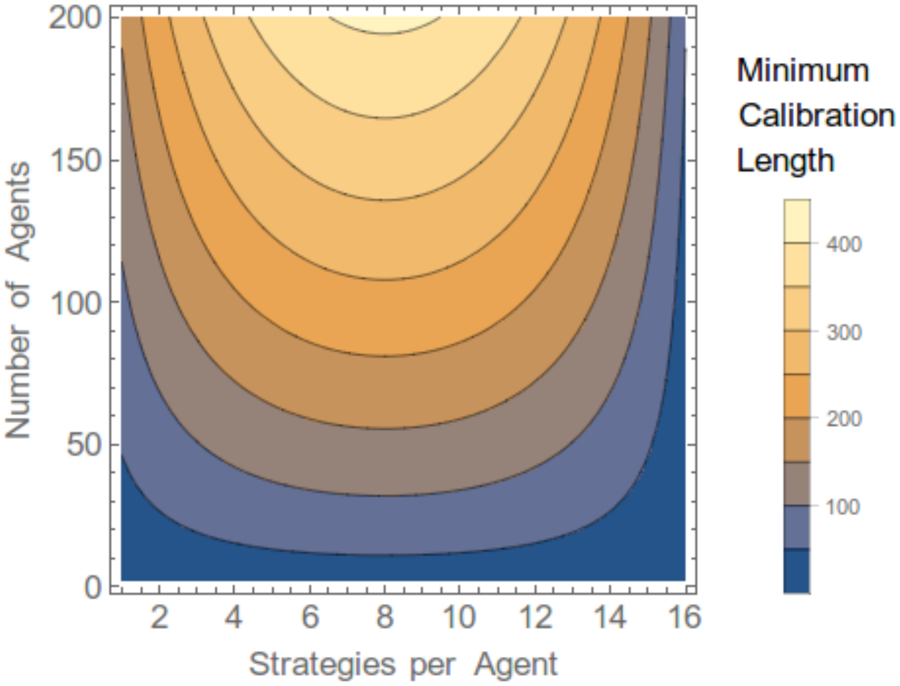


Figure 4.3: Minimum calibration length required for the ABM computed using Equation (4.20) in the case of binary strategy with memory length $m = 2$.

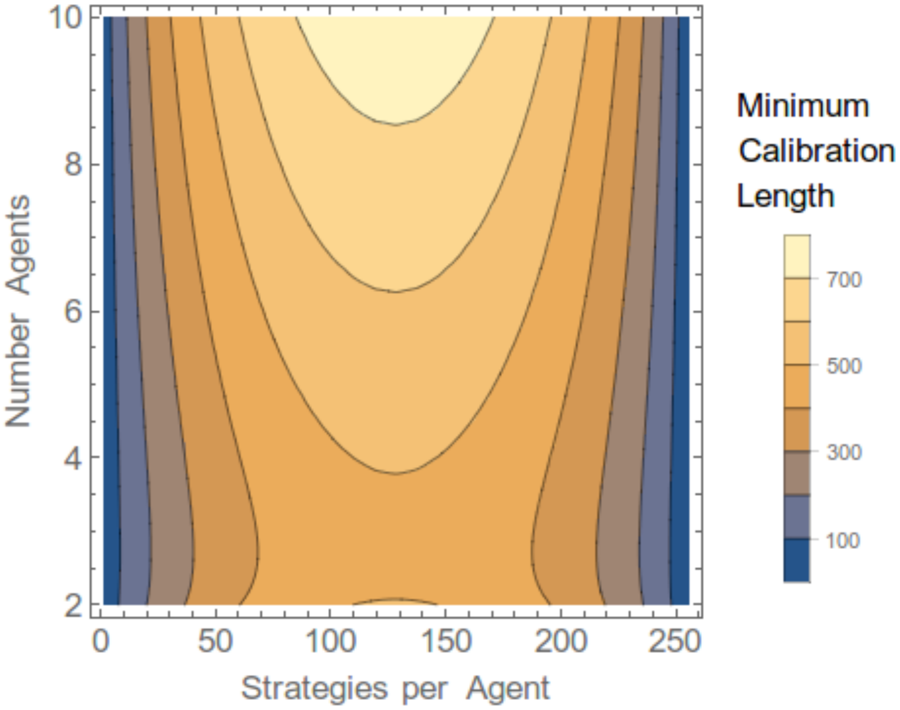


Figure 4.4: Minimum calibration length required for the ABM computed using Equation (4.20) in the case of binary strategy with memory length $m = 3$.

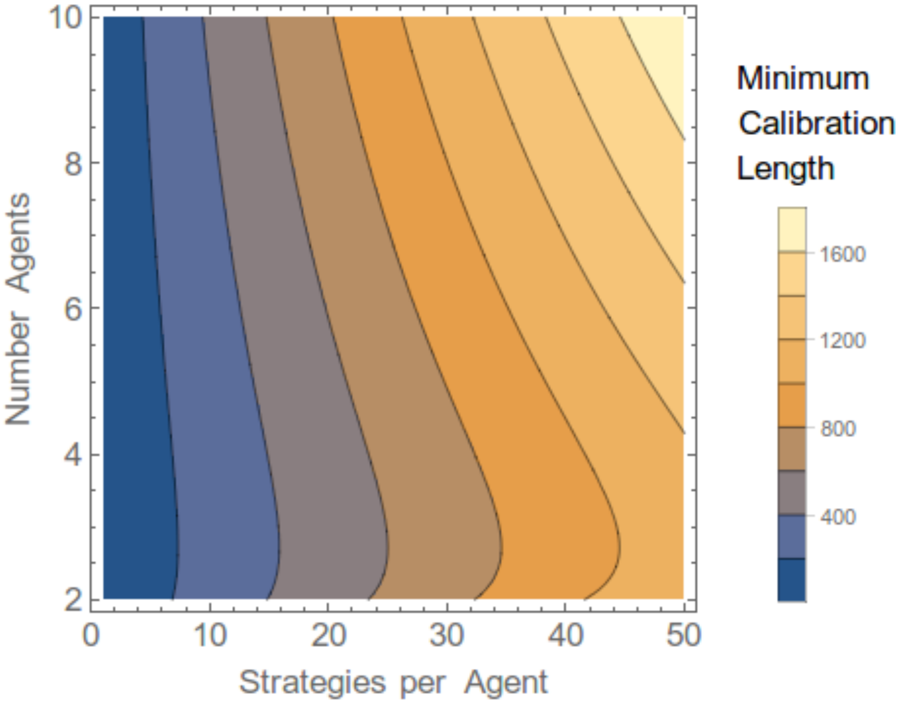


Figure 4.5: Minimum calibration length required for the ABM computed using Equation (4.20) in the case of binary strategy with memory length $m = 4$.

agent to keep track of the complexity of the multi-agent dynamic. While the forecast of an individual agent is entirely determined by the past L returns, the multi-agent model incorporates a states depending on the returns before the performance window of size L.

The parameters of the agent based model should be chosen so that the number of possible states of the model is roughly equal to the number of possible states of the in-sample calibration data. Otherwise, by the pigeonhole principle, there are sequences that are generated by more then one ABM configuration and the calibration becomes ill-defined. The space of all possible strategy configurations $\mathbb{S}^m_{ABM}$ of the ABM is directly related to the number of possible predictors for an individual agent. An individual agent consists of a set $\mathbb{S}^m_i$ of n_s distinct strategies sampled in $\mathbb{S}^m_\mu$, a payoff function $\pi_i \in \{\pi_{MG}, \pi_{MAJG}\}$, and delay $d \in \{0, 1\}$. This leads to an ABM space containing

$$|\mathbb{S}^m_{ABM}| = \binom{\left|\mathbb{S}^m_\mu\right|}{n_s}^N 2^N 2^N = \left(\frac{4 \cdot n^{n^m}!}{(n^{n^m} - n_s)!n_s!}\right)^N \tag{4.18}$$

different strategy configurations. The aggregation function chosen in Subsection 4.3.3 makes N distinct forecasts, equal to the number of agents, and therefore the ABM can generate

$$\#\text{distinct forecast} = N^L \tag{4.19}$$

distinct forecast for the in-sample period of length L. The number of ABM strategies is equal to the number of distinct in-sample sequences the ABM can generate when

$$L = \frac{N}{\log(N)} \cdot \log\left(\frac{4 \cdot n^{n^m}!}{(n^{n^m} - n_s)! n_s!}\right). \tag{4.20}$$

The minimal calibration length L in the case of binary strategies is shown as a function of N and n_s in Figure 4.3 for $m = 2$, in Figure 4.4 for $m = 3$, and in Figure 4.5 for $m = 4$. The results show that an ABM with memory length $m = 2$ can be calibrated for several hundreds of agents and any number of strategies per agents. At memory length $m = 3$, the maximal number of agents is limited to a few dozens, and the number of strategies per agents should be chosen sufficiently small or large depending on the number of agents. At memory length $m = 4$, only a small number of agents with a few dozens of strategies are meaningful to calibrate.

The analysis of the number of configurations of the mixed multi-agent model showed that a viable parameter region exists, in which the likelihood function for any calibration data should have a unique optimum. The calibration procedure to find the minimum has to minimize the mean squared error

$$MSE = \sum_{i=1}^{L} \left(\tilde{r}_i^{ABM} - r_i\right)^2 \tag{4.21}$$

between the predicted and true returns of the in-sample calibration data. The minimization is typically performed over the following parameters: the number of agents $N \in \{3, \ldots, 100\}$; the memory length $m \in \{1, 2, 3, 4\}$; the number of strategies per agent $n_s \in \{1, \ldots, 16\}$; the threshold for action $\tau \in [0, 1]$; the duration of the scoring counter $L \in \{1, \ldots, 500\}$; and the strategies of each agent. The reverse-engineering is performed using a genetic algorithm as described in Wiesinger et al. (2012).

The reverse engineering procedure finds the multi-agent configuration that explains best the observed in-sample data. As for the single agent model, the ABM is repeatedly calibrated using a rolling window of size L to predict one or several out-of-sample steps. The out-of-sample prediction of the multi-agent model should possess a collective intelligence outperforming the single agent model.

4.4 Results

We test the predictive performance of the single agent models and the multi-agent model on the equity indices and time periods presented in table 4.1. The two chosen time periods are ten years long and each contain a major bubble and its crash. The first time period, from Jan. 1992 to Dec. 2001, covers the build up to the dotcom bubble and most of the subsequent crash. The second time period, from Jan. 2002 to Dec. 2011, covers the build up to the financial crisis, the crash, and subsequent recovery. Given this choice of time periods, any strategy that can somehow time the crash of the bubble will significantly outperform the buy-and-hold strategy.

Ticker	Start	End	σ_r (%)	$\langle\lvert\rho\rvert\rangle$(%)	$\#_{\uparrow}$(%)	ρ_m(Bps)
^GSPC	2002	2011	1.39	0.91	53.9	0.98
^GSPC	1992	2001	0.99	0.70	53.1	4.66
^DJI	2002	2011	1.29	0.86	52.7	1.32
^DJI	1992	2001	0.98	0.70	53.4	5.07
^NDX	2002	2011	1.68	1.18	53.7	2.36
^NDX	1992	2001	2.10	1.48	54.1	8.66

Table 4.1: Equity indices and time periods used to test the single agent models and the mixed game ABM. The table provides: the daily volatility σ_r; the mean absolute return $\langle|\rho|\rangle$; the percentage of up moves ($\#_{\uparrow}$); and the mean daily market return ρ_m.

To search in a systematic manner for anomalous models on each equity index and time period, we test the single agent models and the multi-agent model for the memory lengths

$$m \in \{1, 2, \ldots, 4\}$$

and scoring counters

$$L \in \{20,\ 40,\ \ldots,\ 500\}.$$

This yields a total of $4 \times 25 = 100$ different models, tested on each of the six combinations of a equity index and time period.

The maximal Sharpe ratio difference with respect to the buy-and-hold strategy (i.e. lowest p-value)

$$\Delta\hat{\mathrm{Sr}}_{\max} = \max_{m,\, L} \hat{\mathrm{Sr}}_{\mathrm{model}(m,\ L)} - \hat{\mathrm{Sr}}_{\text{buy-and-hold}}$$

is shown in Table 4.2 for each combination of an asset and model. The distribution of the Sharpe ratio differences is shown in Figure 4.6 for the single agent majority models without delay, in Figure 4.7 for the single agent majority model with a one day delay, and in Figure 4.8 for the mixed game agent based model.

Five, out of the six experiments, with single agent majority models, without delay, contained instances with significant trading performance at the 95% confidence level. Four, out of the six experiments, with one day delayed single agent majority models contained instances with significant trading performance at the 95% confidence level. None of the mixed game agent based models had statistically significant trading performance, which even under-performed the single agent models in all cases. The reverse engineering of the complex multi-agent model was unable to emulate the abnormal performance (before transaction costs) that can be achieved with trend following, as performed by the single agent models. The genetic algorithms have not been successful at extracting potential emergent phenomena in the stock markets.

To analyze the robustness of the ABM, we compared multiple runs predicting the same time series. It turned out that the predictions of the individual runs had little correlation

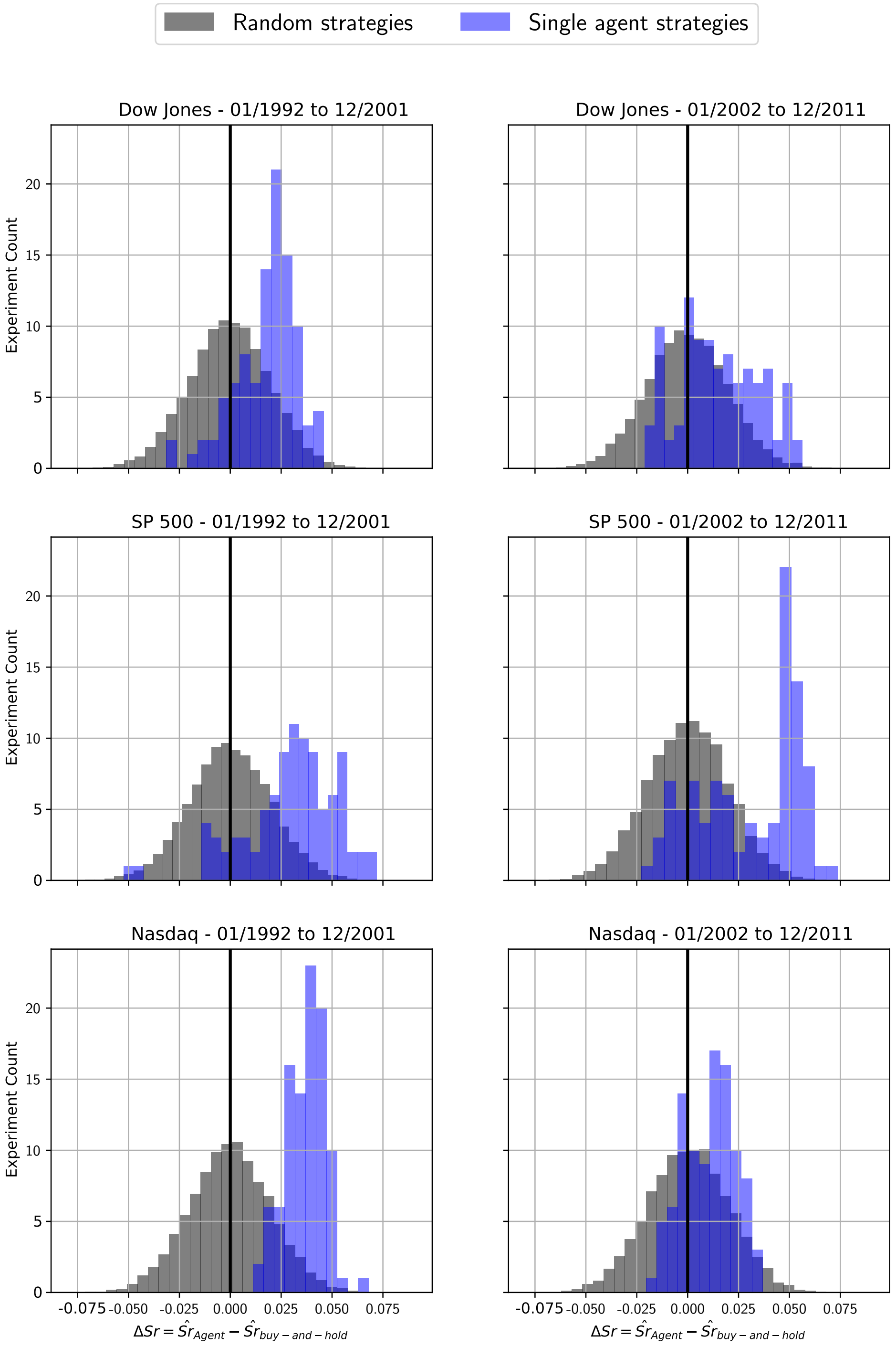


Figure 4.6: Sharpe ratio difference of the single agent models with respect to the buy-and-hold strategy. The single agent models are tested for the majority game (**without delay**), $m \in \{1, 2, 3, 4\}$, and $L \in \{20, 40, \ldots, 500\}$, yielding a total of 100 models. As reference, the distribution of 100 random strategies is shown, strategies that go randomly short or long on each day.

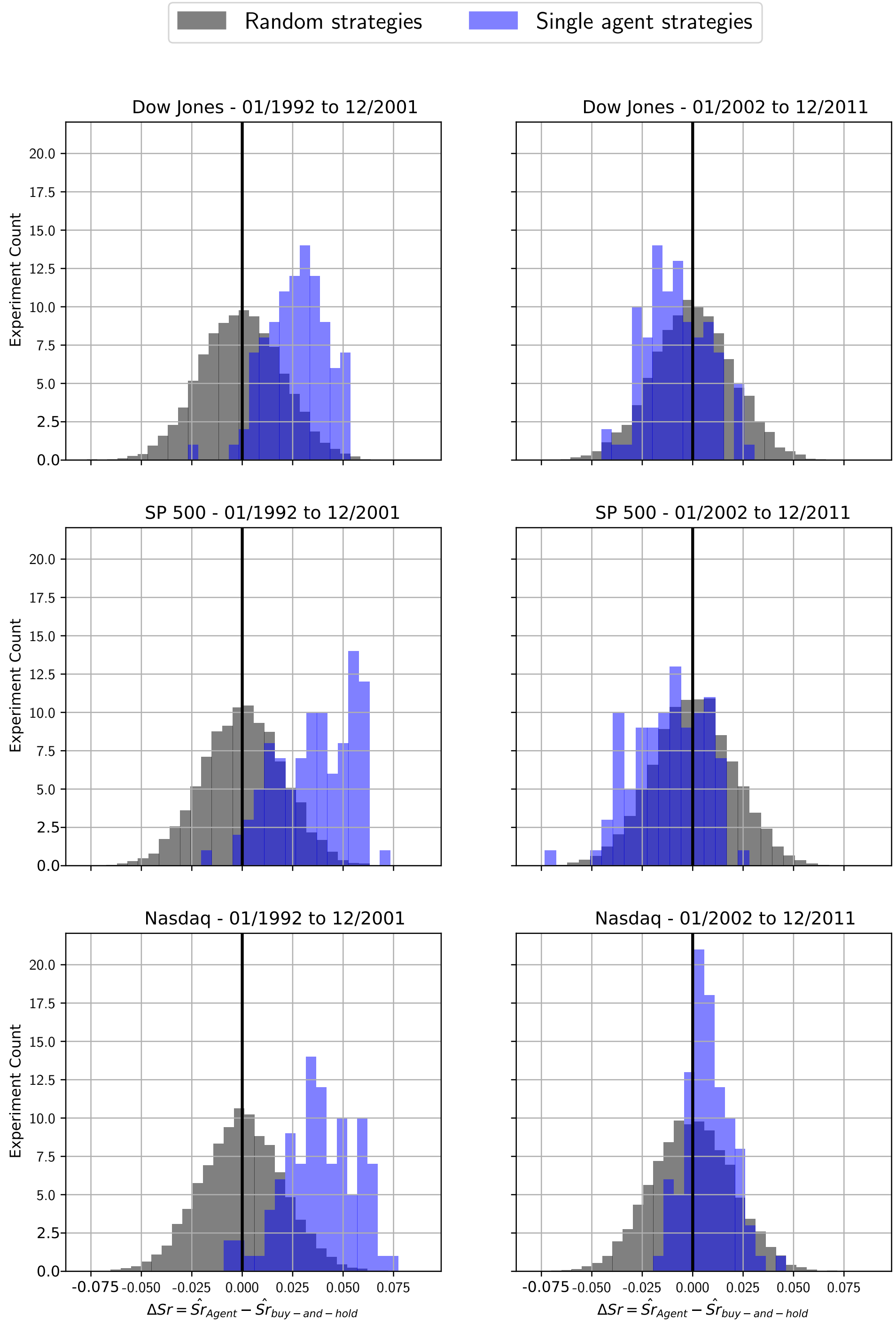


Figure 4.7: Sharpe ratio difference of the single agent models with respect to the buy-and-hold strategy. The single agent models are tested for the majority game (**with delay** $d = 1$), $m \in \{1, 2, 3, 4\}$, and $L \in \{20, 40, \ldots, 500\}$, yielding a total of 100 models. As reference, the distribution of 100 random strategies is shown, strategies that go randomly short or long on each day.

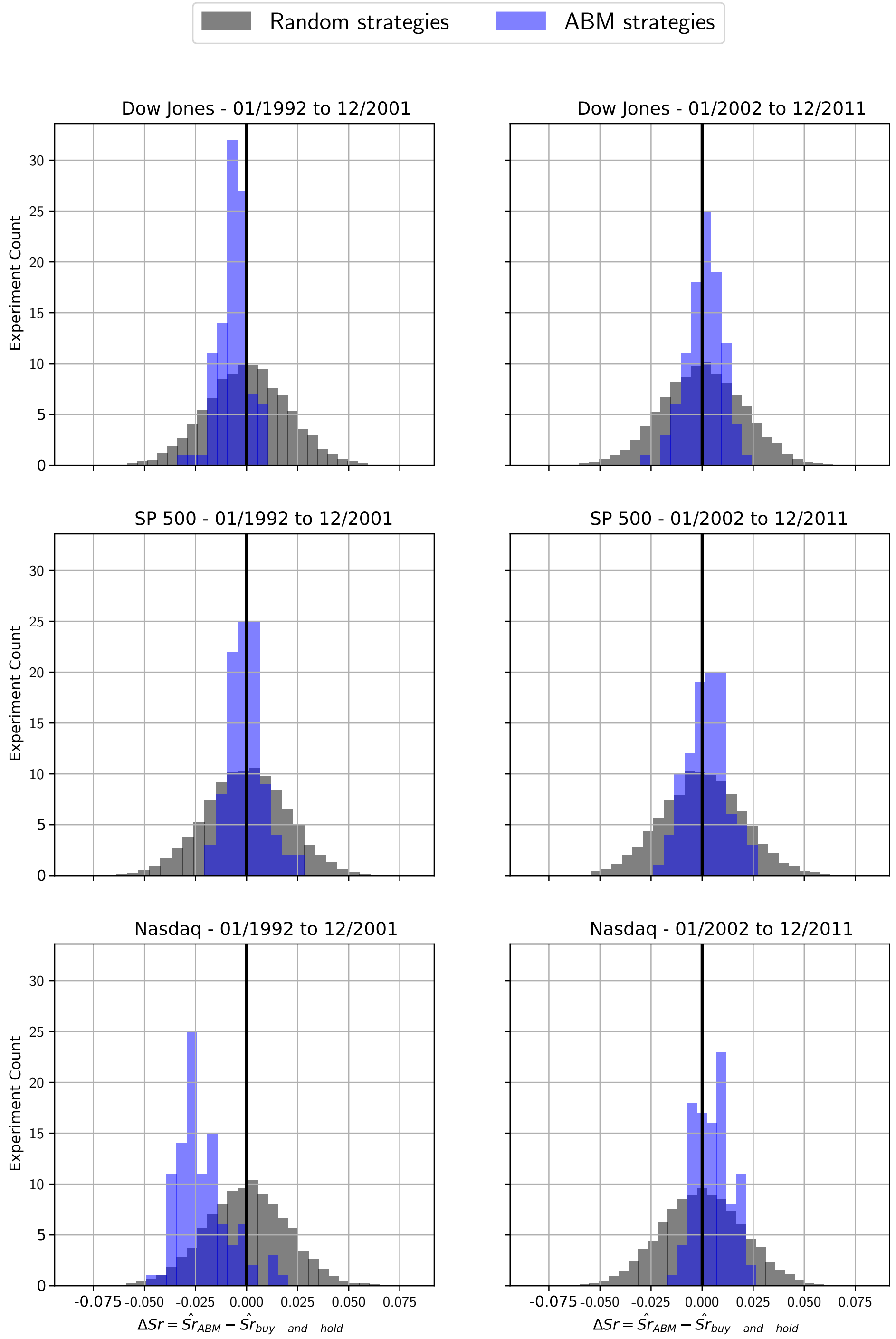


Figure 4.8: Sharpe ratio difference of the mixed game multi-agent model with respect to the buy-and-hold strategy. The ABM is tested for fixed $m \in \{1, 2, 3, 4\}$, and $L \in \{20, 40, \ldots, 500\}$, yielding a total of 100 models. As reference, the distribution of 100 random strategies is shown, strategies that go randomly short or long on each day.

Ticker	Start	End	Model	Game	$\Delta\widehat{\mathrm{Sr}}_{\mathrm{max}}$	Min p-value
^GSPC	2002	2011	1-Agent	MAJ	0.073	0.00*
			1-Agent	dMAJ	0.025	0.34
			ABM	Mixed	0.024	0.21
^GSPC	1992	2001	1-Agent	MAJ	0.071	0.00*
			1-Agent	dMAJ	0.072	0.00*
			ABM	Mixed	0.026	0.19
^DJI	2002	2011	1-Agent	MAJ	0.065	0.00*
			1-Agent	dMAJ	0.027	0.28
			ABM	Mixed	0.023	0.26
^DJI	1992	2001	1-Agent	MAJ	0.047	0.05*
			1-Agent	dMAJ	0.052	0.02*
			ABM	Mixed	0.009	0.46
^NDX	2002	2011	1-Agent	MAJ	0.032	0.14
			1-Agent	dMAJ	0.072	0.00*
			ABM	Mixed	0.024	0.19
^NDX	1992	2001	1-Agent	MAJ	0.068	0.00*
			1-Agent	dMAJ	0.075	0.00*
			ABM	Mixed	0.021	0.31

Table 4.2: Highest Sharpe ratio excess with respect to the buy-and-hold strategy, and bootstrapped p-value, for each of the single agent and ABM experiments shown in Figure 4.6, Figure 4.7, and Figure 4.8. The experiments with significant trading performance are marked with an asterisk (*).

among each other, but all achieved very high in-sample fitness. A typical fitness being a directional accuracy of more then 90% of the in-sample days. This shows that the likelihood function of the ABM has many local maximum, and that the reverse engineering algorithm merely selects randomly a maximum depending on the starting values of the optimization.

A correlation analysis between the ABM and the equivalent single agent model shows identical prediction on 54% of days, averaged across all experiments. This implies that the agents in the ABM are heterogeneous in their beliefs, and none of the four agent groups (MAJ, dMAJ, MG, dMG) dominates the aggregate prediction. Hence, the ABM reverse engineering does find complex agent configurations that reproduce well the in-sample returns. Nonetheless, the aggregate prediction did not outperform the single agent model, and therefore the multi-agent configurations in the ABM cannot be interpreted as capturing some true hidden state of the agents acting on the stock market.

4.5 Conclusion

In this chapter, we presented a mathematical definition of an individual agent and the mixed game multi-agent model. Single game multi-agent models have already been studied by Challet et al. (2001), Satinover and Sornette (2012a,b), and the positive results provide incentive to examine the predictive power of mixed game models (Wiesinger et al., 2012, Zhang, 2013).

The mixed game ABM is characterized by the following parameters: the number of agents, the memory length of the agents, the number of strategies tracked by each agents, the trading threshold of the agents, the scoring duration for the strategies, the strategies tracked by each agent, and the game played by each agent. An exhaustive evaluation of all possible parameter combinations is computationally impossible. Nevertheless, to gain some systematic inside, we evaluated the ABM for all combinations of a statistically meaningful memory length and scoring counter.

The ABMs and matching single agent models were then tested, for all memory length and scoring counter combinations, on two distinct time periods for the S&P500, NASDAQ and Dow Jones. The comparison between the single agent models and the ABM revealed that the complexity added by the ABM could not improve the predictive performance. Actually, the opposite is true, the single agent models significantly outperformed the mixed game ABM. Hence, the approach of using genetic algorithms to reverse engineer large agent based models is not practicable. The analysis showed that the reverse engineering overfitted the in-sample data, without being able to extract the true hidden state of agents acting on the stock market.

Given the observed abnormal Sharpe ratio of single agent models, the next chapter is focused on better understanding single agent models. The detailed study of abnormal performance in single agent model can potentially provide insights on how to combine multiple agents, so as to truly extract emergent predictability.

Chapter 5

From Agent Based Modeling to Time Series Learning

5.1 Agent as decision tree

The agents defined in Subsection 4.2 track an individual set of strategies, and at every time step trade based on the strategy with best past performance for their payoff function. The space of possible strategies is determined by the chosen discretization $\mathcal{R}$ of the returns and the memory length m. An individual strategy is a function $f : \mathcal{R}^m \to \mathcal{R}$ that assigns to each of the $|\mathcal{R}|^m$ histories a prediction in $\mathcal{R}$. For each return in the history of length m, the class in $\mathcal{R}$ of that return can be determined by a sequence of inequalities. Hence, a strategy can be visualized as a decision tree, which is a common statistical learning method (James et al., 2014). A randomly chosen strategy for $\mathcal{R} = \{-, +\}$ and $m = 3$ is represented as a decision tree over the past three returns $\{r_1, r_2, r_3\}$ in Figure 5.1. The extension to a larger number of classes defined by cuts at arbitrary thresholds, and arbitrary memory length, is straightforward by introducing the equivalent nodes. Each branch corresponds to a single history, and the action following that history is assigned to the terminal node of the branch.

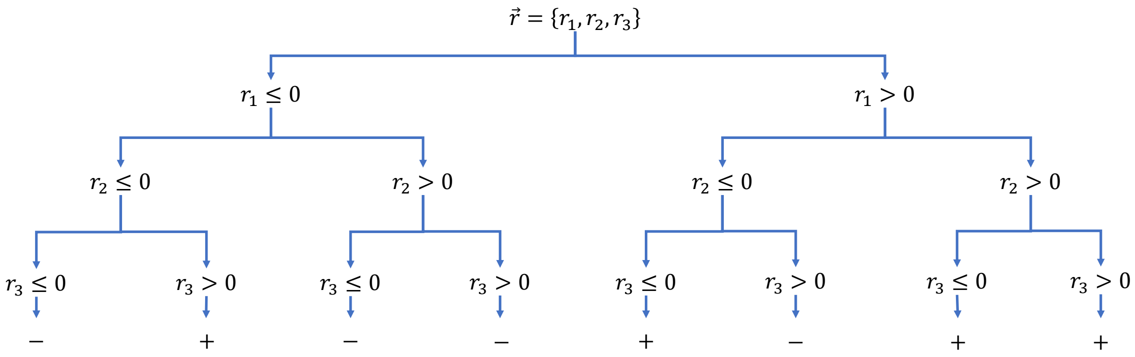


Figure 5.1: **Binary strategy with memory length $m = 3$ represented as a decision tree.** Each of the $2^3 = 8$ histories is assigned to a long or short trade.

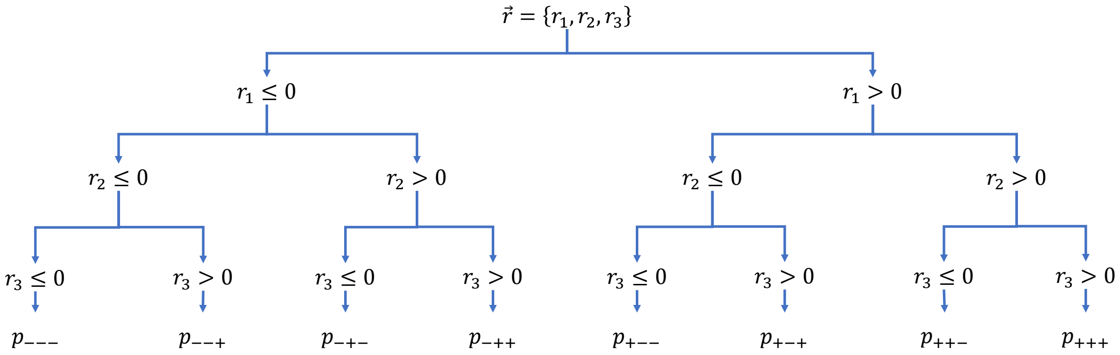


Figure 5.2: **Up move probabilities for a binary decision tree with memory length $m = 3$.** The probabilities in the terminal nodes indicate the probability of an up move. For example, $p_{---} = P(---\to +)$.

Given an in-sample return sequence $\{r_1, \ldots, r_L\}$ of length L, and a decision tree with fixed branches, the classification and regression tree algorithm (Hastie et al., 2001) computes the class probabilities in each branch. For example, in the case of binary classes, the probability of three down moves being followed by an up move is given by $p_{---} = P(---\to +)$. This probability is computed by finding all instances of three down moves in a given in-sample sequence and taking the ratio of the number of instances followed by an up move over the total number of instances. A generic decision tree for binary classes and memory length $m = 3$ is represented in Figure 5.2. The normalization of the probabilities imposes $P(h \to -) = 1 - P(h \to +)$ for all histories $h \in \{-, +\}^m$. Hence, the generic decision tree with binary classes is determined by a total of 2^m probabilities, one for each branch. In general, the class probabilities are constrained by

$$\sum_{a \in \mathcal{R}} P(h \to a) = 1, \, \forall h \in \mathcal{R}^m, \tag{5.1}$$

leading to a total of $|\mathcal{R}|^{m+1} - |\mathcal{R}|^m$ independent probabilities defining the decision tree.

The decision tree predicts for each history (i.e. branch) the class with the highest probability. This prediction mechanism is equivalent to selecting the strategy with the highest performance for the majority game as defined in Subsection 4.2.7. The probabilities $p_{\vec{r}, r}$ used in Equation (4.15) to determine the strategy with the highest performance are the class probabilities of the equivalent decision tree. For the majority game, the class with highest probability is predicted, while for the minority game the class with lowest probability is predicted. Consequently, an agent endowed with the knowledge of all possible strategies, selecting at every time step the strategy with the highest performance on the past L steps, is equivalent to a decision tree calibrated on the L past returns.

However, in the agent based model, the heterogeneity in believes among the agents is modeled by having each agent track a subset of all possible strategies. Therefore, the optimal strategy defined by the decision tree may not be available to all agents. Nonetheless, the classification and regression tree algorithm allows for an efficient selection of the best strategy in a subset of all strategies. First, the optimal strategy is determined, then the best strategy in a subset of strategies is determined by finding the strategy with the smallest Hamming distance to the optimal strategy. Formally, the strategy selection reads

$$f_{best} = \underset{f \in \mathbb{S} \subseteq \mathbb{S}_{\mu}^{m}}{\operatorname{argmax}} \sum_{h \in \mathcal{R}^{m}} \left| f\left(h\right) - \mathcal{T}\left(h\right) \right|, \tag{5.2}$$

where $\mathcal{T}\left(\cdot\right)$ is the tree prediction function. In the case of the minority game, the tree prediction function is defined to return the class with the lowest probability. Using this algorithm, the performance of each individual strategy no longer needs to be determined. It suffices to compute the class probabilities once, and find the strategy with the smallest distance to the optimal strategy.

5.2 Motivating tree based forecasting

5.2.1 Linear filter models & technical trading rules

Wold's decomposition theorem (Mills and Markellos, 2008, Hamilton, 1994) states that every weakly stationary, purely non-deterministic stochastic process $\{X_t\}$ can be written as a linear filter with infinite lag. All time series analysis models defined as a finite-order stochastic difference equation derive from this general concept. The simplest stationary models being the autoregressive model AR(ϱ) and moving average model MA(ϱ) of order ϱ. Non-stationary models can be built using nested stationary models for the mean and variance.

These models of stochastic processes are foremost characterized by the autocorrelation function of returns and absolute returns. While these characteristics are predominantly used to described financial returns, several stylized facts such as gain/loss asymmetry and heavy tails are well document (Cont, 2001). The gain/loss asymmetry is often described by the skewness of the return distribution and the heavy tails by the kurtosis. However, possible non-linear dependencies are often neglected.

Non-linear stochastic processes arise when their representation is obtained by some non-linear mechanism, for example polynomial dependencies in the innovation, asymmetric innovations, correlated innovations, time varying parameters, or regime switching. Unfortunately, testing for non-linearity is a challenging task and not possible in general when the functional form of the non-linearity is unknown (Mills and Markellos, 2008, chap. 6). Past efforts have been concentrated on modeling the common stylized facts such as skewness, fat tails, volatility clustering, and regime switching, using non-linear combinations of linear filter models (Mills and Markellos, 2009). This continued use of linear models as building blocks for non-linear models leaves the possibility that deterministic non-linearly separable patterns have remained undetected in past studies

To forecast trends in stock markets, traders have developed an extensive taxonomy of technical trading rules expressing their beliefs about the market behavior. Some of these rules rely on a linear filter model such as moving average, which forecasts a trend persistence. While a majority of rules define trading triggers based on support and resistance bands, channels, and oscillators, these rules do not focus on a potential intrinsic structure of market returns, but on prices levels psychologically important to a large number of traders. Assuming that a majority of market participants acts based on these price levels, these rules should have significant profitability.

A common feature of linear filter models and technical trading rules is that they do not test for some potential non-linear dependencies. An example are the return sign correlations, which have been well documented by Christoffersen and Diebold (2003) and Christoffersen et al. (2006). The presence of non-linear return sign correlation in a large number of financial assets is further supported by the findings of Niederhoffer and Osborne (1966), Zhang (1999), Leung et al. (2000), Zunino et al. (2009), and James et al. (2014). The existing assessments of the EMH are focused on testing the performance of technical trading strategies and funds. However, the non-linear anomalies around return sign correlations have not undergone a rigorous analysis to the best of our knowledge. These anomalies are particularly interesting because many linearly inseparable patterns, as for example the logical exclusive-OR function (XOR), would remain undetected with linear models such as moving averages. Nonetheless, such patterns can be detected using statistical learning models such as decision trees.

This section argues that return sign correlations can remain undetected in the autocorrelation function, and propose decision trees as a forecasting model to detect return sign correlations. The limitations of the autoregressive models to capture the predictable XOR pattern are derived, and it is shown how decision tree based models overcome this limitation. The different variants of decision trees are discussed, and the issue of overfitting is addressed. Especially, a connection between fixed decision trees and Markov chains is derived. The connection is used to create a binary Markov process of order ϱ, analogous to the autoregressive process of order ϱ, but with more intricate non-linear autocorrelation patterns. It is then proven that the parameters of an autoregressive process of order ϱ have zero expectation for a binary Markov process satisfying certain conditions.

5.2.2 From autoregressive to tree based models

The autoregressive model AR(ϱ) defines the evolution of the time series $\{X_t\}$ with ϱ lags[1] as

$$X_t = \phi_0 + \sum_{i=1}^{\varrho} \phi_i X_{t-i} + a_t, \tag{5.3}$$

with parameters $\phi = (\phi_0, \phi_1, \ldots, \phi_\varrho)$, and i.i.d. innovations a_t. The shortcoming of autoregressive models can be illustrated with the two argument exclusive-OR function $XOR(r[-1], r[-2])$ that returns true ($= 1$) when exactly one of the arguments is true and

---

[1] ϱ is used as order parameter (i.e. lag) instead of the common p to avoid confusions with probabilities denoted by p.

false ($= -1$) otherwise (see Figure 5.3). An example of XOR like data is

$$\mathbf{X} = \left\{ \left(\left(1,\, 1 \right),\, -1 \right),\, \left(\left(1,\, -1 \right),\, 1 \right),\, \left(\left(-1,\, 1 \right),\, 1 \right),\, \left(\left(-1,\, -1 \right),\, -1 \right) \right\}, \tag{5.4}$$

where the four samples are assumed to be at independent times. The notation of Equation (5.4) is taken from the statistical learning literature (James et al., 2014), where $\mathbf{X}$ is the training data available, and $\left(\left(X_{t-2},\, X_{t-1} \right),\, X_t \right)$ denotes the sample at time t with input $x_t = \left(X_{t-2},\, X_{t-1} \right)$ and output (=response) $y_t = X_t$. Calibrating the autoregressive model of order two to the data of Equation (5.4) yields the parameters

$$\left(\phi_0 = 0,\, \phi_1 = 0,\, \phi_2 = 0 \right), \tag{5.5}$$

which fail at capturing the deterministic XOR function. As we will discuss in section 5.3.2, almost XOR like patterns can arise in time series data.

Non-linearly separable patterns can be modeled using a partition $R = \{R_1,\, \ldots,\, R_n\}$ of the input (or feature) space into n regions, and assigning the constant values $\{c_1,\, \ldots,\, c_n\}$ to each region. The resulting evolution of the time series X_t can then be written as

$$y_t = X_t = \sum_{i=1}^{n} c_i \cdot I \left\{ x_t \in R_i \right\} + a_t, \tag{5.6}$$

where I is the indicator function, and a_t are i.i.d. innovations. This modeling approach allows for an arbitrary flexibility, as any function can be approximated to any precision with a sufficient number of regions. For example, the XOR data from Equation (5.4) can be modeled exactly by the two regions $R_1 = \left\{ x \in \mathbb{R}^2 | x_1 x_2 \geq 0 \right\}$ with $c_1 = -1$, and $R_2 = \left\{ x \in \mathbb{R}^2 | x_1 x_2 < 0 \right\}$ with $c_2 = 1$. The downside of this modeling approach is that the number of parameters increases arbitrarily with the number of regions, and a procedure to control for overfitting is required.

What remains unspecified in the stochastic process of Equation (5.6) is the algorithm to estimate the regions based on a given realization $\{X_t\}_1^t$ of a process. In general, regions of arbitrary shape and overlaps can be used. However, for the purpose of this study, decision tree models provide sufficient flexibility. Decision trees find rectangular regions, using a recursive splitting algorithm of the input space that minimizes a loss function for the given training data. While several variations of the splitting algorithms exist, the most popular Classification And Regression Tree (CART) algorithm described by Hastie et al. (2001, chap. 9) will be used.

In the context of financial returns with low signal to noise ratio, finding the best tree is an NP-complete problem (Hyafil and Rivest, 1976) that is not computationally feasible. This stands in stark contrast to autoregressive models were the global optimum can be estimated. However, this is not an issue as the CART algorithm is nonetheless deterministic by using recursive binary splitting, and subsequent pruning. The algorithm always finds the same tree for a given training data, allowing practitioners to independently obtain the same forecast. The definition of the splitting algorithms for regression and

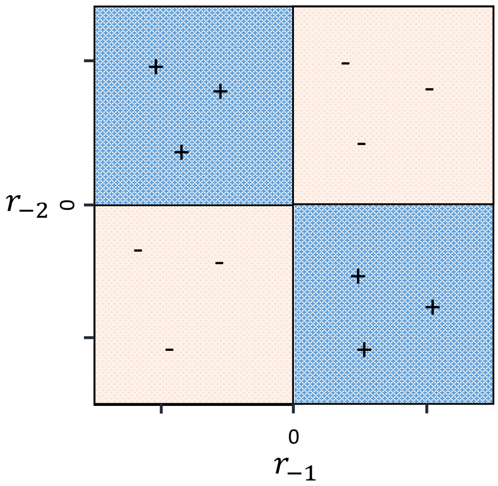


Figure 5.3: **Illustration of the XOR pattern**

classification are given below, as well as a discussion of robustness in the context of highly stochastic data.

5.2.3 Regression tree algorithm

The CART algorithm constructs a regression tree using the Mean Squared Error (MSE) as the loss function $Q_{MSE}(y, \hat{y}) = (y - \hat{y})^2$. For a given training data

$$\mathbf{X} = \{(x_1, y_1), \ldots, (x_N, y_N)\} \tag{5.7}$$

and partitioning of the input space into regions $R = \{R_1, \ldots, R_n\}$, the algorithm assigns the response

$$\hat{c}_i = \langle \{y_i | x_i \in R_i\} \rangle \tag{5.8}$$

to region R_i, which is the mean observed output in that region. The binary splitting algorithm adds one new region at each iteration by splitting an existing region into two. At each iteration an existing region R is split into two halves with a plane defined by variable j and split point s as

$$R^1(j, s) = \{x | x_j \leq s\} \text{ and } R^2(j, s) = \{x | x_j > s\}. \tag{5.9}$$

The optimal split is given by

$$\min_{j,\, s} \left[\sum_{x_i \in R^1} Q(y_i, \hat{c}_1) + \sum_{x_i \in R^2} Q(y_i, \hat{c}_2) \right], \tag{5.10}$$

which can be determined efficiently by running through all the input variables and split points defined by the inputs.

In many scenarios, for example of XOR like training data, a split with not improvement in the loss can be followed by a split with a large reduction of the loss. This prevents the introduction of termination criterion in the splitting procedure. To overcome this issue, the fully grown tree $\mathcal{T}_0$ is subsequently pruned. The pruning procedure finds an optimal subtree $\mathcal{T}$ without the internal nodes that do not affect significantly the loss. The loss of a subtree $\mathcal{T} \subset \mathcal{T}_0$ is given by

$$Q\left(\mathcal{T},\, \alpha\right) = \sum_{m=1}^{|\mathcal{T}|} N_m Q_m\left(t\right) + \alpha \left|\mathcal{T}\right|, \tag{5.11}$$

where $Q_m\left(\mathcal{T}\right) = \frac{1}{N_m} \sum_{x_i \in R_m} Q\left(y_i,\, \hat{c}_m\right)$ is the loss of the terminal node m,

$$N_m = \#\left\{x_i \in R_m\right\} \tag{5.12}$$

is the number of samples in that node, and $|\mathcal{T}|$ is the number of terminal nodes in the tree. The regularization coefficient α of the tree size determines to which degree smaller trees are favored with respect to higher loss. The value $\alpha = 0$ would yield the fully grown tree $\mathcal{T}_0$.

In the context of predicting financial returns, the signal to noise ratio in the samples is intrinsically low, and without appropriate pruning the data would be overfit. However, the used Scikit-learn library (Pedregosa et al., 2011) does not have a pruning parameter to control for overfitting, but instead sets a lower bound on the number of samples in a terminal node. Fortunately, determining the adequate number of samples in a terminal node is simpler then determining an equivalent pruning parameter. The lower bound on the number of samples in a terminal node of a classification tree is computed in the following Subsection 5.2.4. The lower bound for a regression tree is set to the lower bound value of the equivalent classification tree.

We remark that the objective of the MSE loss function, namely minimizing variance within each region, is not necessarily optimal for forecasting financial returns. The sub-optimality arises when the returns within a region exhibit low variance and mean close to zero. The insignificant mean implies that no forecast can be made. However, this does not negatively impact performance as return forecasts of small amplitude can be discarded as insignificant.

5.2.4 Classification tree

The dominantly stochastic behavior of financial returns translates into particularly low statistical significance of regression model forecasts. However, the statistical significance of a potential signal can eventually be improved by mapping the outputs to categorical outputs. Under the assumption that the mapping removes more noise then signal, the resulting classification problem is more robust. The simplest such mapping is $r \to \operatorname{sign}\left(r\right)$ that results in predicting only an up and down moves. The CART algorithm described in

Section 5.2.3 only needs one modification for classification, namely a different loss function because the MSE loss is not suitable for classification. Typically, the classification of K classes (e.g. $\{-, +\}$ with $K = 2$) is performed using the Gini index $Q_{Gini}(\vec{p}) = \sum_{i \neq j}^{K} p_i p_j$, which maximizes the probability p_i of a single class in a region.

In the context of predicting market up or down moves, the Gini index naturally sets the right objective, as it will maximize the predictability of an up or down move in a given region. However, the Gini index by itself is prone to overfitting, as terminal nodes with a few samples of the same kind are favored despite such nodes being a likely occurrence under the null hypothesis of no predictability. To alleviate this problem, a lower bound on the number of samples in a terminal node is introduced.

The probability distribution of k up moves ($\uparrow$) and $n - k$ down moves ($\downarrow$), assuming equal class probability, is given by the binomial distribution

$$P\left(\# \uparrow = k,\, \# \downarrow = n - k\right) = \left(\frac{1}{2}\right)^{n} \begin{pmatrix} n \\ k \end{pmatrix}. \tag{5.13}$$

For sufficiently large n, this binomial distribution can be approximated by the normal distribution $\mathcal{N}\left(\frac{n}{2}, \frac{1}{4n}\right)$. Requiring that the prediction is significant at the one standard deviation and the class probabilities satisfy $|p_+ - p_-| = 2 \cdot 0.05$, a lower bound on n can be computed. The constraints imply $\frac{1}{4n} = 0.05^2$, or $n \geq 100$, which would require at least 100 samples per terminal node. This illustrates the difficulty of obtaining statistically significant forecasts and the large amount of calibration data required. In general, the minimum number of samples per terminal node should be picked as large as possible.

5.2.5 Fixed tree

Similarly to the classification presented in Section 5.2.4, the signal to noise ratio in the input variables can be improved by a mapping of the inputs onto the two categories $\{-, +\}$. In case of inputs with ϱ lags, this results in the discrete input space $\{-, +\}^{\varrho}$ with 2^{ϱ} elements. For a training set $\mathcal{D}$, the limited number of possible inputs controls the expected number of samples in a leaf R as $E\left[N_R\right] = \frac{|\mathcal{D}|}{2^p}$. Therefore, with an appropriate choice of ϱ with respect to the sample size $|\mathcal{D}|$, the problem of overfitting does not arise. This allows us to remove the lower bound on the number of samples per leaf, resulting in the CART algorithm to generate the fully grown tree with regions $\{R_1, \ldots, R_{2^{\varrho}}\}$ (i.e. one region for each input). In other terms, the tree is fixed by the number of lags ϱ, and the CART algorithm reduces to compute the class probabilities in each leaf. A fixed tree can be used for classification, prediction of the majority vote inside each leaf, or regression by predicting the mean return of a leaf.

5.2.6 Probit models & polynomial features

Probit models could be used as the autoregressive equivalent of classification trees, calibrating the parameters using only binary up and down returns. Nonetheless, just as the autoregressive model, the probit model cannot capture linearly inseparable patterns like the XOR function. To capture non-linear patterns, the autoregressive or probit models have to be extended with non-linear features. For example, the XOR function can be described with the quadratic feature $r[t-1] \cdot r[t-2]$.

However, polynomial features are not robust with respect to skewed distributions or multiple patterns offset with respect to the origin (e.g. offset as $(r[t-1]-x_1)\,(r[t-2]-x_2)$). In contrast, decision trees are non-linear predictors robust with respect to both issues. The CART algorithm can find localized non-linear predictability independently of its location in the input space, and use multiple regions to approximate skewness.

5.3 Comparing the forecasting power of autoregressive & tree models

5.3.1 Connecting fixed trees to Markov chains

Given a fixed decision tree, the time evolution defined by Equation (5.6) defines as well the transition probability p from the sample at time t to the sample at time $t+1$ as

$$\left(\left(X_{t-\varrho},\, \ldots,\, X_{t-1}\right),\, X_t\right) \xrightarrow{p} \left(\left(X_{t-\varrho+1},\, \ldots,\, X_t\right),\, X_{t+1}\right). \tag{5.14}$$

The sample inputs $(X_{t-\varrho},\, \ldots,\, X_{t-1})$ define 2^ϱ possible states $\{S_1,\, \ldots,\, S_{2^\varrho}\}$, and the time evolution defines the two possible transitions from each state to another state, while the remaining $2^\varrho - 2$ states are unattainable. In other terms, the fixed decision tree defines a ϱ-th order Markov chain with 2^ϱ states, and two time dependent non-zero outgoing transition probabilities from each state.

We denote by $\boldsymbol{P} \in [0,\,1]^{2^\varrho \times 2^\varrho}$ the time independent transition matrix of the Markov chain defined by a fixed tree $\mathcal{T}$, resulting from the mapping of the returns onto the states $\{-,\,+\}$. The elements $\boldsymbol{P}_{ij}$ of the transition matrix describe the probabilities $P\left(S_j \to S_i\right)$ to go from state S_j to state S_i. The probabilities to transition from one state to any of the states must always sum to one, imposing $\sum_{i=1}^{2^\varrho} \boldsymbol{P}_{ij} = 1,\ \forall j$. Due to the particular structure of the tree, each state has two incoming transitions, and two outgoing transitions. This implies that each row and column in $\boldsymbol{P}$ has two non-zeros entries.

This Markov chain admits a stationary distribution if there exists a vector of probabilities $\pi \in [0,\,1]^{2^\varrho}$, satisfying $|\pi| = \sum_{i=1}^{2^\varrho} |\pi_i| = 1$ and

$$\boldsymbol{P}\pi = \pi. \tag{5.15}$$

The component π_i of the vector π represents the probability to be in state S_i at the stationary regime. The equality in Equation (5.15) asserts the stationarity condition that the probability of each state is invariant when transitioning to the next state. The stationary distribution π is an eigenvector of $\boldsymbol{P}$ with unit eigenvalue. The eigenvalues $\{\lambda_1,\, \ldots,\, \lambda_{2^\varrho}\}$ of $\boldsymbol{P}$ all satisfy $0 \leq \lambda_i \leq 1$, and consequently the stationary distribution is always determined by the eigenstates (i.e. eigenvectors) with unit eigenvalues, while the eigenstates associated to eigenvalues $\lambda < 1$ decay away. The second largest eigenvalue determines how quickly the stationary distribution is reached. Multiple eigenstates with unit eigenvalue arise in the case of reducible chains composed of several independent Markov chains. This can be shown by the orthogonality property of eigenstates and all their components being strictly positive in the context of Markov chains.

5.3.2 Binary Markov based processes

In the case of binary categories $\{-, +\}$, namely up ($X_t \geq 0$) and down ($X_t < 0$) moves, the Markov chain defined in Section 5.3.1 has 2^ϱ possible states. Each state $S_i \in \{-, +\}^\varrho$ has two outgoing transitions probabilities p_{i+} and p_{i-}, the probability of an up move, respectively a down move. These two outgoing probabilities are subject to $p_{i+} + p_{i-} = 1$, and we can characterize them by a single variable $\Delta p_i = p_{i+} - p_{i-}$, where $p_{i+} = \frac{1}{2}(1 + \Delta p_i)$ and $p_{i-} = \frac{1}{2}(1 - \Delta p_i)$. Each state S_i has as well two incoming transition probabilities, which we denote by $p_{+i} = P(S_{+i} \to S_i)$ and $p_{-i} = P(S_{-i} \to S_i)$, where the first index ($+$ or $-$) stands for the first sign of the previous state (e.g. $S_i = (-, +)$, $S_{+i} = (+, -) \to S_i$, and $S_{-i} = (-, -) \to S_i$). We remark that the sets of outgoing and incoming probabilities are in one-to-one correspondence. The duplicate definition of the transition probabilities is introduced for subsequent convenience.

The stationary distribution π satisfies

$$\pi_{+i} \cdot p_{+i} + \pi_{-i} \cdot p_{-i} = \pi_i, \; \forall 1 \leq i \leq 2^\varrho, \tag{5.16}$$

where π_{+i} and π_{-i} stand for the probabilities at stationarity of the state S_{+i}, respectively S_{-i}. I remark that the outgoing probabilities of each state must always sum to one, but the incoming probabilities can be arbitrary $p_{+i}, p_{-i} \in [0, 1]$. In the special case of incoming probabilities $p_{+i} + p_{-i} = 1$, the stationary distribution is given by $\pi_i = \frac{1}{2^\varrho}, \forall i$. In other terms, the stationary distribution with uniform probability for all states arises when the incoming probabilities of each state (i.e. rows in $\boldsymbol{P}$) sum to one. This result can be derived from Equation (5.15) by setting $\pi = \frac{1}{2^\varrho}\vec{1}$, which leads to $\boldsymbol{P}\frac{1}{2^\varrho}\vec{1} = \frac{1}{2^\varrho}\vec{1} \Rightarrow \sum_{j=1}^{2^\varrho} \boldsymbol{P}_{ij} = p_{+i} + p_{-i} = 1, \forall i$.

Based on the Markov chain, I define the Data Generating Process (DGP)

$$X_{t+1} = \begin{cases} + \left|\mathcal{N}\left(a_{t+1} \middle| 0, \sigma\right)\right| & \text{with probability } p_{t+} \\ - \left|\mathcal{N}\left(a_{t+1} \middle| 0, \sigma\right)\right| & \text{with probability } p_{t-} \end{cases}, \tag{5.17}$$

where p_{t+} and p_{t-} are the outgoing transition probabilities for the state S_t at time t.

In order for this DGP to generate a continuous normal distribution, without jump at $X_t = 0$, the number of up and down moves needs to be equal, which is enforced by the

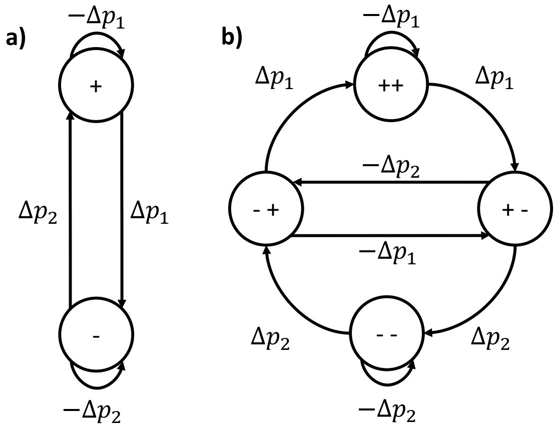


Figure 5.4: **Illustration of binary Markov processes and associated parameters: a) one lag ($\varrho = 1$); b) two lags ($\varrho = 2$) and stationary regime with equal probabilities for all states.**

condition

$$B(\boldsymbol{P}) = \sum_{i=1}^{2^{\varrho}} \pi_i \Delta p_i = 0. \tag{5.18}$$

As the stationary vector π is a function of $\boldsymbol{P}$, the Equation (5.18) is a non trivial polynomial equation in $\boldsymbol{P}$. However, in the case where all states are equally likely, the balance of up and down moves is achieved when $\sum_{i=1}^{2^{\varrho}} \Delta p_i = 0$.

I remark that the balancing condition is not the only possibility to obtain a continuous distribution. The balance between up and down moves can be broken locally as long as the balancing Equation (5.18) has expectation zero: $E[B(\boldsymbol{P})] = 0$. Assuming sufficiently low autocorrelation over time for the balancing term $B(\boldsymbol{P})$, the discontinuity of the return distribution in finite samples cannot be detected in a statistically significant manner. Another possibility is to remove the discontinuity with an asymmetry of the left and right tail distribution. The later case is regularly observed in equity indices where up moves are more likely but the negative returns have a larger tail (i.e. distribution skewness).

Figure 5.4.a shows the binary Markov process with lag one ($\varrho = 1$). Figure 5.4.b shows the two degrees of freedom in the binary Markov process with lag two ($\varrho = 2$), uniform state probability, and equal number of up and down moves. When picking $\Delta p_1 = 0.5$ and $-0.5 < \Delta p_2 \ll 0$ the binary Markov process with two lags produces almost exactly the XOR pattern. The binary Markov processes with three and four lags have more then two degrees of freedom when fulfilling the uniform state probability and balancing conditions.

5.3.3 Expected autoregressive forecast for binary Markov processes

The Markov DGP defined in Equation (5.17) violates the Martingale condition

$$E[X_{t+1} | X_{t-\varrho+1}, \ldots, X_t] = 0 \tag{5.19}$$

when there exists a state S_i such that $\Delta p_i \neq 0$. This subsection determines to which degree the deterministic pattern can be predicted using an AR(ϱ) autoregressive model as defined in Equation (5.3). Assuming homoskedasticity of the inputs $\boldsymbol{X}$ (not to be confused with the samples $\mathbf{X}$) and outputs $\boldsymbol{Y}$, the unbiased least square regression estimator reads

$$\hat{\phi} = \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{Y}. \tag{5.20}$$

To compute the expectation of this estimator for the binary Markov DGP I notice that for every state $S \in \boldsymbol{S} = \{-, +\}^{\varrho}$ there is a reverse state $-S$. Therefore, the set of states can be split into two as $\boldsymbol{S} = \boldsymbol{S}^+ \cup \boldsymbol{S}^-$, where $|\boldsymbol{S}^+| = |\boldsymbol{S}^-|$ and $S_i^+ = -S_i^-$, $\forall 1 \leq i \leq 2^{\varrho-1}$. Each state is paired to its reverse. Assuming n observed samples, the inputs $\boldsymbol{X}$ belong to the space $\boldsymbol{S}^{\times n}$, and the distribution of $\boldsymbol{Y}$ is defined by Equation (5.17). Using these assumptions we compute

$$E\left[\boldsymbol{X}^T \boldsymbol{Y}\right] = \sum_{i=1}^{2^{\varrho-1}} \alpha \left(\pi_i^+ S_i^+ p_{i+}^+ + \pi_i^- S_i^- p_{i-}^-\right), \tag{5.21}$$

where α is a normalization factor stemming from the half-normal distribution of the variables. In the case of uniform probabilities for all states, the expectation simplifies to

$$E\left[\boldsymbol{X}^T \boldsymbol{Y}\right] = \frac{\alpha}{2^{\varrho}} \sum_{i=1}^{2^{\varrho-1}} S_i^+ \left(p_{i+}^+ - p_{i-}^-\right). \tag{5.22}$$

The condition $\Delta p_i^+ = \Delta p_i^-$ implies $E\left[\boldsymbol{X}^T \boldsymbol{Y}\right] = 0$ as

$$p_{i+}^+ - p_{i-}^- = \frac{1}{2}\left(1 + \Delta p_i^+ - 1 - \Delta p_i^-\right) = 0,$$

and at the same time fulfills the balancing condition of Equation (5.18). As a result, the estimated parameters $\hat{\phi}$ are zero, and the autoregressive model is unable to predict the deterministic pattern in the Markov process.

The binary Markov process with lag $\varrho = 1$ only has the two states $\{-, +\}$, which can be split into $\boldsymbol{S}^+ = \{+\}$ and $\boldsymbol{S}^- = \{-\}$, and is determined by the two parameters Δp_1^+ $(= \Delta p_1)$ and Δp_1^- $(= \Delta p_2)$. This process has zero expectation for an autoregressive model when $\Delta p_1^+ = \Delta p_2^+$, which implies the stationary regime $\pi_+ = \pi_- = \frac{1}{2}$. However, the balancing condition of Equation (5.18) is incompatible as it is satisfied when $\Delta p_1^+ = -\Delta p_2^+$. Nonetheless, as discussed in Section 5.3.2, the balancing condition is not necessary as the return distribution can be skewed.

The binary Markov process with two lags can satisfy the balancing condition and have zero expectation for an autoregressive process. The DGP including the degrees of freedom fulfilling these conditions are shown in Figure 5.4.b.

5.3.4 Expected fixed tree forecast for autoregressive processes

Assuming an autoregressive DGP AR(ϱ), I want to determine how well a fixed regression tree with ϱ lags is able to predict the deterministic component arising when $\|\phi\| > 0$. The fixed regression tree assigns to every of the 2^ϱ input states $S_i = \left(S_{i,1}, \ldots, S_{i,\varrho}\right) \in \boldsymbol{S}$ the mean

$$\hat{c}_i = E\left[X_{t+1} | \operatorname{sign}\left(X_{t-\varrho+1}\right) = S_{i,1}, \ldots, \operatorname{sign}\left(X_t\right) = S_{i,\varrho}\right]. \tag{5.23}$$

In an autoregressive process with normally distributed innovations, the lagged variables are distributed as $x_t = \left(X_{t-\varrho+1}, \ldots, X_t\right) \sim \mathcal{N}\left(0, \Lambda\left(\phi\right)\right)$, where $\Lambda\left(\phi\right) \in \mathbb{R}^\varrho$ is the covariance matrix of the ϱ lags, assuming the independent innovations are distributed as $a_t \sim \mathcal{N}\left(0, \sigma\right)$.

Due to the autocorrelation, there is a directional bias of the return dependent of the previous ϱ returns. This bias can be computed for a region $R_i = \left\{x \in \mathbb{R}^\varrho | x_j S_{i\,j} \geq 0, \forall j\right\}$ to be

$$P\left(y_t \geq 0 | x_t \in R_i\right) = \int_{R_i} \mathcal{N}\left(x | 0, \Lambda\left(\phi\right)\right) \int_{-\phi\cdot x}^{+\infty} \mathcal{N}\left(a | 0, \sigma\right) dx da. \tag{5.24}$$

In the general case of an arbitrary autoregressive parameter ϕ, this integral cannot be evaluated in closed form.

The simplest case AR(1) has the analytically solution

$$P\left(y_t \geq 0 | x_t \in R_+\right) = \frac{1}{2} + \frac{1}{\pi} \arctan\left(\frac{\phi}{1-\phi^2}\right), \tag{5.25}$$

where $R_+ = \mathbb{R}^+$. As expected, the bias is independent of the variance of the innovations. The directional bias for the region $R_- = \mathbb{R}^-$ is obtained by symmetry as $P\left(y_t \geq 0 | x_t \in R_-\right) = 1 - P\left(y_t \geq 0 | x_t \in R_+\right)$. Assuming the calibration data is sufficient for a decision tree to predict the bias correctly every time, the directional accuracy of the tree will be $P\left(y_t \geq 0 | x_t \in R_+\right)$. An AR(1) predictor with correct parameter ϕ always makes the same prediction as a decision tree in this scenario, and therefore the two predictors have identical directional accuracy. As an example, for $\phi = 0.1$ the directional accuracy is ≈ 0.532.

Determining the directional accuracy for the AR(2) case leads to integrals over multivariate distribution that cannot be solved in closed form. Hence, only numerical solutions are possible and would require a lengthy computation analyzing each region separately. Nonetheless, let us remark what happens in the regions R_{+-} and R_{-+} for the parameter choice $\phi = \left(\phi_0 = 0, \phi_1, \phi_2 = \phi_1\right)$. The value of $\phi \cdot x$ for $x = \left(x_1, x_2\right) \in R_{+-}$ is anti-symmetric with respect to the axis $x_2 = -x_1$. As a consequence of this anti-symmetry, the integral defined in Equation (5.24) takes the constant value $P\left(y_t \geq 0 | x_t \in R_i\right) = \frac{1}{2}$, and this region has no predictability bias for a fixed tree. However, the regions R_{++} and R_{--} exhibit a predictability bias similar to the AR(1) case, and therefore the fixed tree predictor will have roughly half the directional accuracy of the autoregressive predictor in

Model	Description	
$\boldsymbol{AR}$	**A**uto-**R**egressive	$\varrho \in \{1, 2, 3, 4\}$ $L \in \{10, 20, \ldots, 500\}$
$\boldsymbol{RT_{MSE}}$	**R**egression **T**ree with MSE loss	
$\boldsymbol{CT_{Gini}}$	**C**lassification **T**ree with Gini index loss	
$\boldsymbol{FRT}$	**F**ixed **R**egression **T**ree with mean prediction	
$\boldsymbol{FCT}$	**F**ixed **C**lassification **T**ree with majority vote	

Table 5.1: **Overview of the strategies compared in this study**. A strategy is defined by a model, the order parameter ϱ, and the calibration window length L. A total of $5\,\text{models}\times 4\,\text{lags}\times 50\,\text{lengths} = 1000\,\text{strategies}$ are tested.

this specific case. Hence, fixed trees can be a sub-optimal choice to predict autoregressive processes with two lags or more. Nonetheless, they are more robust then autoregressive models, which are entirely unable to predict a binary Markov process. Additionally, the CART algorithm can generate a regression tree approximating arbitrarily well the autoregressive process, assuming sufficient calibration data is available. In reverse, the autoregressive model is always to rigid to capture the binary Markov process.

5.3.5 Simulation study

To confirm the theoretical results, the statistical significance of the competing models is simulated on an autoregressive process of order two and a binary Markov process of order two. The statistical significance is studied based on the studentized Sharpe ratio as a function of the sample size, the calibration window length, and the autocorrelation parameters.

At each time step, a model M is calibrated on the past L returns to forecast one step ahead as $\tilde{X}_{t+1} = M\left(\{X_t\}_{t-L+1}^{t}\right)$, where $\tilde{X}_{t+1}$ is the forecast of the model. For the purpose of this study, the forecast $\tilde{X}_{t+1}$ is then mapped to a long or short trading signal defined by $s_{t+1} = \text{sign}\left(\tilde{X}_{t+1}\right)$[2]. This produces a sequence of binary signals $\{s_t\}_{L+1}^{T} \in \{-1, 1\}^{t-L-1}$ that define the corresponding trading strategy on the returns $\{X_t\}_1^t$. Hence, a strategy is defined by a model, the number of lags ϱ, and the calibration window length L. Table 5.1 presents an overview of the models and the parameters of the associated strategies.

The first L returns in the sequence $\{X_t\}_1^t$ are part of the calibration window of size L, and only the subsequent returns are forecast one step ahead using a rolling window of size L. Nevertheless, for further convenience, the number of forecast returns is defined by T, implicitly assuming that L initial returns have been cropped before the first forecast.

---

[2] We remark that in general continuous trading signals can be constructed based on the value of a regression forecast or the class probabilities of a classifier. However, within this paper we evaluate forecasting models only based on binary signals, as more complex trading strategies are out-of-scope for the purpose of studying the predictive power of decision trees.

5.3.5.1 Setup

As finite sample realizations of a DGP have large variance, the p-value for a strategy on a DGP has to be computed as an average over multiple runs. To obtain a reasonably small confidence interval, a p-value has to be averaged over 5000 runs with 500 bootstrap iterations for the data matrix of each run. The p-value for a given strategy and DGP depends on three parameters, namely the autocorrelation parameters ϕ or Δp of the DGP (implicitly defining the lag ϱ), the number of forecast returns t, and the calibration window length L. Given the large computational expense of a single p-value, and the large parameter space, a small subset of simulations have been chosen to study the impact of the different parameters.

Following the theoretical analysis of Section 5.3.1, the two compared data generating processes are: DGP_1, an autoregressive process AR(2) with the two degrees of freedom $\phi = (\phi_1,\, \phi_2)$; and DGP_2, a binary Markov based process M(2) defined in Equation (5.3.2), with the two degrees of freedom $\Delta p = (\Delta p_1,\, \Delta p_2)$ as shown in Figure 5.4.b. Realizations of both processes are generated with normally distributed innovations $a_t \sim \mathcal{N}\,(0,\, \sigma = 1)$.

First, I study the behavior of the statistical significance as a function of the autocorrelation parameters, expecting significance to increase with autocorrelation. Two years of trading ($t = 500$ days) are simulated with a calibration length of $L = 50$.

Second is the verification that significance increases with the duration T as true skill becomes less likely to be a result of luck. In particular, the typical T needed to achieve a given significance level is determined. A moderately explosive process is picked with autocorrelation $\phi = \Delta p = (0.2,\, 0.2)$ and a calibration length of $L = 50$.

Third, the impact of the calibration length L on the significance level is explored. In finite samples, the estimation error of small autocorrelations is large, and therefore the calibration window length plays an important role. To illustrate this behavior, a small autocorrelation $\phi = \Delta p = (0.1,\, 0.1)$ and a sufficient number of time steps $t = 500$ is chosen.

5.3.5.2 Results

The simulated p-values as a function of the autocorrelation parameters ϕ and Δp are presented in Figure 5.5, showing the increase in statistical significance as the autocorrelation becomes stronger. To obtain a 95% confidence level for an individual strategy on a two year time frame, autocorrelations larger then (0.3, 0.3) need to be present.

Figure 5.6 shows the increase in statistical significance with the sample size T. Even for autocorrelation parameter $\phi = (0.2,\, 0.2)$, which is much higher then average on financial markets, at least $T = 800$ samples are needed before a 95% confidence level is achieved.

Figure 5.7 shows the decrease in significance with the calibration window length L. At least $L = 175$ samples for calibration are needed before the autocorrelation parameter $\phi = (0.1,\, 0.1)$ can be estimated robustly.

As computed in Section 5.3.3, the autoregressive forecasting model does not pick up

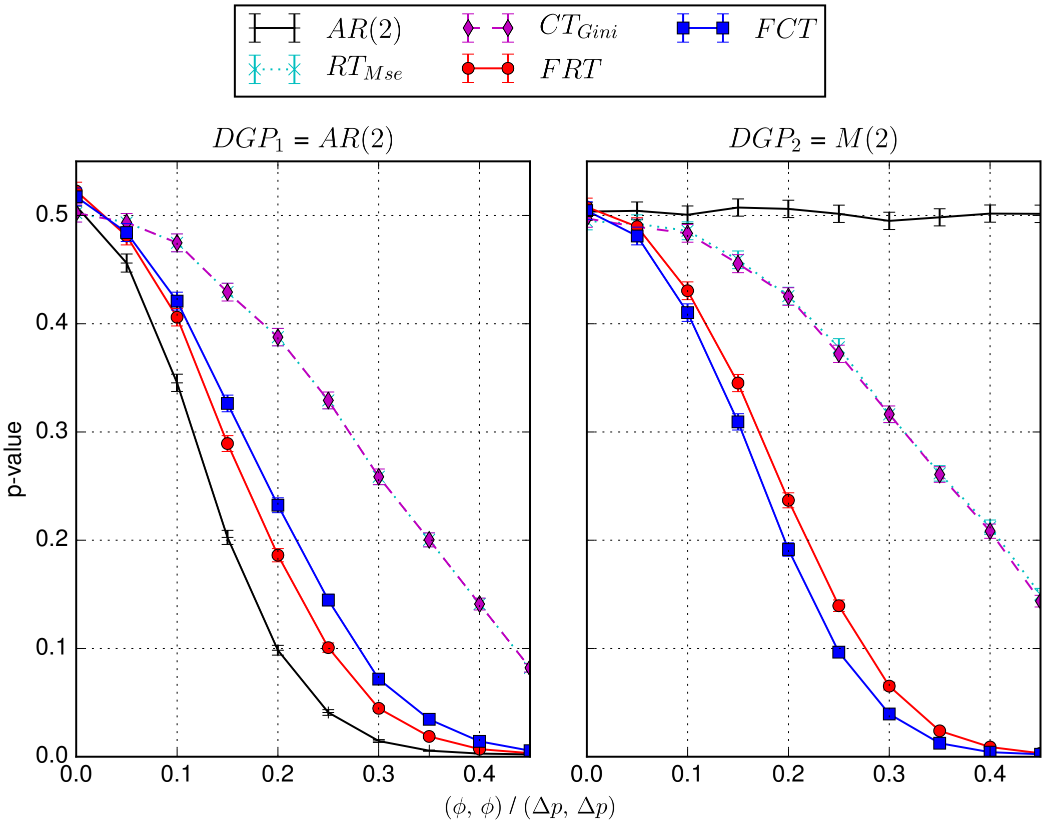


Figure 5.5: **P-values of all models when predicting an autoregressive process AR**(2) **and a binary Markov process M**(2) **as function of the parameters** $(\phi_1,\, \phi_2 = \phi_1)$ **and** $(\Delta p_1,\, \Delta p_2 = \Delta p_1)$**.** The prediction is made for sample size $T = 500$ and calibration window length $L = 50$. The regression tree with MSE loss and the classification tree with Gini index loss perform similarly on both DGPs, but have the lowest performance on the autoregressive processes. The fixed regression and classification tree perform almost identically on both DGPs. The autoregressive process is optimal at predicting itself, however fails entirely at predicting the binary Markov process. The fixed trees exhibit a small bias at $\phi = 0$ resulting from the effect described in Appendix 3.4.5.

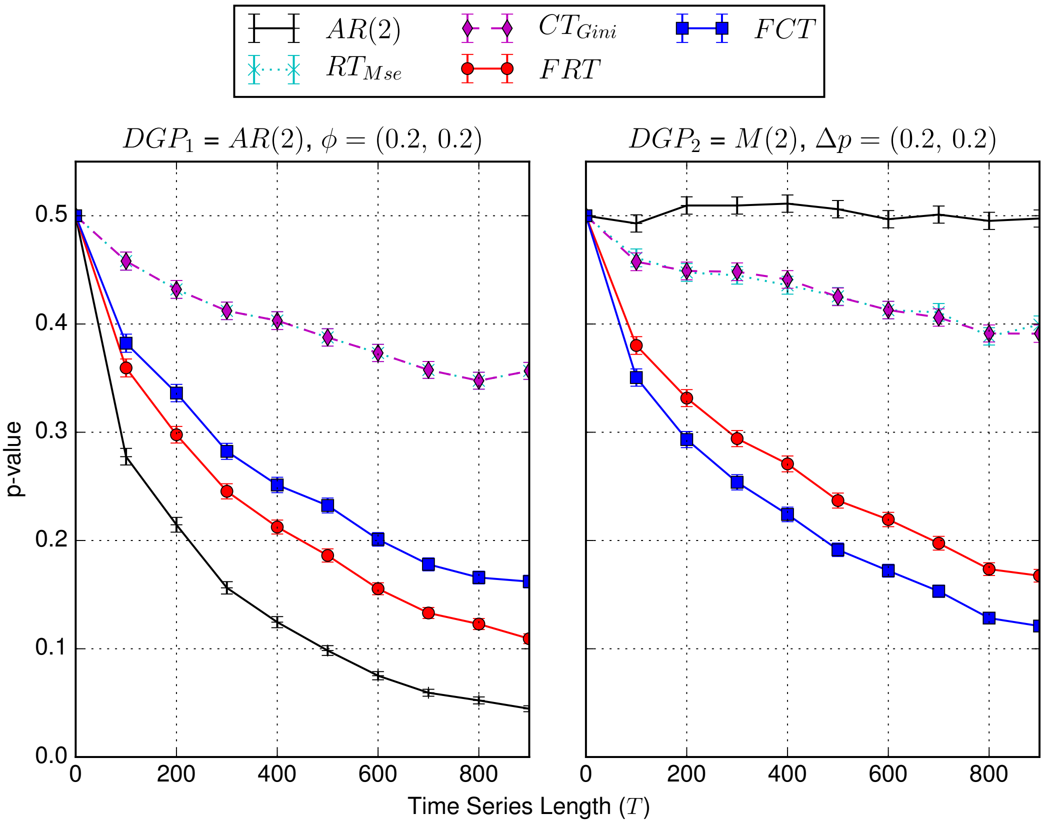


Figure 5.6: **P-values of all models with calibration length** $L = 50$ **when predicting an autoregressive process AR**(2) **and binary Markov process M**(2)**, with parameters** $(\phi_1 = 0.2,\, \phi_2 = 0.2)$**, respectively** $(\Delta p_1 = 0.2,\, \Delta p_2 = 0.2)$**.** The significance of the performance increases with the sample size T. The tree based models performs similarly on both DGPs. The autoregressive predictor only work on the autoregressive DGP.

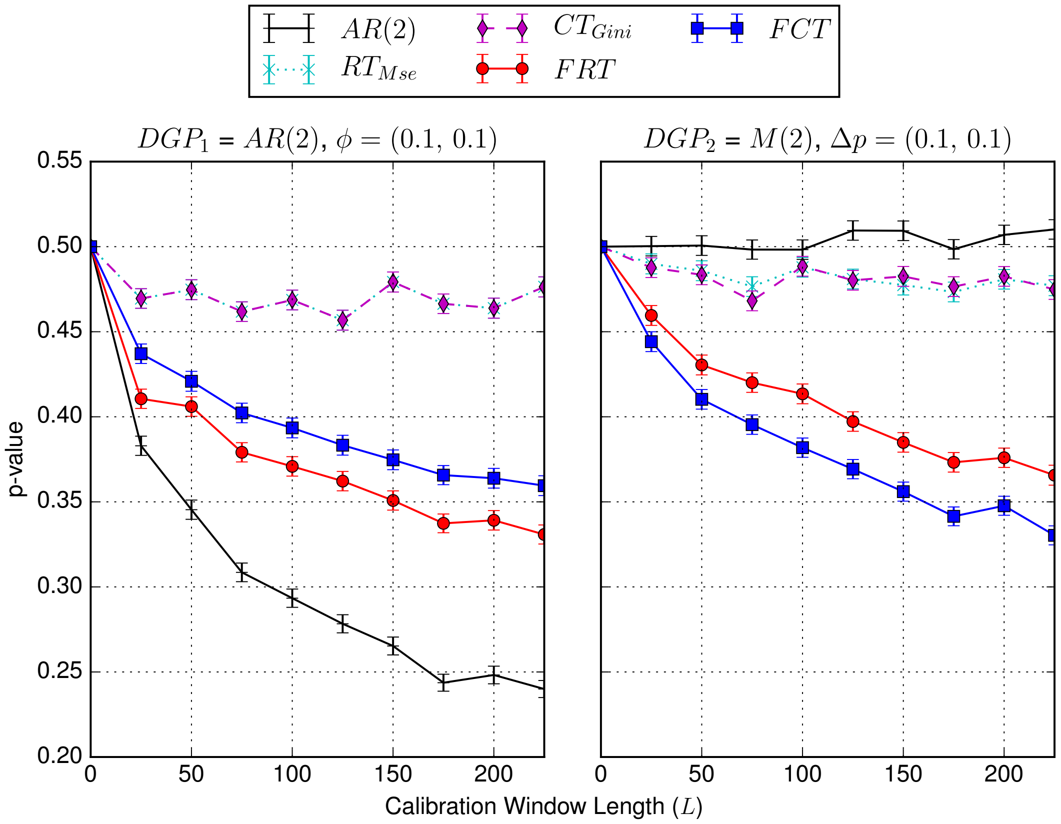


Figure 5.7: **P-values of all models for sample size $T = 500$ when predicting an autoregressive process AR(2) and binary Markov process M(2), with parameters** $(\phi_1 = 0.1, \phi_2 = 0.1)$**, respectively** $(\Delta p_1 = 0.1, \Delta p_2 = 0.1)$**.** The significance of the predictability increases with the calibration window length L as the estimation of the parameters becomes more robust.

the deterministic pattern in the binary Markov based process. The p-value converged to 0.5 for all tested parameters. In contrast, the tree based performance is almost identical on the autoregressive and binary Markov process. The regression tree performs slightly better then the classification tree for the autoregressive DGP, and vice versa for the binary Markov DGP. The dynamic trees significantly underperform the fixed trees, especially at small autocorrelations, as they overfit the noise.

5.3.6 Empirical results

5.3.6.1 Data & parameters

To demonstrate how decision tree based models can outperform autoregressive models on real data, the forecasting performance of all strategies is analyzed on daily returns of the S&P 500 during the 20 year time period Jan. 1, 1995 to Dec. 31, 2015. The daily returns are obtained from Thomson-Reuters Eikon with dividend adjustment. The studied universe of strategies arises from the models in Table 5.1, backtested for all calibration window lengths $L \in \{10,\, 20,\, \ldots,\, 500\}$ and lags $\varrho \in \{1,\, 2,\, 3,\, 4\}$.

Two strategies that only differ by their calibration window lengths L_1 and L_2 are highly correlated when $|L_2 - L_1| < 10$. A step size of $\Delta L = 10$ finds the optimal length with sufficient accuracy, while avoiding to compare almost identical strategies. As well, the bootstrap algorithm maintains the correlation structure, and therefore the multiple testing adjusted p-values are not impacted by the choice of ΔL. The upper bound of 500 trading days on L results from the fact that the best strategies were found well below this bound.

The limit $\varrho \leq 4$ on the number of lags ϱ arises from the condition $\frac{L}{2^{\varrho}} \geq 20$, which imposes the lower bound of 20 on the an average samples per leaf in a decision tree, or equivalently a $\approx 5\%$ accuracy on the class probabilities. At lag $\varrho = 5$, the class probabilities would be determined with a 6.5% accuracy, which is too high for a meaningful prediction.

To perform the bandwidth selection for the HAC covariance matrix estimation, as well as the bootstrap block size selection, the parametric $\mathrm{AR}(\varrho) - \mathrm{GARCH}\,(\varrho,\, \varrho)$ model is chosen. The model parameters are obtained by regression on the daily returns of the S&P 500 on the entire time period. The optimal parameters, determined by simulation, were found to have a kernel bandwidth $S^*_{5000} = 2.7$ and block size $b = 5$, roughly independent of the lag ϱ.

5.3.6.2 Performance of the top strategies

The individual p-values of all 1000 strategies are shown in Figure 5.8. The individual p-values are shown instead of the multiple testing adjusted p-values because in the later most p-values are simply one, and the relative performance is not visible anymore. The best performing predictor is the fixed classification tree (FCT) with lags $\varrho = 2$ and calibration length $L = 370$, which had a studentized Sharpe ratio performance of $\Delta_S^{1st} = 2.12$, noticeably higher then the second best predictor at $\Delta_S^{2nd} = 2.05$. This predictor was

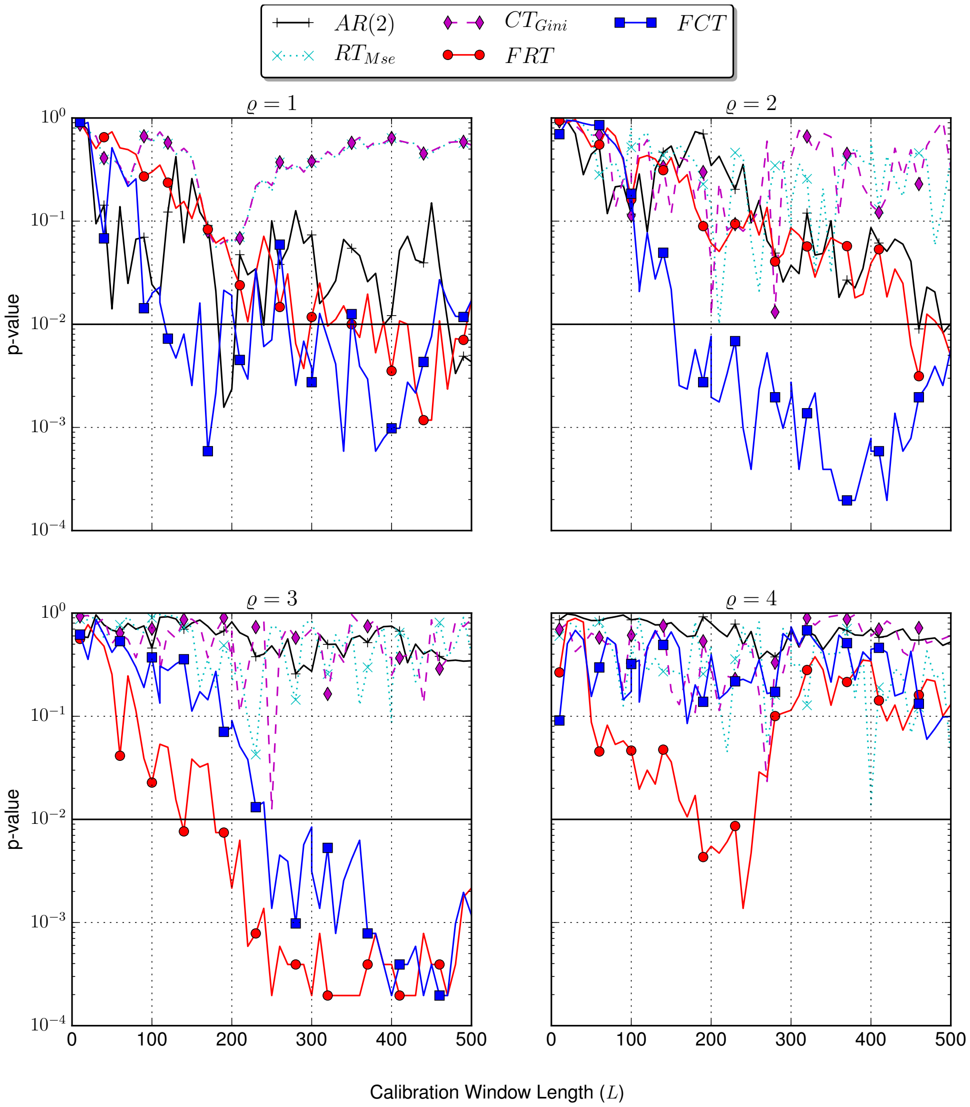


Figure 5.8: **Overview of the individual p-values of all strategies on the S&P 500 during the 20 year time period Jan. 1, 1995 to Dec. 31, 2015.** The five forecasting models are: autoregressive (AR); regression and classification tree (RT_{MSE}, CT_{Gini}); and the fixed regression and classification trees (FRT, FCT). Each forecasting model is run for the lags $\varrho = \{1, 2, 3, 4\}$ and calibration window lengths $L = \{10, 20, \ldots, 500\}$. The p-values are obtained by 5000 bootstrap simulations. The fixed trees perform above the 95% confidence level at lags $\{1, 2, 3\}$ and calibration window length $\varrho \geq 250$. The other models never perform significantly.

Model	ϱ	L	r_y(%)	Sh_y	**MD (%)**	p	p^{adj}	**BE (bps)**	**#RT**
B&H			5.15	0.49	-30.5			0	0
FCT	2	370	7.87	0.96	-12.2	$< 2.0 \cdot 10^{-4}$	0.004	16.4	1246
FCT	3	400	7.80	0.94	-10.1	$< 2.0 \cdot 10^{-4}$	0.031	10.6	1870
FCT	1	340	7.65	0.86	-10.6	$\leq 5.9 \cdot 10^{-4}$	0.031	21.4	860
FRT	3	330	7.86	0.92	-12.6	$< 2.0 \cdot 10^{-4}$	0.043	8.3	2417
AR	1	200	6.15	0.53	-23.4	0.002	0.750	2.2	2728
RT_{MSE}	2	210	5.75	0.55	-13.7	0.010	0.920	1.3	2653
CT_{Gini}	3	250	5.58	0.88	-15.4	0.012	0.953	0.9	2651

Table 5.2: **Summary of the top performing strategies.** The key values in order are: the model family; the number of lags ϱ; the calibration window length L; the compounded annual return r_y; the yearly Sharpe ratio Sh_y; the maximum draw down MD; the individual p-value; the p-value adjusted for multiple-testing in the entire universe of models; the break-even transaction costs (BE) with the buy-and-hold strategy; and the number of round trips #RT.

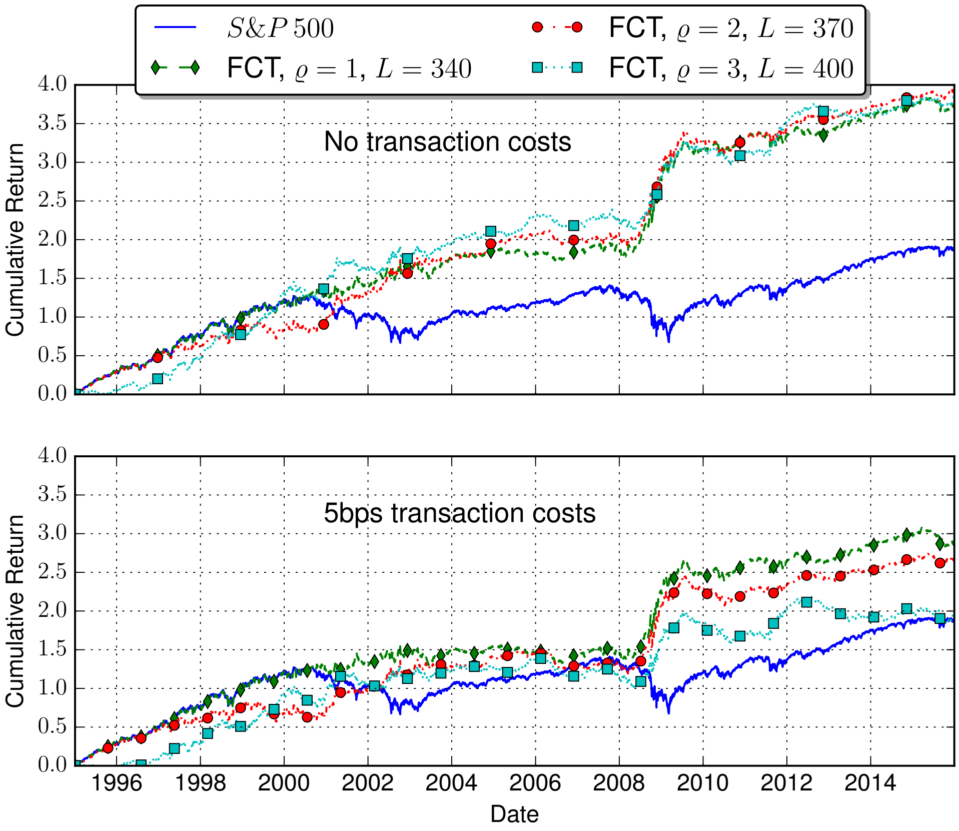


Figure 5.9: **Cumulative returns without (upper plot) and with (lower plot) transaction costs of the top performing model (*FCT*) at three different lags on the S&P 500 during the 20 year time period Jan. 1, 1995 to Dec. 31, 2015.** The highest predictability arises during the crash of the financial crisis in 2008 for all lags. Noticeable predictability arises as well during the implosion of the dotcom bubble from 2000 to 2003 at lag $p = 2$ and during the European debt crisis in late 2011 at lag $p = 3$. The lower plot shows the same strategies with 10bps of round trip transaction costs.

FCT	$\alpha\ (t_\alpha)$	$\beta_{MKT}\,(t_{MKT})$	$\beta_{SMB}\,(t_{SMB})$	$\beta_{HML}\,(t_{HML})$	$\beta_{MOM}\,(t_{MOM})$	R^2
$\varrho = 1$	1.06 (3.5)	0.31 (4.7)				0.08
$L = 340$	1.03 (3.4)	0.31 (4.5)	0.10 (1.1)	0.10 (1.0)		0.09
	0.90 (2.9)	0.38 (5.2)	0.08 (0.8)	0.16 (1.6)	0.16 (2.7)	0.11
$\varrho = 2$	1.22 (4.2)	0.20 (3.1)				0.04
$L = 370$	1.20 (4.1)	0.19 (2.9)	0.07 (0.8)	0.04 (0.4)		0.04
	1.21 (4.1)	0.19 (2.6)	0.08 (0.8)	0.03 (0.3)	-0.02 (-0.3)	0.04
$\varrho = 3$	1.19 (3.7)	0.20 (2.7)				0.03
$L = 400$	1.20 (3.7)	0.16 (2.2)	0.13 (1.2)	-0.07 (-0.7)		0.04
	1.17 (3.5)	0.18 (2.2)	0.12 (1.2)	-0.06 (-0.6)	0.04 (0.5)	0.04

Table 5.3: **Intercepts and slopes in variants of regression for the three top performing FCT strategies on the S&P 500**. The table shows the monthly intercepts (α) and regression slopes (β_{MKT}, β_{SMB}, β_{HML} and β_{MOM}, for $r_m - r^f$, SMB, HML, and MOM, respectively), as well as their t-statistics, for the CAPM, three-factor, and four-factor versions of regression. The factors are estimated for the three top performing FCTs on the S&P 500 between Jan. 1, 1995 and Dec. 31 2015 at zero transaction cost, as shown in Figure 5.9. The monthly intercepts are significant for all strategies and regression models. As well the market factor is significant in all cases, and particularly strongly at lag $\varrho = 1$. The FCT with lag one correlates significantly with momentum.

found to have a performance above the 99% confidence level when adjusting for multiple testing of all 1000 strategies. The fixed classification and regression trees are the only predictors that reach a performance above the 95% confidence level for a large range of lags and calibration lengths, when adjusted for multiple testing. This thick set of outperforming rules shows the presence of a robust signal, which has low probability of being a spurious phenomenon due to data-snooping. At lag $\varrho = 4$, the length L is to small for a robust calibration length and no predictor reaches significant results. This confirms the choice of tested lags. The results are in line with the simulation were the fixed trees where the most robust predictors, and longer calibration length provide a more robust parameter estimation.

The cumulative returns over time for the buy-and-hold strategy and the best fixed classification tree predictors are shown in Figure 5.9 without and with typical transaction costs. The one-way transaction costs of 0.05% applied at each buy or sell is supported by the discussion of Hsu et al. (2010, sec. 4.2). The strongest predictability arises during the burst of the dotcom bubble, the crash of the financial crisis, and the European debt crisis. Some performance metrics for the best strategy in each model class are shown in Table 5.2. The best strategy achieves twice the buy-and-hold return, twice the Sharpe ratio, and roughly a third of the maximum draw down. The break even costs of 21.4 bps per round trip are higher then the typical 10 bps of costs on actual markets .

5.3.6.3 Further analysis

The statistical significance of the best performing strategies does not imply that the EMH is violated, as it may have been impossible to select one of these strategies ex-ante. Following the EMH definition of Timmermann and Granger (2004), there needs to be a search technology that would have selected the winning strategy. The EMH is tested using the search technology that invests at every point in time into the strategy with the best Sharpe ratio after 10bps of round trip transaction costs. The search technology uses an expanding window over all past returns. As can be seen in the lower plot of Figure 5.9, the FCT strategy with lag $\varrho = 1$ and $L = 340$ starts to outperform the two next best strategies early on around 1997. The search technology confirms that no other strategy in the universe of 1000 strategies would have hindered the ex-ante generation of economic profits in excess of the buy-and-hold strategy. Consequently, the result does seem to violate the EMH, or at least raises questions about what limits to arbitrage could have prevented the implementation of such strategies, at least in the last two decades. The test period, which includes two major crashes, is unlikely to hide extreme events that could drastically change the downside risks of the strategy with respect to the market. However, an accurate simulation of transaction costs and potential market friction would have to be performed to confirm the result.

To determine if this abnormal performance can be explained by known factors, the performance is evaluated with the CAPM, the three-factor model of Fama and French (1993), and the four-factor model of Carhart (1997). The full four-factor model measures performance as a time-series regression of

$$r - r^f[t] = \alpha + \beta_{MKT}\left(r_m[t] - r^f[t]\right) + \beta_{SMB}SMB_t + \beta_{HML}HML_t + \beta_{MOM}MOM_t + e_t. \tag{5.26}$$

In this regression, r is the strategy return on month t, $r^f[t]$ is the risk-free rate (the 1-month U.S. Treasury bill rate), $r_m[t]$ is the market return, SMB_t and HML_t are the size and value-growth returns of Fama and French (1993), MOM_t is the momentum return, α is the average return not explained by the benchmark model, and e_t the residual error term. The values for $r^f[t]$, $r_m[t]$, SMB_t, HML_t and MOM_t are taken from Ken French's data library (French, 2012), and derive from underlying stock returns data from the Center for Research in Security Prices (CRSP). The three-factor model is obtained by leaving out the momentum term, and the CAPM is obtained by further leaving out the SMB and HML factors. Table 5.3 shows the intercept and regression slopes (load) for all three models, including their t-statistics. The returns of the three best strategies are partially explained by the market returns (significant β_{MKT}), but nevertheless all three have significant intercept α that remains unexplained by known factors. The return sign correlation uncovered by the decision tree strategies qualifies as a new anomalous factor.

To obtain a better understanding of the return dynamics that generate the predictability during the burst of the different bubbles, Figure 5.10 presents a zoom-in on this periods.

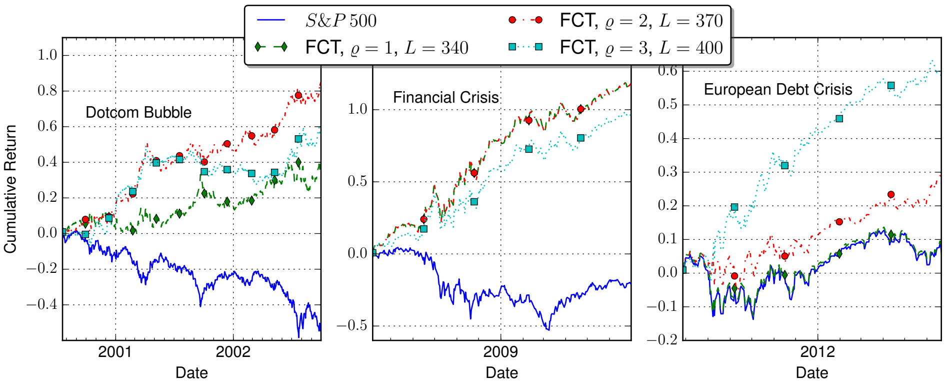


Figure 5.10: **Cumulative returns without transaction costs of the top performing model (FCT) at three different lags on the S&P 500 during the anomalous periods.** The left plot shows the anomalous performance of the fixed classification tree with lag $\varrho = 2$ and calibration length $L = 340$ during the burst of the dotcom bubble. The center plot shows the anomalous performance of the fixed classification tree with lag $\varrho = 1$ and calibration length $L = 370$ during the financial crisis. The right plot shows the anomalous performance of the fixed classification tree with lag $\varrho = 3$ and calibration length $L = 400$ during the European debt crisis.

Dotcom Bubble		Financial Crisis		European Debt Crisis	
$P\,(+) \approx 0.48$		$P\,(+) \approx 0.52$		$P\,(+) \approx 0.56$	
$P\,(-- \rightarrow +)$	≈ 0.54	$P\,(- \rightarrow +)$	≈ 0.61	$P\,(- - - \rightarrow +)$	≈ 0.56
$P\,(-+ \rightarrow -)$	≈ 0.53	$P\,(+ \rightarrow -)$	≈ 0.55	$P\,(- - + \rightarrow +)$	≈ 0.59
$P\,(+- \rightarrow -)$	≈ 0.54			$P\,(- + - \rightarrow -)$	≈ 0.50
$P\,(++ \rightarrow -)$	≈ 0.56			$P\,(+ - - \rightarrow -)$	≈ 0.56
				$P\,(- + + \rightarrow +)$	≈ 0.67
				$P\,(+ - + \rightarrow +)$	≈ 0.60
				$P\,(+ + - \rightarrow +)$	≈ 0.67
				$P\,(+ + + \rightarrow -)$	≈ 0.55

Table 5.4: **Daily return sign correlations of the S&P 500 during the burst of the dotcom bubble, financial crisis, and European debt crisis.** The sign correlations are computed based the whole time periods shown in Figure **5.10**, and do not account for potential non-stationarity. The burst of the dotcom bubble had more down moves then up moves, with significant two day directional accuracy. The crash of the financial crisis is impregnated by a daily reversal of the return sign, with up moves being 11% more likely then down moves after a down move. The European debt crisis exhibits a more intricate three day sign correlation, with an up move being 10 to 17% more likely then a down move after two up moves and one down move in arbitrary order.

During each period, the best performing strategy is significant at the 99.9% level when performing a bootstrap of returns under the stationarity condition (see Section 3.4.2). The burst of the dotcom bubble is dominated by a predictable return sign pattern at lag two (break-even cost of 45 bps per round trip). The crash of the financial crisis is dominated by a one day return sign dependency (break-even cost of 83 bps per round trip). Finally, the European debt crisis exhibits a more intricate three lag pattern that cannot be captured by lower lag strategies (break-even cost of 32 bps per round trip). The abnormal transition probabilities of these patterns are presented in Table 5.4. These stationary snapshots over the whole duration of each bubble confirm the significant directional accuracy.

Applying the theoretical model developed in Section 5.3.2 and 5.3.3, to the stationary snapshots of the three periods of abnormal returns, confirms the fundamental concept that autoregressive models poorly capture return sign correlations. The expected autoregressive parameters computed with Equation (5.22) based on the values of Table 5.4 are

$$\begin{aligned} E_{\text{Dotcom}}\left[\phi\right] &= \left(0.006,\ -0.013\right), \\ E_{\text{Financial Crisis}}\left[\phi\right] &= \left(-0.027\right), \text{ and} \\ E_{\text{European Debt Crisis}}\left[\phi\right] &= \left(-0.000,\ 0.008,\ 0.013\right). \end{aligned}$$

The expected parameters ϕ have been scaled so as to be comparable with the directional accuracies presented in Table 5.4. The values show that the autoregressive model captures at most 2.7% of excess directional accuracy, far smaller then the maximal 17% of directional accuracy capture by the fixed classification tree. Hence, the theoretical model explains why the autoregressive underperforms the fixed trees on the S&P 500 as shown by Figure 5.8.

5.3.6.4 Conclusion

The EMH is an assumption about financial market at the heart of many regulatory decisions. This hypothesis has been verified to hold true for a large range of regressive forecasting models, technical trading rules and asset portfolios. However, recent developments in statistical learning have not yet undergone a rigorous test.

In this paper, we presented a common non linearly separable pattern that can arise, but cannot be forecast using autoregressive models. Decision tree models possess arbitrary flexibility and are well suited to capture these non linearly separable patterns. The issue of overfitting can be addressed with an adequate lower bound on the number of samples per leaf in a decision tree. We provide a connection between fixed decision trees and Markov chains. The presented class of binary Markov processes with a deterministic component are proven to be unpredictable with autoregressive models. In contrast, the fixed classification trees only marginally underperform an autoregressive forecast on an autoregressive DGP for most parameter choices. A simulation study confirmed the theoretical results and the robustness of fixed classification and regression trees.

The models are tested on daily returns of the S&P 500 for different lags and calibration window lengths, giving rise to a universe of 1000 strategies. The multiple testing adjusted p-value of each strategy, benchmarked against the buy-and-hold strategy, is computed using the methodology of Romano and Wolf (2005b) and Ledoit and Wolf (2008). The fixed classification tree (FCT) at lags $\varrho \in \{1, 2, 3\}$ and calibration window length $L \in [300, 460]$ are the best strategies, significant above the 95% confidence level. This confirms the simulation results where the fixed trees were robust predictors for autoregressive and Markov based processes. The analysis showed that the theoretical model holds true on the S&P 500 to explain the performance difference between decision trees and autoregressive strategies.

Without transaction costs, the best fixed tree strategies more then double the cumulative return and Sharpe ratio of the buy-and-hold strategy. They break even with the buy-and-hold strategy at transaction costs as high as 21 bps per round trip. A simple best Sharpe ratio search technology could have selected ex-ante the best performing strategy. The strategies all have significant intercept for the four-factor regression model. Therefore, the EMH appears to be violated, in particular during the dotcom bubble (2000-2003), the financial crisis (2008-2009) and the European debt crisis (2012). During bull markets, the performance of decision tree based strategies is roughly equal to the buy-and-hold strategy. While no certain explanation can be given, it would not be surprising to find that behavioral heuristics, as exposed by Tversky and Kahneman (1974) even among trained statisticians, have significant impact during a market crash. The strong return sign correlations speak in favor of biases in human heuristics relying more on past return signs then on the amplitude.

The finding of this study stands in contrast with multiple prior market efficiency studies that included the S&P 500. The studies by Sullivan et al. (1999), Hsu and Kuan (2005) found no significantly performing technical trading rule on the S&P 500, while our fixed classification trees perform significantly on a longer test period of 20 years. The study by Hsu et al. (2010), as well testing technical trading rules, further concluded that market efficiency has increased after the introduction of ETFs in the year 2000, and that emerging markets are less efficient then more mature markets. Our study provides a solid counter example of a mature market that exhibits large inefficiencies. As well, the largest inefficiencies appear long after the introduction of ETFs, casting doubt on the impact of ETFs on market efficiency. The discrepancy with prior studies is a result of their focus on technical trading rules. Our work shows that market efficiency cannot be measured only using technical trading rules.

Noticeable research has gone into detecting explosive regimes in stock markets (Phillips et al., 2011, Kaizoji et al., 2015, Sornette and Cauwels, 2015), and studying the growth phase of bubbles. In contrast, the return dynamics during the burst of a bubble seem under researched as shown by the recent findings of an acceleration factor (Ardila et al., 2015) and the novel return sign predictability established in this paper. Future research needs to study more in detail the predictability dynamics during bubbles on other stock indices. As well, it needs to be better understood which stocks in an equity drive the predictability

5.4 Towards time series learning

The binary strategies used by the agents in the agent based model happened to be equivalent to a decision tree. The decision trees were shown to capture sign correlations that would go unnoticed with autoregressive models. It turned out that such correlations were statistically significant in the S&P 500 during the past 20 years. Tree based trading strategies could have exploited these sign correlations in a profitable manner after adjusting for transaction costs. Given this success of decision trees in finding non-linearly separable patterns in the returns of the S&P 500, there are two follow up questions: are these return sign correlations present in other assets as well? and are there more non-linear patterns that could be found with other models?

The forecasting methodology used for the decision tree and autoregressive models, always predicting one step ahead using a rolling window of size L, is straightforward to extend for other time series or statistical learning models. The research by Fernández-Delgado et al. (2014) showed that a small number of statistical learning models provide the highest performance across a large range of data-sets. This section implements an evaluation of a larger universe of strategies constructed from these top performing statistical learning models. Each models gives rise to several strategies, which are parametrized by the number of lags and in-sample length used for calibration. Additionally, a AR-GARCH model is used as a time series reference model. This led to a total of 3136 uniquely defined strategies.

The strategies are evaluated on daily returns of the CSI 300 during the period 2005-2015, and the FTSE and S&P 500 during the period 1995-2015. This choice of three major equity indices, well separated geographically (and too some extend economically), provides a good test set to determine if non-linear patterns are common in equity indices.

5.4.1 Statistical learning & nonlinear finite lag processes

A general nonlinear stochastic process $\{X_t\}$ is described by some arbitrary function $f(\cdot)$ mapping uncorrelated random variables $\{a_t\}$ as

$$X_t - \mu = f(a_t, a_{t-1}, a_{t-2}, \ldots). \tag{5.27}$$

In practice, this general representation is of limited use because the curse of dimensionality makes it difficult or even impossible to determine f based on past data. To calibrate f on past data, the common approach is to chose a specific parametric form for f, with a few parameters that can be determined from the data. For example, the linear filter models are obtained by a Taylor expansion at first order as

$$X_t - \mu = f\left(0,\, 0,\, \ldots\right) + a_t \cdot \partial_{x_1} f\left(0,\, x_2,\, \ldots\right) + a_{t-1} \cdot \partial_{x_2} f\left(x_1,\, 0,\, \ldots\right) + \ldots, \tag{5.28}$$

where the first order partial derivatives of f evaluated at zero are related to the coefficients of the autoregressive process. Common nonlinear models can be obtained by considering the second order terms.

Besides polynomial processes obtained by Taylor expansion, one can as well consider non-parametric nonlinear autoregressive models of order ϱ of the form

$$Y_t = f\left(Y_{t-1},\, \ldots,\, Y_{t-\varrho}\right) + \epsilon_t. \tag{5.29}$$

For a discussion, see Chapter 6.3 in Mills and Markellos (2008). With the advent of cheap computation, a large number so called statistical learning (or machine learning in computer sciences) models have been developed, which solve special cases of finding f in a non-parametric setting. These methods are discussed in detail by Hastie et al. (2001) and James et al. (2014).

In the context of statistical learning, the problem is to estimate the systematic relation f between a set of inputs X and dependent outputs Y, such that

$$Y = f(X) + \epsilon. \tag{5.30}$$

The error term ϵ has mean zero, and is independent of X. To illustrate, the output could be the wage of a person, and the inputs are some characteristics of that person, such as age and years of education. The difference to the nonlinear autoregressive models is that the input and output variables a separate entities.

A statistical learning algorithm learns a non-parametric function f based on a training dataset

$$\mathcal{D} = \left\{\left(X_1,\, Y_1\right),\, \ldots,\, \left(X_N,\, Y_N\right)\right\} \tag{5.31}$$

with a sufficient number of samples. In the wage example, a sample would be a data tuple $\left(X = \left(age,\, education\right),\, Y = \left(wage\right)\right)$ associated to a person. The set of statistical learning models include parametric and non-parametric methods, ranging from low flexibility (e.g. linear regression) to high flexibility (e.g. support vector machine with non-linear kernels). The challenge in the estimation of f for flexible models is to avoid over-fitting the error term associated with the dataset $\mathcal{D}$. The method has to only extract the systematic information, so that the estimated function works on any dataset $\mathcal{D}'$ with new samples of the same type.

5.4.1.1 Issues in cross-validating time series data

The statistical learning models all find some non-parametric approximation of the data used for calibration. However, the intrinsic signal to noise ratio in financial time series is small, which makes it difficult to distinguish between truly predictable patterns and randomly occurring transient patterns. The standard technique to avoid over-fitting is to split the dataset into a training and a test set. The calibration is then performed in multiple steps. At each step the calibration on the training set is refined and the performance on the test set is computed. When the performance on the test set reaches a plateau, the calibration is considered to be optimal. More extensive cross-validation can be performed by generating multiple (*training*, *test*) sets based on the dataset, and repeating the calibration process on each of them.

A core assumption of the cross-validation techniques is the sample independence, which does not hold in the time series context. For a discussion, see Chapter 6.3.3 in Mills and Markellos (2008). In nonlinear autoregressive models the samples are sequential in time, which makes it difficult to justify cross-validation methods that consider the samples as independent. The only test set that respects the time order uses the last samples in the time series data used for calibration. However, assuming some form of nonlinear autocorrelation, the last samples have the highest correlation with the predicted next step output and should therefore be used for calibration. Consequently, any type of test set or cross-validation is likely to affect negatively the predictive power.

5.4.1.2 Calibration window length and lag selection

Most statistical learning models split the input space into multiple regions similarly to the decision trees in Subsection (5.2.2) and predict the mean or majority vote in each region. The discussion for classification trees in Subsection 5.2.4 showed that at least 100 samples are needed in each region to obtain a 95% confidence level for a binary prediction, assuming independent random samples as null hypothesis. As this number of samples does not dependent on the shape of the region, this lower bound is meaningful for most statistical learning models and is applied throughout this thesis.

Although the sample size should be maximized to avoid over-fitting, it may not be optimal to keep all past data because financial time series are usually not strictly stationary (Mills and Markellos, 2008). Non-stationarity negatively impacts predictive performance when the calibration is done across multiple distinct regimes. To minimize the mixing of different regimes, the number of samples used for calibration is limited to a constant in-sample calibration length L. For the decision tree prediction experiment on the S&P 500 in Subsection 5.3.6 the best results where obtained well below the in-sample length of two years ($L \leq 500$ trading days). As this length characterizes the typical duration of a stationary regime in an equity index, and is independent of the prediction model, this bound is used for all statistical learning models.

The curse of dimensionality strongly limits the number of lags that can be calibrated robustly. Similarly to the decision trees, one can argue that a meaningful prediction model should distinguish at least two distinct regions in each lag dimension, otherwise the lag variable has no influence on the prediction. Two regions correspond to the best case scenario of binary input and output variables. Hence, in the best case scenario, the complexity of a prediction model increases as 2^{ϱ}. The limit $\varrho \leq 4$ on the number of lags ϱ follows from the condition $\frac{L}{2^{\varrho}} \geq 20$, which imposes the lower bound of 20 on the an average samples per region, or equivalently a $\approx 5\%$ accuracy on the class probabilities. At lag $\varrho = 5$, the class probabilities would be determined with a 6.5% accuracy, which is too high for a meaningful prediction.

5.4.1.3 Regression vs classification

Asset returns are quantitative outputs that are typically modeled using regression. The degrees of freedom in regression models have a variety of origins: the parameters of a function, for example the coefficients of a linear regression; the parameters of a kernel transforms of the inputs, for example the exponent in a homogeneous polynomial kernel; and the number of partitions of the input space, for example the leaves in a regression tree. Models with too few degrees of freedom will have high bias, failing to model relevant relations between inputs and outputs. Models with too many degrees of freedom will have high variance, modeling the random noise in the training data. In the context of financial time series, the signal to noise ratio is small and therefore the optimal models should have a low number of degrees of freedom and maximize the signal to noise ratio.

Subsection 5.2.4 argued that the signal to noise ratio of financial data can be improved by mapping the returns to a finite number of classes. The classification problem of maximizing the probability of a single class in each region is better suited to maximize predictability than the mean squared error loss typically used in regression models. As well, the statistical significance of a prediction is easier to estimate.

From a behavioral perspective of the stock market participants, one could argue that the traders do not care about the precise value of returns. Fundamentalists trade based on their fundamental analysis of companies, and chartists mostly look at trends and a finite set of patterns. Consequently, it is meaningful to map returns onto a set of classes that relate closely to the discrete mental models used by the agents. Tversky and Kahneman (1974) showed extensively that humans, even when trained in statistics, often rely on heuristics to answer statistical questions. Further on, Gary and Wood (2010) studied human mental models in solving strategic problems, and reported that good heuristics often suffice to achieve good performance. Within this heuristic perspective, the agents trading on stock markets make the strongest distinction between negative and positive returns. Therefore, the most sensible preprocessing is a binary map of the outputs to down and up moves. Such binary outputs can be predicted using classifiers. In a second step, the same binary preprocessing can as well be applied to the inputs.

Beyond the binary map of up and down moves, a infinite number of generalized binary maps and high order maps are possible as discussed in Subsection 4.2.9. However, testing all these maps individually defeats the purpose of statistical learning. Models such as decision trees are able to learn the optimal map dynamically with the adequate pruning parameter.

5.4.2 Constructing the universe of strategies

There are hundreds of statistical learning models available in the literature (Fernández-Delgado et al., 2014). In James et al. (2014), several of the most common methods have been tested on returns of the S&P 500 between the year 2001 and 2005, with the highest performing method being the quadratic discriminant analysis. The present chapter evaluates the set of models presented in Table 5.5, covering the families of linear regression (AR-GARCH), discriminant analysis, logistic regression, nearest neighbors, support vector machines, decision trees, and bagging and boosting of decision trees. These are the most commonly used methods and known to be among the top performers across a large variety of problems. Neural networks have been excluded, as they are computationally much more intensive and difficult to interpret.

Classification models (and most regression models) can be visualized as separating the samples with decision boundaries maximizing the predictability within each group formed by the boundaries. The difference between the methods mostly lies in the shape of the decision boundaries. For the reader unfamiliar with statistical learning, some visual examples are provided in Figure 5.11. These figures represent a two dimensional input and a three class categorical output (e.g. negative returns, small returns and positive returns). Subfigure a) shows the optimal decision boundary from a linear discriminant analysis separating three multivariate normal distributions. Subfigure b) shows some optimal hyperplane boundaries used in support vector machines. Finally, subfigure c) represent the boundaries of a decision tree obtained by recursively splitting the dataset.

The selected regression models are applied to the quantitative inputs and outputs. The classification methods are applied to quantitative inputs but binary outputs of down and up moves. The case of binary inputs organizes the samples neatly into ϱ dimensional hypercubes as illustrated in subfigure d). In such a configuration, the shape of the decision boundaries becomes mostly irrelevant as many models will make the same prediction. An input space partitioned into hypercubes can be exactly matched by the decision boundary of decision tree based model. Therefore, binary inputs are predicted using a decision tree. The prediction of a decision tree is the majority vote (classification) or mean (regression) within each hypercube. Table 5.6 provides an overview of all the 16 possible combinations of using regression and classification. Given the four different lags $\varrho \in \{2, 3, 4, 5\}$ and the in-sample length $L \in \{10, \ldots, 500\}$ in steps of $\Delta L = 10$, the universe M has $4 \times 49 \times 16 = 3136$ uniquely defined strategies.

As for the decision trees in Subsection 5.3.5, a model prediction is simply converted to a long or short trading signal. A positive regression forecast is converted to a long position and a negative regression forecast to a short position. Down and up class predictions of a classifier are straightforwardly converted to a short position, respectively a long position. The strategies are always long or short, and never out of the market. However, the

Model	Description	Parameters
Regression		
AR(p)-GARCH(p, q)	Linear regression model, with a nested linear regression model for the variance of the error terms.	Lags $p = q = m$
Classifiers		
Linear **D**iscriminant	Assumes **k** classes from independent multivariate distributions. Computes the optimal linear boundary between the distributions.	No shrinkage
Quadratic **D**iscriminant	Assumes **k** classes from independent multivariate distributions. Computes the optimal quadratic boundary between the distributions.	
Logistic **R**egression	Computes the logistic decision boundary between two classes.	
Regression/Classification		
Nearest **N**eighbors	Finds the **k** nearest neighbors of a data point and computes their mean output or majority vote.	$k = 5$, Minkowski metric, uniformly weighted
Support **V**ector **M**achine	Computes the maximum-margin separating hyperplane.	Linear kernel
Decision **T**ree	Partitions the features space by recursive splitting at optimal values along a single feature. Computes the mean or majority vote in each subset.	Any depth, minimum one sample per leaf
Random **F**orest	Averages the prediction of many decision trees trained on subsamples of the training data to avoid over-fitting.	10 trees
Gradient **B**oosting	Recursively creates decision trees to reduce the remaining errors.	Max depth 3, 100 boosts, learning rate 0.1

Table 5.5: **Description and relevant parameters of the statistical learning methods evaluated in the present work to forecast daily returns.** The table separates regression only, classification only (e.g. down/up moves), and dual purpose methods.

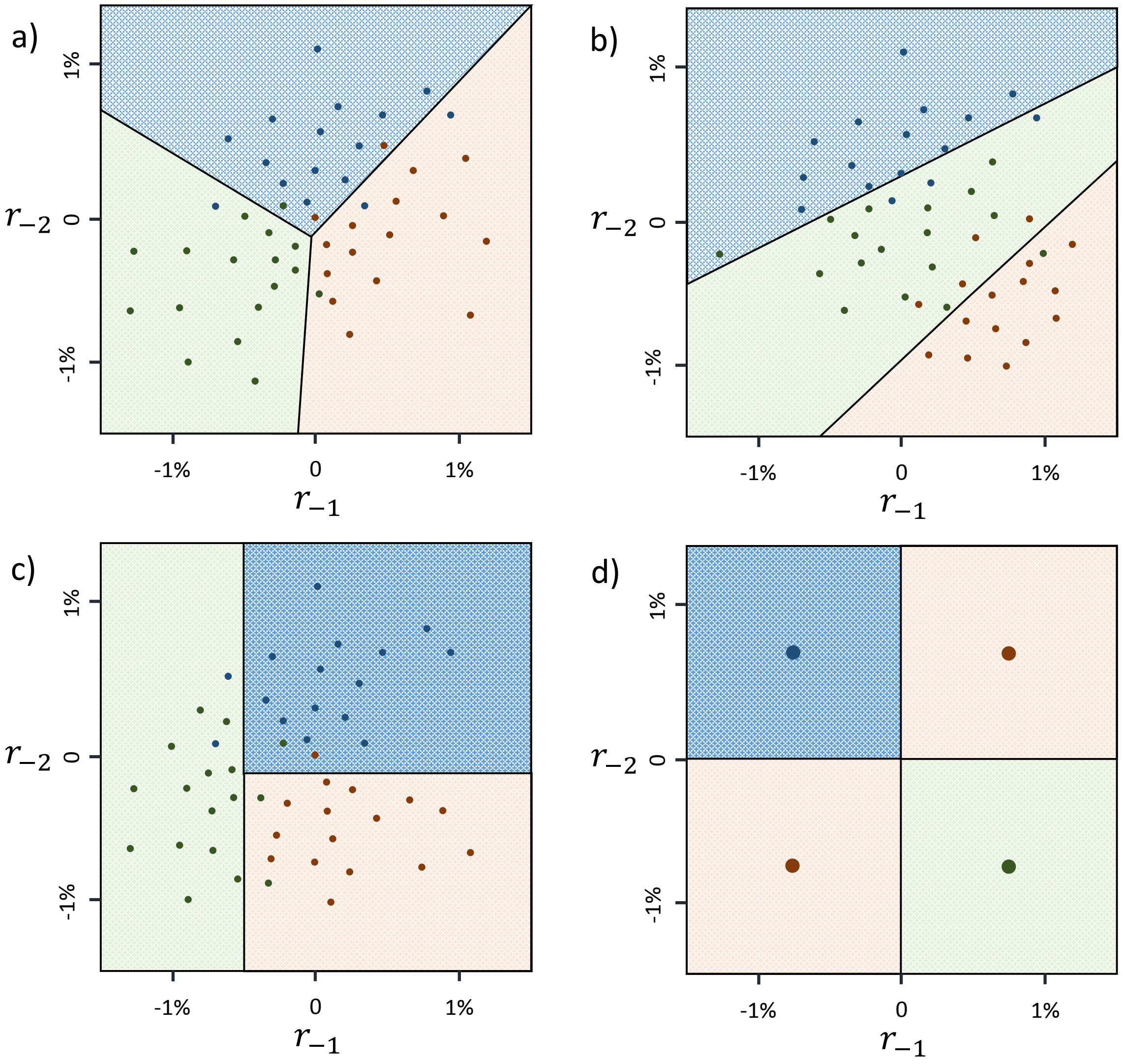


Figure 5.11: **Schematic examples of decision boundaries for several statistical learning methods fitted on the inputs (r_{-2}, r_{-1}) of two past quantitative returns and three possible output classes.** The methods are: a) linear discriminant analysis; b) support vector machine; c) decision tree; d) decision tree fitted on binary returns (the choice of having twice the red class as the majority vote is arbitrary).

Inputs	**Outputs**	**Methods**
Quantitative	Quantitative	6 Regressors
Quantitative	Binary	8 Classifiers
Binary	Quantitative	1 Regression Tree
Binary	Binary	1 Decision Tree
	Total:	16

Table 5.6: **Summary of all the tested combinations of quantitative and binary inputs and outputs.** The number of available classifiers and regressors is taken from the selection presented in Table 5.5.

amplitude of a regression forecast or the probability of the predicted class could be used to construct more fine grained trading signals. For example taking larger positions upon certain forecasts, and staying out of the market upon uncertain forecasts. Nonetheless, such an exploration of different trading signals is not at the heart of this thesis that will only evaluate strategies based on unit long and short positions.

5.4.2.1 Models discussion

The AR-GARCH model predicts the next step return based on the autocorrelations at different lags and volatility of the past returns. However, in the context of highly stochastic financial data, the regression most of the time captures the current market trend, and the resulting strategies are highly correlated with the market and momentum.

The linear and quadratic discriminant analysis can determine if the ϱ-variate distribution of returns followed by a down move significantly differs from the ϱ-variate distribution of returns followed by an up move. Whenever the two distributions are distinguishable, the prediction probability will rise above randomness. The largest difference between the two distributions is located in the tails, where the predictions with the highest probabilities will be made. The tails of the distributions are associated to large returns, consequently the discriminant methods could detect predictability linked to high volatility.

The logistic regression detects directions in the input space that are dominantly followed by a down or up move. Due to the approximate normal distribution of returns around zero, the reference point will be found near zero. Predictions with high probability will be made following large returns, consequently the logistic regression can detect predictability linked to high volatility in a certain direction in the input space.

The nearest neighbor model is evaluated for $k = 5$, computing for a given input the mean or majority vote of the five nearest samples (using an euclidean metric) in the training set. In the classification case, the odd number of neighbors ensures an unambiguous prediction. The choice of only five neighbors guarantees the estimation of a neighborhood

Index	Start	End	W_y(%)	SR_y	↑(%)	r_m (%)	r_c (%)
CSI 300	Apr. 8, 2005	Dec. 31, 2015	13.12	0.45	54.1	$5.60 \cdot 10^{-3}$	1.34
FTSE	Jan. 1, 1995	Dec. 31, 2015	3.45	0.11	52.2	$0.88 \cdot 10^{-3}$	0.82
S&P 500	Jan. 1, 1995	Dec. 31, 2015	7.37	0.31	53.7	$2.64 \cdot 10^{-3}$	0.82

Table 5.7: **Summary statistics of the equity indices used for the empirical study.** This table provides a summary statistics of the equity indices used to test the statistical learning methods on daily returns. The key values in order are: the compounded annual growth rate W_y; the yearly Sharpe ratio SR_y; the number of positive days; the average daily market return r_m due to the market trend as defined in Equation 3.30; and the average daily return r_c on a winning trade as defined in Equation 3.32.

close to the evaluated input even in the case of $\varrho = 5$ lags. This model can detect pockets of predictability localized in a small neighborhood of the input space.

The support vector machine (SVM) model separates the features space by a hyperplane, maximizing the class asymmetry between the two areas defined by the hyperplane. This method is ideal to detect any imbalance of the distribution of up or down moves in the input space. On the downside, the standard SVM classifier does not support the computation of a prediction probability.

The decision tree model partitions the input space into disjoint n-orthotopes (hyperrectangles), maximizing the proportion of a single class in each of them. As discussed in Subsection 5.2, without a well selected pruning parameter, decision trees are prone to overfitting, which can be addressed by bagging and boosting models. The bagging model, called a random forest, averages multiple trees calibrated on subsets of the training data to reduce the variance. The boosting method recursively reduces the error by adding new trees calibrated on the remaining error. To avoid a perfect fit from the beginning, the tree depth is limited to the number of lags in the boosting case.

5.4.3 Empirical results

The universe of strategies is tested on daily returns of equity indices in three geographical areas: Asia, Europe and U.S. The selected stock indices are the CSI 300 Index in Shanghai, the FTSE in London and the S&P 500 in the U.S. The daily returns are loaded from the Thomson Reuters data stream, while the monthly risk free rates are taken from the Ken French's data library French (2012). An overview with relevant key values is provided in Table 5.7. The starting date of the CSI 300 correspond to the first ever trading day of that index.

5.4.3.1 Best models

Table 5.8 presents for each model family the statistical significance of the best model. The performance measures are the directional accuracy, wealth (i.e. compounded returns), and Sharpe ratio as defined in Section 3.3. The test statistics are computed using a one-sided

	S&P 500				**FTSE**				**CSI 300**			
Family	$p^{\min}_{\chi^2_d}$	$p^{\min}_{W}$	$p^{\min}_{SR}$	$\#^{95}_{W}$	$p^{\min}_{\chi^2_d}$	$p^{\min}_{W}$	$p^{\min}_{SR}$	$\#^{95}_{W}$	$p^{\min}_{\chi^2_d}$	$p^{\min}_{W}$	$p^{\min}_{SR}$	$\#^{95}_{W}$
GARCH	$\mathbf{6.4}\cdot 10^{-3}$	0.24			0.19	0.67	0.68		$\mathbf{1.3}\cdot 10^{-4}$	0.05		
NN-R	0.84	0.25			0.10	$\mathbf{9.1}\cdot 10^{-3}$		**163**	$\mathbf{2.0}\cdot 10^{-2}$	0.35		
SVM-R	0.11	0.25			$\mathbf{2.2}\cdot 10^{-3}$	$\mathbf{2.2}\cdot 10^{-3}$		**187**	$\mathbf{6.4}\cdot 10^{-4}$	0.35		
DT-R	0.78	0.48			$\mathbf{3.9}\cdot 10^{-2}$	$\mathbf{3.6}\cdot 10^{-2}$		**1**	0.22	$\mathbf{2.1}\cdot 10^{-2}$		**1**
RF-R	0.47	0.55			0.67	0.11			0.43	0.15		
GB-R	0.22	0.20			0.33	0.07			0.16	$\mathbf{1.0}\cdot 10^{-2}$		**1**
LDA-C	0.14	0.26	0.23		$\mathbf{3.3}\cdot 10^{-2}$	$\mathbf{3.7}\cdot 10^{-3}$	$\mathbf{3.4}\cdot 10^{-3}$	**186**	$\mathbf{3.0}\cdot 10^{-2}$	0.56	0.57	
QDA-C	0.41	0.66	0.65		$\mathbf{3.6}\cdot 10^{-2}$	0.06	0.05		$\mathbf{4.4}\cdot 10^{-3}$	0.17	0.18	
LR-C	$\mathbf{4.4}\cdot 10^{-3}$	0.37	0.28		0.19	0.68	0.67		$\mathbf{1.1}\cdot 10^{-4}$	$\mathbf{4.7}\cdot 10^{-2}$	$\mathbf{6.8}\cdot 10^{-2}$	**1**
NN-C	0.78	0.98			0.84	0.10			$\mathbf{9.0}\cdot 10^{-6}$	$\mathbf{2.0}\cdot 10^{-2}$	$\mathbf{1.9}\cdot 10^{-2}$	**150**
SVM-C	$\mathbf{8.7}\cdot 10^{-4}$	0.04	0.07		0.22	0.25	0.36		$\mathbf{3.0}\cdot 10^{-4}$	0.07		
DT-C	0.80	0.78	0.77		$\mathbf{7.3}\cdot 10^{-3}$	0.24			0.86	0.44		
RF-C	0.80	0.89	0.87		0.65	0.28			0.07	0.18	0.16	
GB-C	0.73	0.92	0.91		0.58	0.05			0.11	0.21		
DT-BR	$\mathbf{2.9}\cdot 10^{-4}$	$\mathbf{3.6}\cdot 10^{-3}$	$\mathbf{5.6}\cdot 10^{-3}$	**166**	$\mathbf{7.5}\cdot 10^{-3}$	$\mathbf{2.8}\cdot 10^{-3}$		**196**	$\mathbf{3.5}\cdot 10^{-3}$	0.45		
DT-BC	$\mathbf{7.0}\cdot 10^{-6}$	$\mathbf{6.4}\cdot 10^{-4}$	$\mathbf{9.3}\cdot 10^{-4}$	**119**	$\mathbf{1.6}\cdot 10^{-2}$	$\mathbf{5.9}\cdot 10^{-3}$	$\mathbf{5.5}\cdot 10^{-3}$	**196**	0.14	0.12		

Table 5.8: **Summary statistics by model family: out-of-sample S&P 500, FTSE and CSI 300.** This table provides for each model family the three best p-values for the directional accuracy ($p_{\chi^2_d}$), the compounded wealth (p_W), and the Sharpe ratio (p_{SR}). The Sharpe ratio p-value is only indicated if $p_{SR} \neq p_W$. The model families are **N**earest **N**eighbors (NN), **S**upport **V**ector **M**achine (SVM), **D**ecision **T**ree (DT), **R**andom **F**orest (RF), **G**radient **B**oosting (GB), **L**inear **D**iscriminant **A**nalysis (LDA), **Q**uadratic **D**iscriminant **A**nalysis (QDA), and **L**ogistic **R**egression (LR). The family names are suffixed as follows: "-R" for regressors; "-C" for classifiers; and "-B" for binary. The p-values are adjusted for multiple testing within each family according to the algorithm in Section 3.4. Each family has $49 \times 4 = 196$ models, parametrized by 49 in-sample lengths $L \in [20, 500]$ and 4 lags $\varrho \in [2, 3, 4, 5]$. The value $\#^{95}_{W}$ indicates the number of models significant at the 95% level in compounded wealth.

test for excess performance. The p-values have been adjusted for multiple testing within each family using the algorithm of Section 3.4, and indicate if a model family is significant when considered in isolation.

The p-values p_W of the compounded wealth and p_{SR} of the Sharpe ratio are mostly identical. This finding is not surprising as both performance measures depend primarily on the mean daily return. Exceptions to the identical wealth and Sharpe ratio significance are found among the top performing models. For example, for the best DT-B model on the S&P 500 the large upside volatility, seen on Figure 5.12, penalizes the Sharpe ratio performance. On the contrary, for the best LDA-C model on the FTSE, staying out of the market during high volatility days improves the Sharpe ratio performance.

This study focuses on models with statistically significant wealth, the measure of profits used to test the EMH. Some model families stand out as significant, far above the 95% confidence level. On the S&P 500, the decision tree models with binary inputs (DT-BR and DT-B) are highly significant. On the FTSE, the best model is the support vector regression (SVM-R), second is the decision tree regressor with binary inputs (DT-BR), and third comes the linear discriminant analysis (LDA-C). Further significant models are the nearest neighbor regression (NN-R) and binary decision trees (DT-BC). The single case above the 95% level for the decision regressor model family (DT-R) could be a statistical outlier. On the CSI 300, the nearest neighbor classifier model family (NN-C) is consistently significant. However, it is significantly outperformed by a single GB-R model.

An overview of the models with highest significance on p_W is presented in Table 5.9. The p-values adjusted for multiple testing across the entire universe remain significant for the FTSE and S&P 500. The best model for the CSI 300 is significant at the 94% level after only 8 years of trading, and would likely be highly significant on an equivalent 18 year period if its performance remained at the same level. The best performing model for each equity index is consistently found at lag $\varrho = 3$, while the optimal in-sample length is found in two different regions. For the S&P 500 and the FTSE, the optimal in-sample length clusters around $L \in [390, 400]$, and for the CSI 300 the optimal in-sample length cluster around $L \in [150, 160]$. For the equity indices tested in this study, these combinations of lag and in-sample length provide the best tradeoff in maximizing sample size and minimizing the risk of calibrating across multiple regimes.

The trading performance over time of a selection of the best models presented in Table 5.9 is shown in Figure 5.12. For the FTSE, the top performing DT-BC and LDA-C models are not shown as they correlate highly with the shown models DT-BR, respectively SVM-R. This Figure reveals that the highest abnormal returns on the S&P 500 and FTSE occur during the dot-com bubble (1997-2003), the financial crisis (2008-2009), and the European debt crisis (2012). The GB-R and NN-C models on the CSI 300 have very different dynamics. The abnormal returns of the GB-R model correlate highly with the financial crisis (2008), and the recent Chinese stock market turbulences (2015). The GB-R model is a variant of boosted decision trees, and this result therefore strengthens the finding that decision tree based models have significant predictability during crises. The

Index	**Model**	ϱ	L	$W_y(\%)$	SR_y	$p_{\chi^2_d}$	p_W	$\Delta\rho$ (%)	$2\Delta\varsigma^{=bh}$ **(bps)**
CSI 300	**GB-R**	**3**	**160**	40.09	0.49	0.33	0.06	1.23	29.7
	NN-C	4	150	36.55	1.08	$1.8{\cdot}10^{-3}$	0.09	2.16	28.5
	NN-C	3	380	33.76	1.06	$1.8{\cdot}10^{-3}$	0.09	2.47	25.7
FTSE	**SVM-R**	**3**	**390**	17.92	0.75	$7.2{\cdot}10^{-2}$	$2.4{\times}10^{-2}$	1.36	15.9
	DT-BR	3	250	16.80	0.43	$9.3{\cdot}10^{-2}$	$2.6{\times}10^{-2}$	1.11	11.2
	KNN-R	2	350	15.20	0.47	0.27	$3.5{\times}10^{-2}$	0.99	9.68
S&P 500	**DT-BC**	**3**	**400**	18.55	0.89	$2.0{\cdot}10^{-3}$	$1.9{\times}10^{-2}$	1.46	12.9
	DT-BR	3	330	17.90	0.74	$2.1{\cdot}10^{-3}$	$1.9{\times}10^{-2}$	1.51	9.38

Table 5.9: **Summary of the top performing models on the compounded wealth metric (p_W) for each equity index.** The key values in order are: the model family; the number of lags ϱ; the in-sample length L; the compounded annual growth rate W_y; the yearly Sharpe ratio SR_y; the p-values $p_{\chi^2_d}$ and p_W adjusted for multiple-testing in the entire universe of models; the excess predictability $\Delta\rho$ as defined in Equation 3.21; and the round trip transaction costs $2\Delta\varsigma^{=bh}$ breaking-even in W_y with the buy & hold strategy.

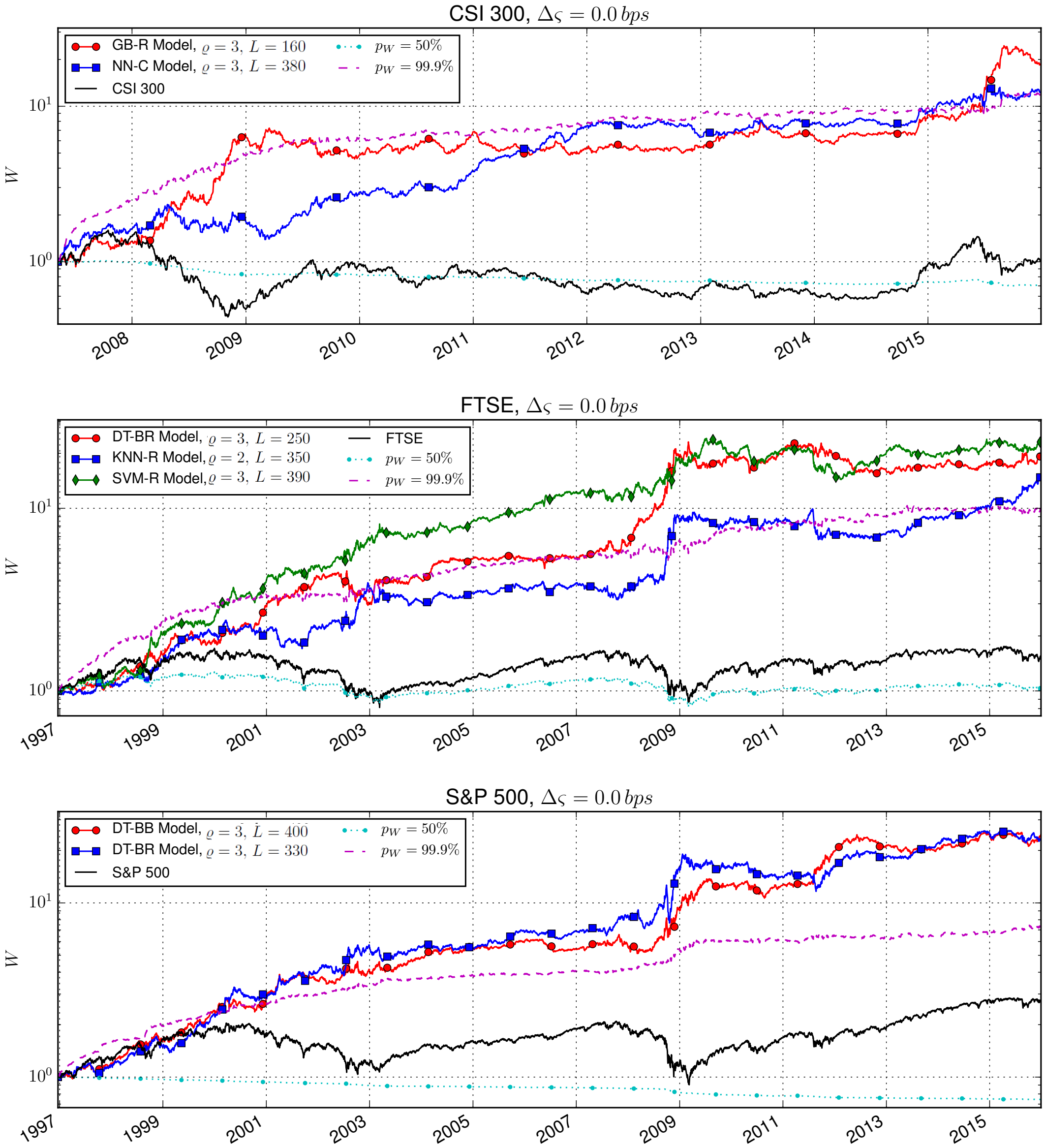


Figure 5.12: **Trading performance of the best model for each equity index.** These three figures show the compounded wealth $W(t)$ of the best model from Table 5.9 for the CSI 300, FTSE, and S&P 500. As reference, the figures show the market return (or buy & hold strategy), as well as the $p_W = 50\%$ and $p_W = 0.1\%$ quantiles of the randomized strategies. All strategies are shown at zero transaction cost $\Delta\varsigma = 0$.

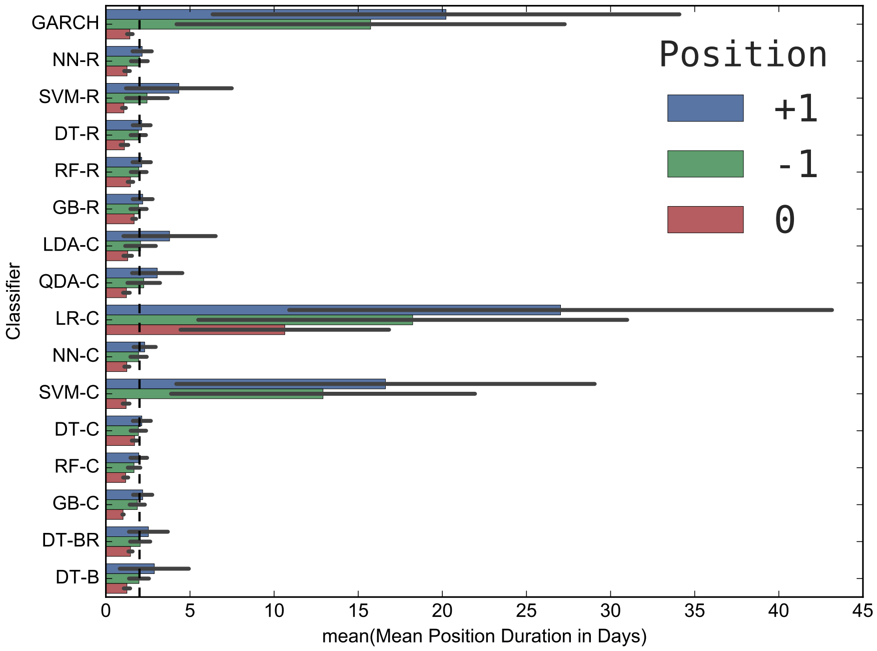


Figure 5.13: **Mean position duration in days by model family.** The position durations have been averaged over all lags, in-sample lengths, and equity indices. The error bars indicate the average one standard deviation of the durations for a single model. The model families can be split into two clusters: GARCH, LR-C, and SVM-C with high mean durations between 10 and 25 days; and all the other models with mean durations around two days (vertical dashed line).

abnormal returns of the NN-C model are made during the period 2009 to 2012, not visibly linked to a crisis. Nonetheless, the financial crisis (2008) and recent Chinese stock market turbulences (2015) are smoothened out in comparison to the buy & hold strategy.

5.4.3.2 Model mean position duration

The p-values adjusted for multiple testing across the entire model universe are computed using the algorithm from Section 3.4. Following the simulation study of Subsection 3.4.4, the block size of the bootstrap algorithm has to be close to the mean position duration to maximize the statistical power of the test. As well, the block size should not exceed the mean position duration of the strategies with shortest positions, otherwise the familiwise error rate would be violated.

To determine the optimal randomization scheme, the mean duration of all model families has been computed, as shown in Figure 5.13. Except for the GARCH, LR-C, and SVM-C models, the mean position duration is always close to two days. Given the low correlation of these three models with the other 13 models, these two groups of models are bootstrapped independently. While sampling independently reduces somewhat the statistical power of the test, this is more then offset by the improvement of removing the three models with long mean position duration. As can be read from Figure 3.6, longer

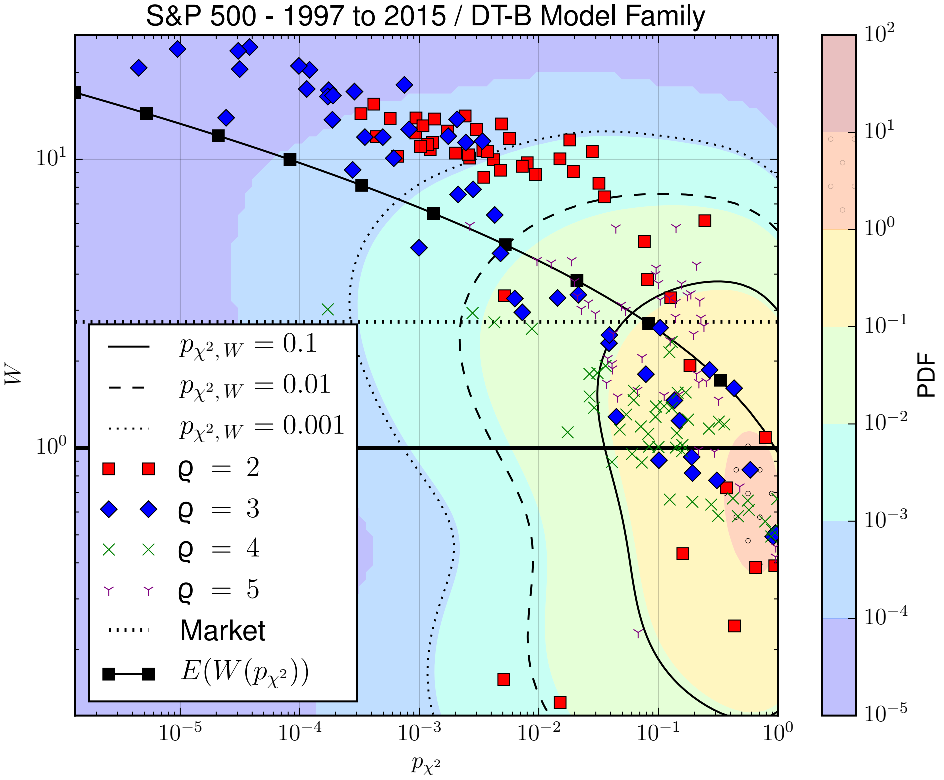


Figure 5.14: **The compounded wealth (W) versus the directional accuracy p-value ($p_{\chi^2_d}$) for the binary decision tree model family (DT-B), on the S&P 500 between 1997 and 2015.** The trading days between 1995 and 1997 are only used for calibration, but not for the out-of-sample performance. The four different lags $\varrho \in [2, 3, 4, 5]$ are differentiated by distinct markers. The market return during the period is given as a reference by a horizontal dotted line. The probability distribution function, as well as the confidence regions $p_{\chi^2_d,W}$, are obtained using one million simulated random strategies. The expected wealth $\mathrm{E}\left(W\left(p_{\chi^2_d}\right)\right)$ as a function of the directional accuracy p-value is computed by Equation 3.32, at zero transaction cost $\Delta\varsigma = 0$.

mean position durations imply significantly lower wealth quantiles, and therefore these strategies do not impact the sampling of the wealth tail of randomized strategies with shorter mean position duration.

5.4.3.3 Wealth correlation with directional accuracy

To better understand the performance of the models in the top performing family for each equity index, their wealth has been plotted against their directional accuracy. Figure 5.14 shows the DT-BC models on the S&P 500. The out-performers are found at lags two and three, with lag three visibly out-performing lag two. Figure 5.15 shows the SVM-R models on the FTSE. None of the lags is significantly out-performing the others, with the exception of one visible outlier at lag three. Figure 5.16 shows the NN-C models on the

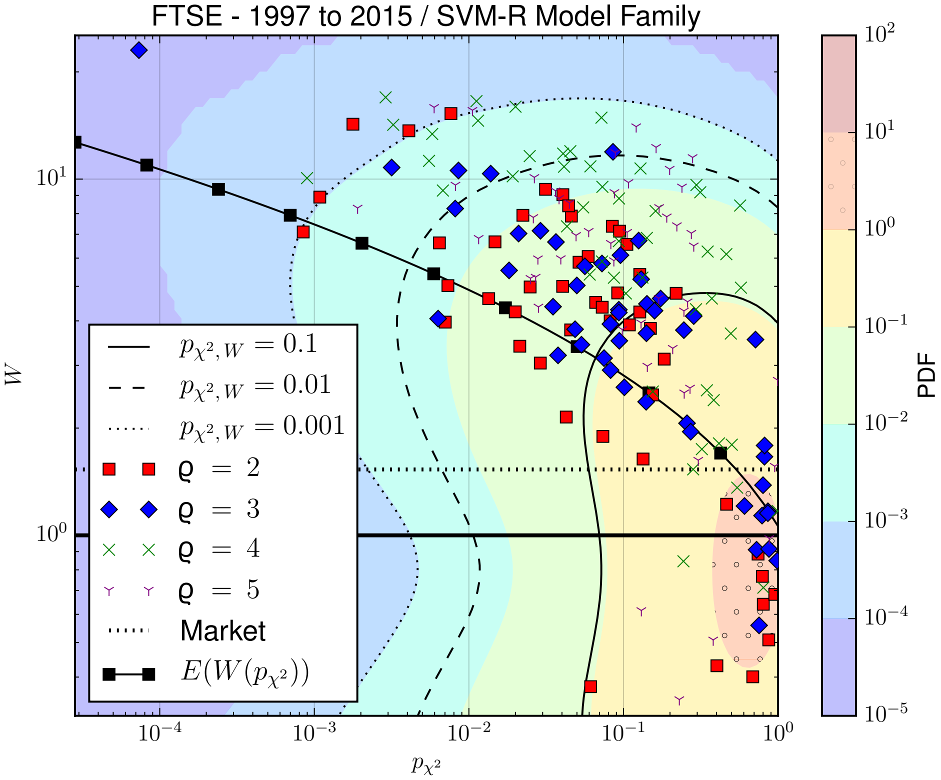


Figure 5.15: The compounded wealth (W) versus the directional accuracy p-value ($p_{\chi_d^2}$) for the support vector machine regression model family (SVM-R), on the FTSE between 1997 and 2015. For details see caption of Figure 5.14.

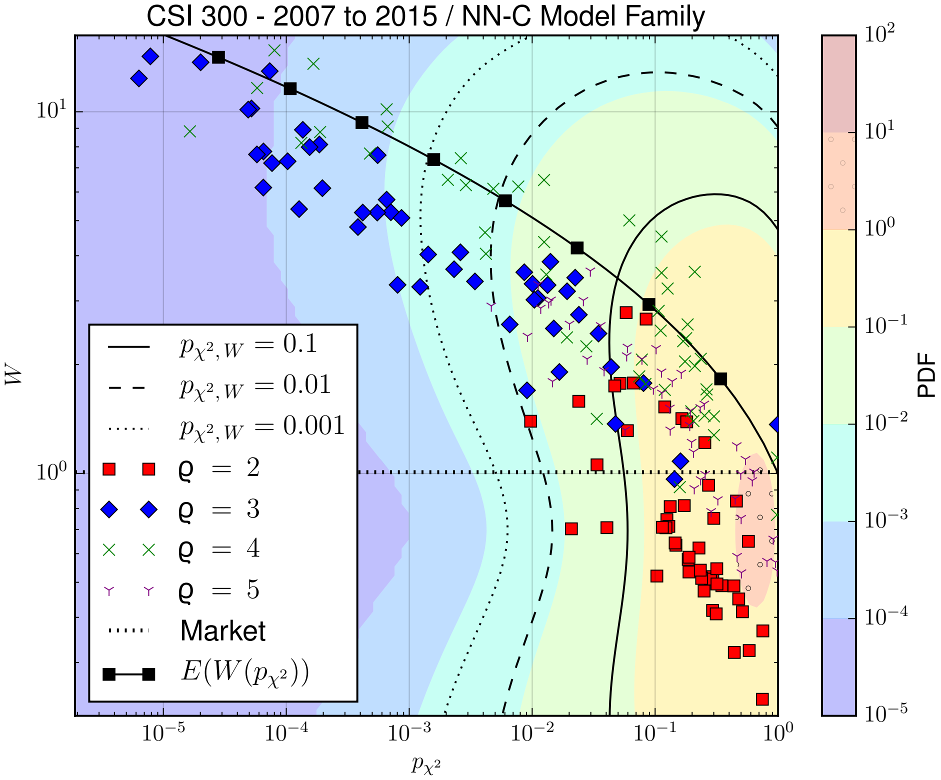


Figure 5.16: The compounded wealth (W) versus the directional accuracy p-value ($p_{\chi_d^2}$) for the k nearest neighbor classifier model family (NN-C), on the CSI 300 between 2007 and 2015. For details see caption of Figure 5.14.

CSI 300, with significant out-performers at lag three and four.

For classification models, the compounded wealth should follow closely the expected value $\mathrm{E}\left(W\left(p_{\chi_d^2}\right)\right)$ of Equation 3.32, as the binary outputs used for training do not contain any information about the return amplitude. Indeed, for the S&P 500 and the CSI 300, the classification models follow quite closely the theoretical relation, up to a constant shift. On the S&P 500, the models' actual wealth outperforms the expected wealth consistently, indicating that predictability was higher during volatile periods. On the CSI 300, the models actual wealth are in good agreement with the expected wealth for lag four, while being slightly lower for lag three. This indicates that the lag three predictors have missed some major tail events.

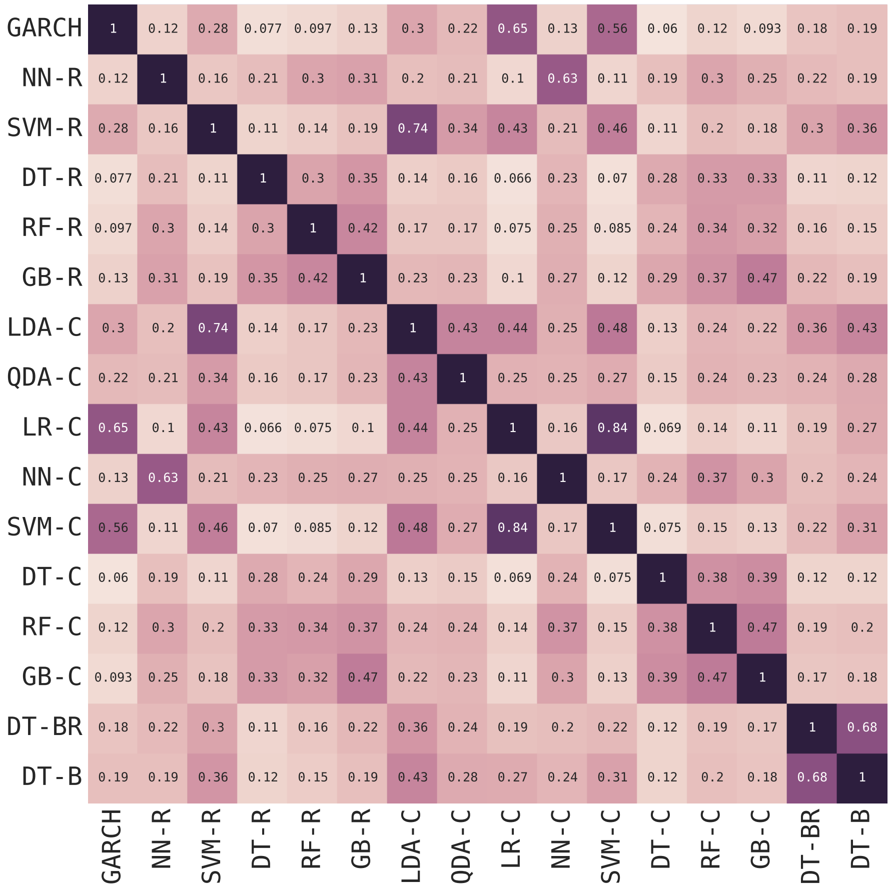
For regression models, the compounded wealth is not necessarily linked to overall directional accuracy, as the prediction can be strongly dependent on the return amplitude. Indeed, for the top performing regression model on the FTSE (SVM-R), there is no strong relationship between directional accuracy and compounded wealth. The models merely lie within the shape of the confidence intervals computed from random strategies.

The confidence regions of randomized strategies on the CSI 300 are strongly stretched along the W axis, which is in part due to the fact that the time period 2007-2015 is less then half the one used for the FTSE and S&P 500. This stretched confidence region is a result of fat tails in the return distribution. Consequently, on the CSI 300, more then half of the randomized strategies lie above the expected wealth curve.

5.4.3.4 Model correlations

To check for redundancy among the models, the Pearson correlation of the model families is computed, averaging all correlations at identical in-sample length and lag. The correlations averaged over all three equity indices are shown in Figure 5.17. Most model pairs have a correlation below 50%, clearly justifying their individual appearance in the model universe.

High correlations are found among the (SVM-R, SVM-C, LDA-C, LR-C) models that correlate up to 84%. This is expected, as these methods all find a hyperplane maximally separating up and down moves. In case of the LDA-C method, the hyperplane is separating the two multivariate distributions of the two classes. For the LR-C method, the hyperplane is maximizing the asymmetry of up and down moves on each side of the plane. The high correlation between the SVM-R and LDA-C models explains why both families are significant simultaneously on the FTSE (see Table 5.8). Two other well correlated pairs are (DT-BR, DT-B) and (NN-R, NN-C) with correlation coefficient equal to 68%, respectively 63%. This is expected too, as both models in these two pairs differ only in taking the average or majority vote among an identical group of samples. The high correlation between the DT-BR and DT-B models explains why both families are significant simultaneously on the S&P 500 and FTSE. Last, the LR-C method correlates at 65% with the GARCH model, indicating that it performs a similar type of trend following.



	GARCH	NN-R	SVM-R	DT-R	RF-R	GB-R	LDA-C	QDA-C	LR-C	NN-C	SVM-C	DT-C	RF-C	GB-C	DT-BR	DT-B
GARCH	1	0.12	0.28	0.077	0.097	0.13	0.3	0.22	0.65	0.13	0.56	0.06	0.12	0.093	0.18	0.19
NN-R	0.12	1	0.16	0.21	0.3	0.31	0.2	0.21	0.1	0.63	0.11	0.19	0.3	0.25	0.22	0.19
SVM-R	0.28	0.16	1	0.11	0.14	0.19	0.74	0.34	0.43	0.21	0.46	0.11	0.2	0.18	0.3	0.36
DT-R	0.077	0.21	0.11	1	0.3	0.35	0.14	0.16	0.066	0.23	0.07	0.28	0.33	0.33	0.11	0.12
RF-R	0.097	0.3	0.14	0.3	1	0.42	0.17	0.17	0.075	0.25	0.085	0.24	0.34	0.32	0.16	0.15
GB-R	0.13	0.31	0.19	0.35	0.42	1	0.23	0.23	0.1	0.27	0.12	0.29	0.37	0.47	0.22	0.19
LDA-C	0.3	0.2	0.74	0.14	0.17	0.23	1	0.43	0.44	0.25	0.48	0.13	0.24	0.22	0.36	0.43
QDA-C	0.22	0.21	0.34	0.16	0.17	0.23	0.43	1	0.25	0.25	0.27	0.15	0.24	0.23	0.24	0.28
LR-C	0.65	0.1	0.43	0.066	0.075	0.1	0.44	0.25	1	0.16	0.84	0.069	0.14	0.11	0.19	0.27
NN-C	0.13	0.63	0.21	0.23	0.25	0.27	0.25	0.25	0.16	1	0.17	0.24	0.37	0.3	0.2	0.24
SVM-C	0.56	0.11	0.46	0.07	0.085	0.12	0.48	0.27	0.84	0.17	1	0.075	0.15	0.13	0.22	0.31
DT-C	0.06	0.19	0.11	0.28	0.24	0.29	0.13	0.15	0.069	0.24	0.075	1	0.38	0.39	0.12	0.12
RF-C	0.12	0.3	0.2	0.33	0.34	0.37	0.24	0.24	0.14	0.37	0.15	0.38	1	0.47	0.19	0.2
GB-C	0.093	0.25	0.18	0.33	0.32	0.47	0.22	0.23	0.11	0.3	0.13	0.39	0.47	1	0.17	0.18
DT-BR	0.18	0.22	0.3	0.11	0.16	0.22	0.36	0.24	0.19	0.2	0.22	0.12	0.19	0.17	1	0.68
DT-B	0.19	0.19	0.36	0.12	0.15	0.19	0.43	0.28	0.27	0.24	0.31	0.12	0.2	0.18	0.68	1

Figure 5.17: **Pearson correlation of returns between the different model families, at identical in-sample length and lag.** The Pearson correlations have been averaged over in-sample length L, lags ϱ, and the three equity indices (S&P 500, FTSE, and CSI 300). The models are all positively correlated due to excess in up moves in the predicted asset, inducing a proportional excess in up predictions by the models. Most model pairs have a correlation below 50%, resulting from their fundamentally different prediction dynamics. None of the models pairs are redundant above the 84% level.

Index	W_y(%)	SR_y	$p_{\chi^2_d}$	p_W	$\Delta\rho$ (%)	$2\Delta\varsigma^{=bh}$ **(bps)**
CSI 300	-3.41	0.19	$2.5{\cdot}10^{-2}$	0.48	1.31	-0.6
FTSE	3.16	0.16	0.40	0.15	0.31	2.2
S&P 500	14.43	0.62	$9.8{\cdot}10^{-3}$	$1.0{\times}10^{-3}$	0.98	9.9

Table 5.10: **Summary of the best Sharpe ratio search technology portfolio performance for each equity index.** The key values in order are: the compounded annual growth rate W_y; the yearly Sharpe ratio SR_y; the p-values $p_{\chi^2_d}$ and p_W; the excess predictability $\Delta\rho$ as defined in Equation 3.21; and the round trip transaction costs $2\Delta\varsigma^{=bh}$ breaking-even in W_y with the buy & hold strategy.

5.4.3.5 Periods of abnormal performance

To systematically detect periods where strategies have abnormal returns, the distribution of the performance is analyzed with respect to the buy & hold strategy. The value $\log\left(\frac{S(t)}{BH(t)}\right)$ is used as a scale free measure of performance, were $S(t)$ is the value of the strategy, and $BH\,(t)$ is the value of the buy & hold strategy. The typical period of abnormal returns is around 6 months, and hence the performance measure $P_t = \Delta_{6m}\log\left(\frac{S_t}{BH_t}\right)$ is used, were Δ_{6m} is the 6 months differentiator. The distribution of P over time is found to be non-normal, and therefore the over-performance is defined as any value of P in the upper non-normal tail. Figure 5.18 shows the performance P over time for the same models as Figure 5.12, as well as the distribution of the performance P aggregated over the models for each equity index. The periods of excess performance confirm their correlation with the crisis periods mentioned previously.

5.4.3.6 Ex-ante performance

The finding of profitable model families, statistically significant after correcting for data snooping, does not guarantee that it would have been possible to select these models ex-ante. To settle this question, the best Sharpe ratio search technology is applied for each equity index to the entire universe of strategies. The performance of the resulting portfolios are presented in Table 5.10. The search technology is only significant for the S&P 500, and fails at selecting the best model for the FTSE and CSI 300.

These results are not surprising given the model family performances of Table 5.8. The S&P 500 has one model family (DT-BB) that outperforms the second best model family (DT-BR), while all the other families are insignificant. This performance distribution among the models makes it easy for the search technology to select the best model and stick to it. For the FTSE and the CSI 300, multiple model families are competing at similar significance levels. This induces the search technology to constantly switch to the new best performing model. Often the switch occurs at the moment were the best performing model suffers a draw-down, leaving the search technology portfolio with a mediocre performance compared to the best performing model.

The Figure 5.19 shows the trading performance over time of the search technology

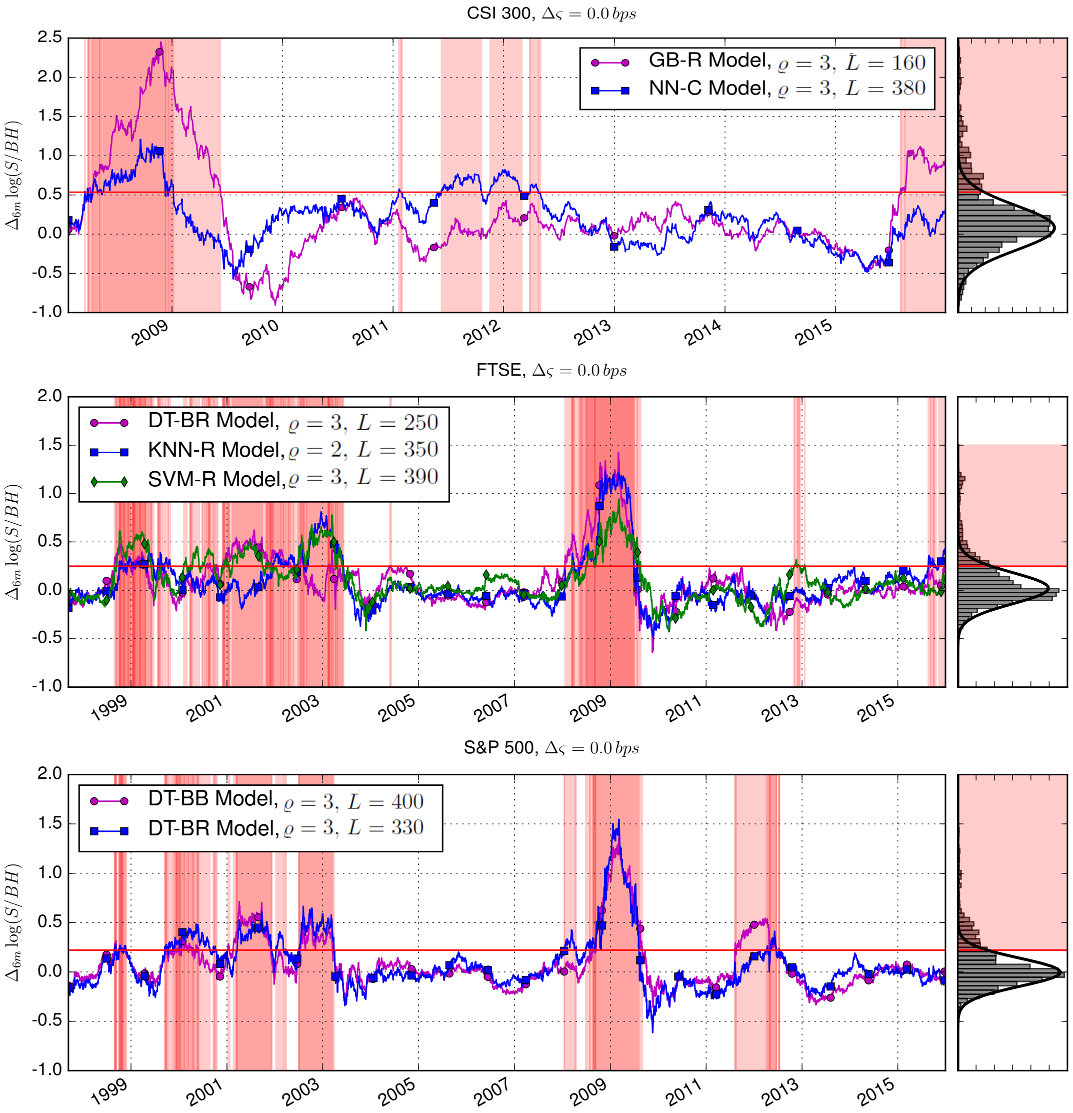


Figure 5.18: **Excess trading performance with respect to buy & hold.** The performance P is measured as $P_t = \Delta_{6m} \log\left(\frac{S_t}{BH_t}\right)$, were S_t is the value at time t of the strategy, BH_t is the value of the buy & hold strategy at time t, and Δ_{6m} is the 6 months differentiator. These three figures show the performance measure of the best models (GB: **G**radient **B**oosting; NN: **N**earest **N**eighbor; DT: **D**ecision **T**ree; SVM: **S**upport **V**ector **M**achine; suffixed as "-R" for regressors, "-C" for classifiers, and "-B" for binary) from Table 5.9 for the CSI 300, FTSE, and S&P 500. Over-performance is determined by the threshold (red line) were the tail of the performance distribution becomes non-normal.

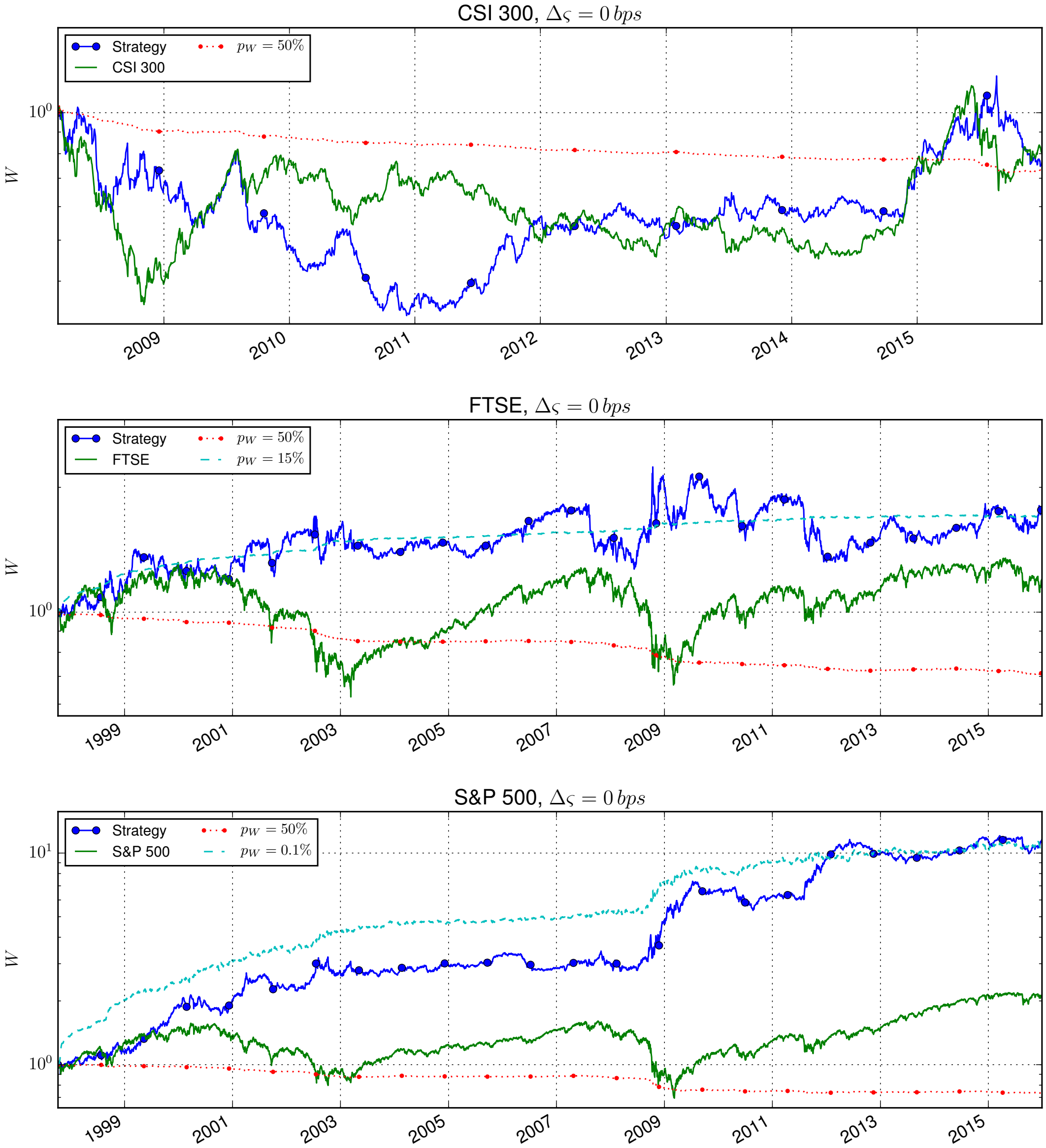


Figure 5.19: **Trading performance of the search technology portfolio for each equity index.** These three figures show the compounded wealth $W(t)$ of the best Sharpe ratio search technology, applied to the whole universe of models, for the CSI 300, FTSE, and S&P 500. As reference, the figures show the market return (or buy & hold strategy), as well as the $p_W = 50\%$ and break even quantiles of the randomized strategies. All strategies are shown at zero transaction cost $\Delta_\varsigma = 0$.

portfolio on all three equity indices. On the CSI 300, the search technology keeps switching models until 2011, with an unfavorable timing producing abnormally low returns. Starting in 2011, the search technology sticks to the best performing model of Table 5.9 (NN-C, $L = 380$, $\varrho = 3$), producing abnormally high returns until the end of 2015. On the FTSE, the search technology keeps switching models during the entire time period, never producing any abnormal returns. On the S&P 500, the search technology selects the best model of Table 5.9 (DT-BB, $L = 400$, $\varrho = 3$) early on, and benefits from the constant abnormal performance of this model.

5.5 Conclusion

This extensive performance analysis of statistical learning models to forecast daily returns found models significant at the 97.5% confidence level on the S&P 500 and FTSE, and 94% level on the CSI 300. The three best models, one for each equity index, were found at lag $\varrho = 3$, in-sample length $L \in [390, 400]$ for the S&P 500 and FTSE, and in-sample length $L = 160$ for the CSI 300. While they could be a statistical fluke, the consistency at lag $m = 3$ strengthens the finding of dependencies in the daily returns of the analyzed equity indices. The results clearly reject the martingale hypothesis $\mathrm{E}\left(r_t | r_{t-3}, r_{t-2}, r_{t-1}\right) = 0$, in line with the argumentation of Grossman and Stiglitz (1980) and the Adaptive Market Hypothesis (Lo, 2012). The S&P 500 and FTSE exhibit significant predictability in directional accuracy of the binary return sequences (DT-BR, and DT-B model families), which go unnoticed in linear regression models such as AR-GARCH. This finding is in good agreement with the study by Christoffersen and Diebold (2003). The CSI 300 exhibits significant local predictability among the five nearest neighbors of sequences of three and four returns (NN-C model families), a result that would go unnoticed as well in linear regression models.

The trading performance over time of the best models was most abnormal during the dot-com bubble (1997-2003), the financial crisis (2008-2009), the European debt crisis (2012), and the recent Chinese stock market turbulence (2015). The EMH seems to hold except during the market crises, where statistically significant deviations are found. Indeed, crisis periods are certainly the most propitious to drastic actions, driven by short-sighted human behavior, and not backed by statistics. As well, in times of fear and panic, investors tend to herd according to the psychology of "being safe in numbers", a behavior reducing the number of competing strategies, and creating a market dominantly driven by a single effective agent (i.e. the herding investors). Coupled with periodic announcements of major monetary institutions, such market dynamics have a high likelihood to introduce systematic biases that can be arbitraged. To verify the EMH as defined by Timmermann and Granger (2004), an ex-ante strategy select was performed using the best Sharpe ratio search technology. The EMH is rejected at transaction costs below 2.2 bps per round trip for the FTSE, and below 9.9 bps for the S&P 500.

The model universe could be extended with trinary decision tree models, and neural networks. However, extracting the found dependencies using unsupervised neural networks could reveal challenging. The data used as input could be extended to include more publicly available data such as dividends, risk free rates, or volatility levels. For equity indices showing significant anomalies, such as the S&P 500, the different sectors should be analyzed to determine which stocks drive the predictability. Further leveraging the randomization test, the search technology could be refined to determine the optimal window size on which to select a model, improving upon the expanding window used in this study.

Chapter 6

Emergent Predictability in Two Agent Models

6.1 Introduction

In the previous chapter, we showed that majority and minority agents can be represented as fixed classification trees. Further on, we have proven that a single agent, with the knowledge of all possible strategies, is equivalent to a fixed classification tree calibrated with the CART algorithm. Each branch of the classification tree uniquely maps to a history of returns, and the best strategy of the equivalent majority agent predicts the class (up or down) with highest probability. A minority agent predicts for each history the class with lowest probability. We then leveraged the highly efficient implementation of the CART algorithm in the scikit-learn library to perform an extensive study of single agent models.

Our study of the trading performance of statistical learning models showed that several fixed classification tree models outperform significantly the buy-and-hold benchmark. The abnormal trading performance holds true over a twenty year time period, across multiple assets, and after adjusting for multiple testing. The analysis of trading performance over time revealed significant pockets of predictability around the dotcom bubble, financial crisis, European debt crisis, and the Chinese stock market turbulence. The pockets of predictability occurred at distinct lags ϱ and calibrations length L.

However, our study only looked at the strategies that out-performed over the entire twenty year time period, and did not search for transient abnormal performance in the other strategies. As well, we did not test delayed predictions, which are known to be crucial in multi-agent models to observe transitions between herding and contrarian regimes (Andersen and Sornette, 2002). In this chapter, we are going to investigate more closely the fixed classification tree models for different delays. In particular, we search for abnormal performance at a yearly scale.

We find significant anomalous predictability on the NASDAQ for majority and minority models alike, after adjusting for multiple testing. The abnormal performance is mostly concentrated around the dotcom bubble and financial crisis. A principal component analysis reveals shared dynamics between the anomalous minority and majority models. The analyzed market crashes are preceded by a significant minority signal at short time scales, which then synchronizes with a majority signal at a longer time scale, rupturing the bubble and triggering the crash.

Current research has not yet answered the question if emergent stock market behavior can be calibrated through the collective intelligence of multiple agents. Hence, the abnormal performance in single agent models raises the question whether multiple such models, at different lag and calibration length, can be combined so as to significantly outperform the individual models. The answer turns out to be positive, the collective intelligence of a heterogeneous two agent model captures emergent dynamics, and has anomalous predictability significantly out-performing the single agent models.

The key to building the two agent model is the correct weighting of the two class probabilities of the individual agents. As the two agents are calibrated with different in-sample length and memory length, the confidence interval on their respective class probability differs. Hence, the aggregation of the two class probabilities into a single prediction needs to weight each probability by the average number of samples per history.

6.2 Representing Agents as Decision Trees

In minority and majority agent models, the global information shared by all agents is the history $\vec{r}_t = \{r_{t-1+\varrho}, \ldots, r_t\}$ of the most recent ϱ outcomes (i.e. memory), given as a binary sequence $\vec{r}_t \in \{-, +\}^{\times\varrho}$ of up and down moves. An individual strategy associates to each history an action to buy or sell. An agent possesses a private set of strategies, and keeps track of their performance on the past window of size L, using a payoff function π, which depends on the game played by the agent. At each time step, an agent buys or sells stock following his best strategy. Liquidity is often modeled using a trading threshold below which an agent would not trade Jefferies et al. (2001), however this feature is unnecessary for the present work.

6.2.1 Strategy Space

For fixed number of lags ϱ, a strategy is a function

$$s : \{-, +\}^{\times\varrho} \to \{-, +\}\,, \tag{6.1}$$

associating an up or down prediction to each history. An example of a two lag strategy is presented in Figure 6.1. The resulting strategy space

$$\mathbb{S}^{\varrho} := \{s \,|\, \forall s : \{-, +\}^{\times\varrho} \to \{-, +\}\} \tag{6.2}$$

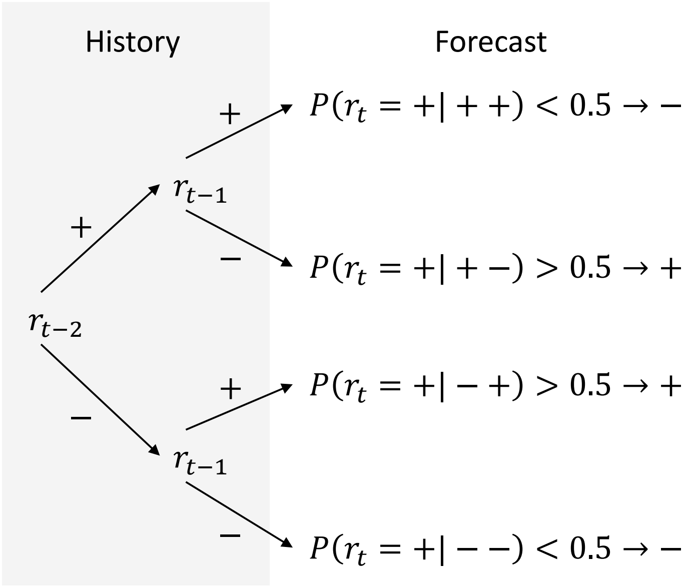

History
Forecast
r_{t-2}
r_{t-1}
r_{t-1}
+
−
+
−
+
−
$P(r_t = +| + +) < 0.5 \rightarrow -$
$P(r_t = +| + -) > 0.5 \rightarrow +$
$P(r_t = +| - +) > 0.5 \rightarrow +$
$P(r_t = +| - -) < 0.5 \rightarrow -$

Figure 6.1: **Strategy with two lags $\varrho = 2$ represented as a decision tree.** Each of the $2^2 = 4$ histories is assigned to a buy or sell action based on the class probabilities in the calibration data.

contains $|\mathbb{S}^\varrho| = 2^{2^\varrho}$ elements. In typical multi-agent simulations, some heterogeneity among agents is introduced by assigning to each agent A_i a subset $\mathbb{S}_i^\varrho \subseteq \mathbb{S}^\varrho$ of all possible strategies.

6.2.2 Delay

Agents with delayed actions are a crucial ingredient in mixed game models to observe speculative up-trends followed by a correction Andersen and Sornette (2002). Therefore, a delay parameter d is introduced, modeling the possible hesitation of agents, which may delay their trade by one or several days when they feel uncertain. A delayed strategy s forecasts the direction of the return d steps ahead as $\tilde{r}_{t+d+1} = s\left(\vec{r}_t\right)$.

6.2.3 Agent as Optimal Decision Tree

When calibrating a single agent model, heterogeneity in the agent's strategies is not required, and the agent can be endowed with the knowledge of all strategies. Models with a partial strategy set are possible, but not optimal from a decision theory point of view. As well, there are 2^ϱ partial strategy sets at lag ϱ, which would drastically increase the number of models to analyze (nonetheless feasible). Hence, the present work focuses on analyzing single agent models with the full strategy space.

A strategy is equivalent to a decision tree as illustrated in Figure 6.1, where each branch corresponds to a single history, and the terminal node of a branch assigns the action following its history. As proven in Subsection 6.2.4, the best performing strategy in $\mathbb{S}^\varrho$ for the majority payoff is always the optimal fixed classification tree $\mathcal{T}\left(\cdot\right)$, as constructed with the Classification And Regression Tree algorithm (CART, see Hastie et al., 2001) applied to the binary returns in the past window of size L. The samples used for training the decision tree at time t are given by

$$\mathbf{X}_t = \left\{ \left(\vec{r}_{t-\tilde{L},\, \varrho},\, r_{t-\tilde{L}+d+1} \right),\, \ldots,\, \left(\vec{r}_{t-1-d,\, \varrho},\, r_t \right) \right\}. \tag{6.3}$$

For training samples $\mathbf{X}$, the CART algorithm computes the class probabilities for $\{-,\, +\}$ in each branch (i.e. history). The class probabilities for a history $\vec{r}$ are determined by the ratio

$$P\left(\vec{r} \to -\right) = 1 - P\left(\vec{r} \to +\right) = \frac{\left|\{X \in \mathbf{X} | X = (\vec{r},\, -)\}\right|}{\left|\{X \in \mathbf{X} | X = (\vec{r},\, \cdot)\}\right|}. \tag{6.4}$$

Hence, a decision tree with binary classes is determined by a total of 2^{ϱ} probabilities, one for each branch. An optimal majority tree predicts for each branch the class with the highest probability, while the optimal minority tree predicts the class with the lowest probability. Ties are settled by a random prediction.

A detailed treatment of autoregressive decision trees and the CART algorithm is provided in Fievet and Sornette (2017). The authors proof that decision trees are robust predictors for linear autoregressive correlations, and additionally for non-linear sign correlations, which can be missed by linear autoregressive models. Hence, decision trees constitute a solid foundation for the prediction mechanism of agents.

6.2.4 Proof of Equivalence Between Agents and Decision Trees

To find the strategy in $\mathbb{S}^{\varrho}$ with the highest performance, as defined in Section 4.2.3, the brute force approach is to compute the performance of every strategy individually. However, the performance computation in Eq. (4.5) can be optimized by rearranging the sums as

$$U_{\pi,\, d,\, L}(s,\, t) \quad \propto \sum_{\vec{r} \in \mathcal{R}^{\varrho}} \sum_{j \in \mathcal{L} | \vec{r}_j = \vec{r}} \pi\left(r_{j+d+1},\, s\left(\vec{r}_j\right)\right), \tag{6.5}$$

where $\mathcal{L} = \{t - \tilde{L} + d,\, \ldots,\, t - 1 - d\}$. Now the inner sum over a fixed history $\vec{r}$ can be maximized independently from the outer sum. Denoting by $p_{\vec{r},\, r}$ the probability of finding the transition $P\left(\vec{r}_t \to r_{t+d+1}\right)$ in the calibration window of length L, the performance function is given by

$$U_{\pi,\, d,\, L}(s,\, t) \quad \propto \sum_{\vec{r} \in \mathcal{R}^{\varrho}} \sum_{r \in \mathcal{R}} p_{\vec{r},\, r} \pi\left(r,\, s\left(\vec{r}\right)\right). \tag{6.6}$$

To maximize the inner sum, the best strategy is now entirely determined as

$$s\left(\vec{r}\right) = \underset{r' \in \mathcal{R}}{\operatorname{argmax}} \sum_{r \in \mathcal{R}} p_{\vec{r},\, r} \pi\left(r,\, r'\right). \tag{6.7}$$

In case several values $r \in \mathcal{R}$ maximize the inner sum, a random choice is made. The computation time for the probabilities $p_{\vec{r},r}$ is linear with respect to the size L of the calibration window, which makes the computation of the best strategy independent from the number of strategies in the space $\mathbb{S}^{\varrho}$.

Updating the best performing strategy from one time step to the next is achieved in constant time. Removing the time step $t - \tilde{L} + d$ from $\mathcal{L}$, and adding the time step $t - d$, will affect at most two of the probabilities $p_{\vec{r},r}$, which are updated without recomputing the other probabilities. This change in the probabilities affects at most two of the inner sums in Equation (6.6), and therefore requires to update at most two of the values of $s(\vec{r})$.

The decision tree predicts for each history (i.e. branch) the class with the highest probability. This prediction mechanism is equivalent to selecting the strategy with the highest performance for the majority game as defined in Subsection 4.2.3. The probabilities $p_{\vec{r},r}$, used in Equation (6.7) to determine the strategy with the highest performance, are the class probabilities of the equivalent decision tree. Consequently, an agent endowed with the knowledge of all possible strategies, selecting at every time step the strategy with the highest performance on the past window of size L, is equivalent to an optimal decision tree calibrated on that window. In the case of the minority game, the tree prediction function is defined to return the class with the lowest probability.

6.3 Single Agent Experiments

6.3.1 Trading Strategies

The performance of a single agent model is evaluated by trading based on the one step ahead forecasts using a rolling window of size L, and taking a long or short position defined by $\mathfrak{s}_{t+1} = \operatorname{sign}(\tilde{r}_{t+1})$. This produces a sequence of binary trading signals $\{\mathfrak{s}_t\}_{L+1}^{T} \in \{-, +\}^{T-L}$ that define the corresponding trading strategy on the market returns $\{r_t\}_{L+1}^{T}$. Hence, a strategy is defined by a payoff function (MIN or MAJ), the number of lags ϱ, the calibration window length L, and the delay d. For further convenience, it is implicitly assumed that L initial returns have been cropped before the first forecast.

6.3.2 Experiments

We search for anomalous predictability by analyzing the trading performance of all meaningful single agent models on the NASDAQ during the 20 year time period Jan. 1, 1995 to Dec. 31, 2015. The daily returns are obtained from Thomson-Reuters Eikon with dividend adjustment. A total of 2000 experiments is run, determined by the two games (MIN and MAJ) and the following parameter space.

At lag ϱ, the class probabilities for the 2^{ϱ} unique histories need to be computed using (6.4). The accuracy of the computed class probabilities depends on the average number of samples per leaf given by $\frac{L}{2^{\varrho}}$. In (Fievet and Sornette, 2017), the best decision tree models on the S&P 500 are found at $\varrho \in \{1, 2, 3\}$ and $L \leq 500$, significantly above an average of 20 samples per leaf. The upper bound of two years on the investment horizon is independently found optimal to reproduce stylized facts (Feng et al., 2012). Hence, a calibration length of up to two years, and a maximum of four lags, constitute a reasonable search space that encompasses all potentially interesting models. Two strategies that only differ by their calibration window lengths L_1 and L_2 are highly correlated when $|L_2 - L_1| < 10$. Therefore, we use a step size of $\Delta L = 10$, which provides a sufficient sampling of strategies, while avoiding to compare almost identical strategies.

For the delay parameter d, we test $d \in \{0, 1, 2, 3, 4\}$. Experiments in behavioral finance show that traders are slow to update their beliefs, a phenomenon known as conservatism (Barberis et al., 1998). It takes between two to five contrarian observations before a subject changes his opinion, which motivates the range of chosen delays.

6.3.3 Anomalous Predictability

Detecting anomalies across a large number of experiments challenging, as the probability of spurious anomalies increases with the number of tests. We use the step-down methodology to correct for multiple testing (White, 2000, Romano and Wolf, 2005b, 2016), which rejects as many null hypotheses as possible, at given significance level, without violating the familywise error rate.

We measure risk adjusted returns with the Sharpe ratio (Sharpe, 1994), and benchmark against the buy-and-hold strategy using a stationary block bootstrap of returns, rejecting at a significance level of 0.05. The bootstrap p-values express the probability of the strategies under-performing the buy-and-hold strategy, when sampling with replacement blocks of returns from the time period of interest. However, during market draw-downs, a large number of lucky strategies are short in the market, hence have significant Sharpe ratio, but without possessing any true skill in predicting daily returns. To further restrict the definition of an anomaly, besides abnormal Sharpe ratio, we require the simultaneous presence of anomalous daily directional accuracy at a significance level of 0.05 (see Section 3.3.1). Hence, a model is anomalous when its Sharpe ratio is statistically significant, and explained by its ability to predict the daily market return direction.

We search for anomalous periods with a rolling window of one year (250 trading days), moving in steps of 10 trading days. In each window, the model returns are tested for the simultaneous occurrence of statistically significant Sharpe ratio and directional accuracy. For anomalous one year windows, the beginning is incrementally cropped, and the end incrementally cropped or extended, in steps of 10 days, until the window maximizes the test statistic. We remark that the bootstrap algorithm maintains the correlation structure between all strategies, and therefore the multiple testing adjusted p-values are not impacted by the choice of ΔL.

The anomalous single agent models for the S&P 500 and Dow Jones are shown in Figure 6.3, respectively Figure 6.4. The results are similar to the NASDAQ, strengthening the robustness of the methodology.

6.3.4 Empirical Results

The experiments revealed a number of anomalous periods in the NASDAQ as shown in Figure 6.2. During up trending markets, instances of models with sufficient directional accuracy to beat the buy-and-hold market are rare, and may be spurious. However, during flat but volatile markets, and market crashes, strongly anomalous periods are found for both games, robust to a large number of lags ϱ, delays d, and calibration lengths L.

Anomalous market dynamics are found precursory to the dotcom bubble and the financial crisis, providing empirical support that large changes in the stock market are preceded by increased predictability as theorized previously (Lamper et al., 2002). The events are preceded by an anomalous delayed market anti-persistency (minority game[1]) at a short time scale of 30 to 40 days. The delayed minority anomaly signals that a sizable fraction of market participants start to be contrarians. This dynamic breaks the growth phase of the bubbles, and transitions the market to a flat regime with high volatility. The contrarian anomalies are followed by an anomalous majority signal at shorter delay and longer time scale. The majority anomaly can be interpreted as a majority of traders flipping to the contrarian side, exiting the market, and hence bursting the bubble.

As a cross-check, the same search for anomalous models was performed on the S&P 500 and Dow Jones. The findings remain robust across all three equity indices, as shown in Figure 6.3, respectively Figure 6.4.

6.3.5 Principal Component Analysis

The dynamics during the dotcom bubble and financial crisis are characterized by the simultaneous occurrence of multiple anomalous single agent models, and the transition over time between anomalies at different time scales, lags, and delays. A better understanding of the interplay between the different anomalous regimes is crucial to the construction of multi-agent models.

To disentangle the multiple overlaid dynamics, and extract the dominant components, we use principal component analysis (PCA, see Hastie et al., 2001, Bishop, 2006). Denoting by $r^i = \{r^i_1, \ldots, r^i_T\}$ the returns of a strategy s^i in a set S of N strategies, the M principal components (PCs) are obtained by the minimization problem

$$\min_{V_q} \sum_{i=1}^{N} \left\| \left(r^i - \bar{r}\right) - V_M V_M^T \left(r^i - \bar{r}\right) \right\|^2, \tag{6.8}$$

which finds the projection V_M, onto a subspace of dimensionality M, which minimizes the reconstruction error. The returns are centered with respect to the mean return $\bar{r}$ of all strategies.

The PCs capture time periods, potentially disjoint, with the highest correlation among

---

[1] We remark that the term "minority" game can be misleading, as an anomalous minority signal implies that a dominant fraction of agents believes in a market reversal during a significant number of days. A more appropriate name would be "contrarian" game, or "reversal" game.

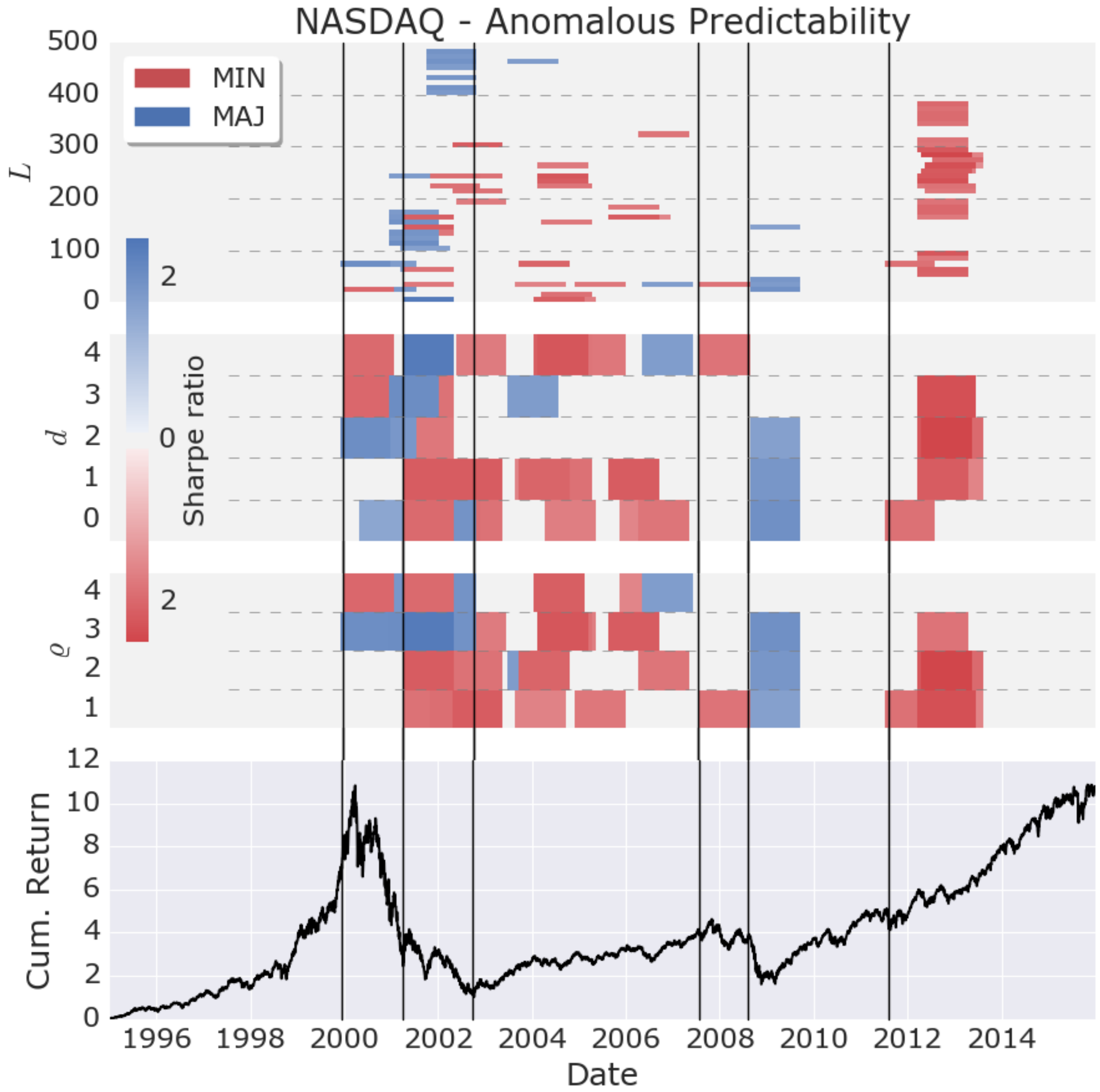


Figure 6.2: **Anomalous time periods found in the NASDAQ with single agent models during the time period Jan. 1995 to Dec. 2015.** A total of 2000 models are tested, resulting from: 2 games (MIN/MAJ) ×4 lags (ϱ) ×5 delays (d) ×50 calibration lengths (L). Using a rolling window of one year, 105 periods are found that exhibit statistically significant Sharpe ratio and directional accuracy, after adjusting for multiple testing. The intensity of the color is determined by the sum of the Sharpe ratios of the out-performing models. The black vertical lines are given as a visual aid for the timing of market anomalies.

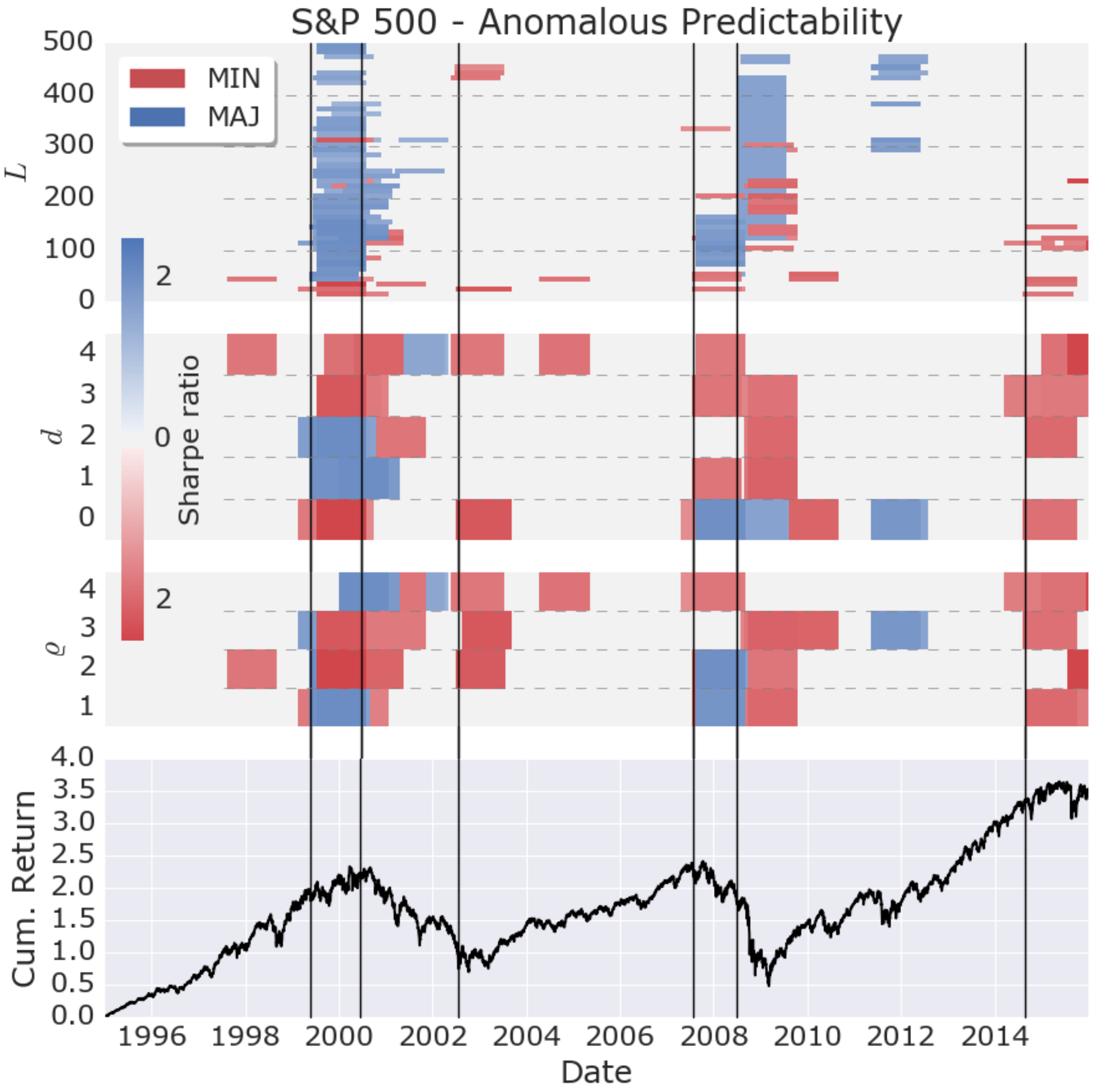


Figure 6.3: **Anomalous time periods found in the S&P 500 with single agent models during the time period Jan. 1995 to Dec. 2015.** A total of 2000 models are tested, resulting from: 2 games (MIN/MAJ) ×4 lags (ϱ) ×5 delays (d) ×50 calibration lengths (L). A total of 233 anomalous periods are found using a rolling window of one year, determining the models that exhibit statistically significant Sharpe ratio and directional accuracy after adjusting for multiple testing. The intensity of the color is determined by the sum of the Sharpe ratios of the out-performing models. The black vertical lines are given as a visual aid for the timing of market anomalies.

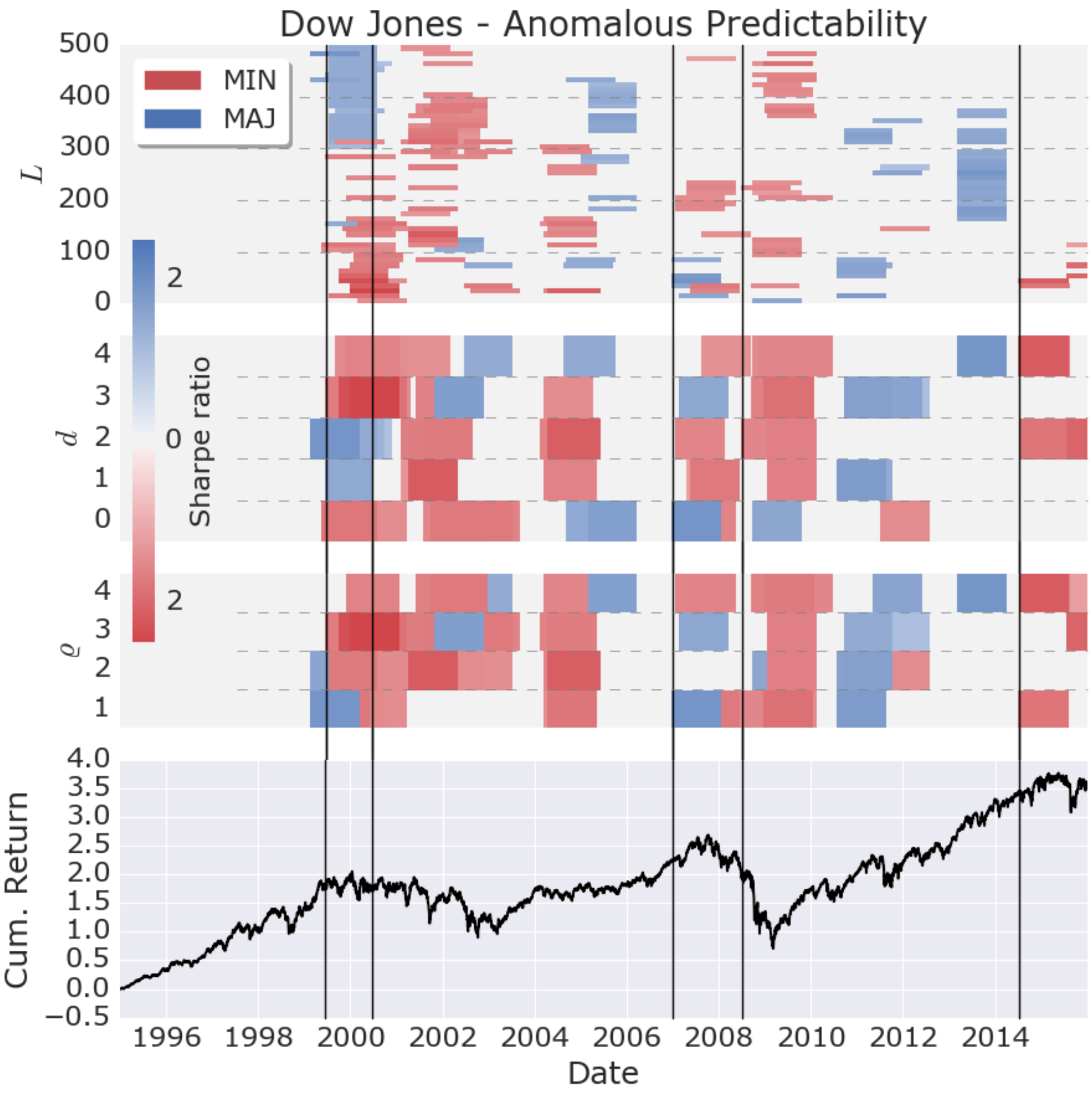


Figure 6.4: **Anomalous time periods found in the Dow Jones with single agent models during the time period Jan. 1995 to Dec. 2015.** A total of 2000 models are tested, resulting from: 2 games (MIN/MAJ) ×4 lags (ϱ) ×5 delays (d) ×50 calibration lengths (L). A total of 297 Anomalous periods are found using a rolling window of one year, determining the models that exhibit statistically significant Sharpe ratio and directional accuracy after adjusting for multiple testing. The intensity of the color is determined by the sum of the Sharpe ratios of the out-performing models. The black vertical lines are given as a visual aid for the timing of market anomalies.

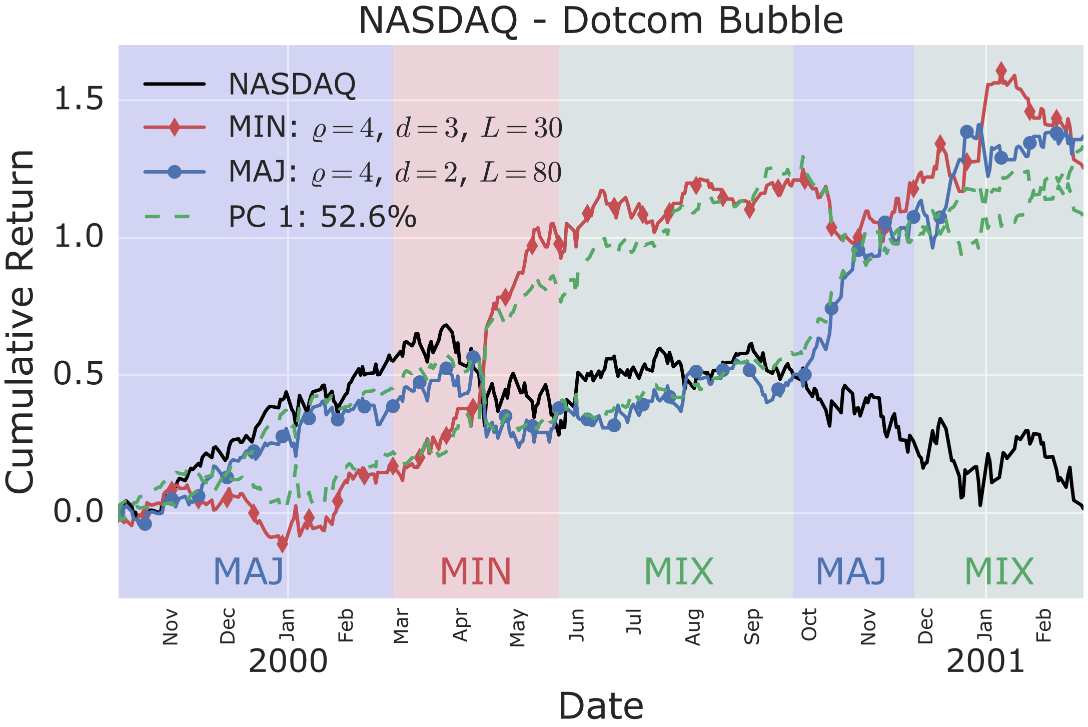


Figure 6.5: **Most anomalous minority and majority model on the NASDAQ during the crash of the dotcom bubble.** Six significant single agent models are found during the early peak and crash of the dotcom bubble. The first principal component (PC 1) explains 52.6% of the variance across these models. In particular, the first component is significantly shared by the minority and majority game models. Several successive regimes are distinguishable: (i) a majority regime (MAJ) building up to the peak; (ii) a strong minority regime (MIN) during the first crash in March 2000; (iii) a mixed regime in which the market remains flat; and finally (iv) a majority regime during the true burst of the bubble in September 2000. A 90% confidence boundary on directional accuracy is shown as reference.

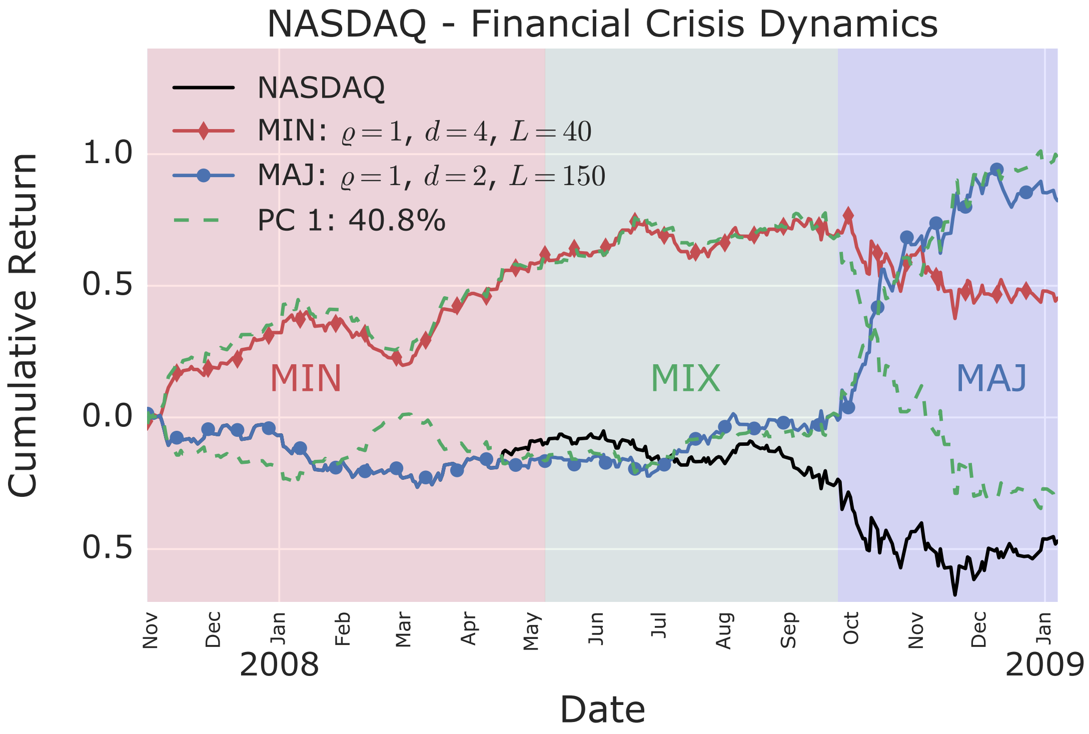


Figure 6.6: **Most anomalous minority and majority model on the NASDAQ during the financial crisis.** Eight significant single agent models are found during the financial crisis. The first principal component (PC 1) explains 40.8% of the variance across these models. In particular, the first component is significantly shared by the minority and majority game models. Three successive regimes are distinguishable: (i) an initial minority regime (MIN); (ii) a mixed regime (MIX) where the majority and minority models synchronize; and (iii) a strong majority regime during the burst of the bubble in September 2008. A 90% confidence boundary on directional accuracy is shown as reference.

a subset of strategies. Taking the mean return of the subset of strategies during the correlated periods, and zero outside the periods, is a PC that minimizes (6.8). The PCs allow us to determine the time periods with strongest inter-strategy correlation, and the projection coefficients determine the involved strategies for each PC.

Determining the first PC of the eight anomalous models during the beginning of the dotcom bubble revealed that 52.6% of the variance is explained by the first PC over a 15 month time frame. The first PC is significantly shared between the minority and majority models at different delays, showing that persistent and anti-persistent predictable patterns are intertwined. Figure 6.5 shows the two most anomalous models, alongside with their projection onto the first principal component. The minority model is gaining momentum during the first half of the year 2000, in particular during the first crash in late March 2000. During the subsequent period, the synchronicity between the majority and minority models keeps maturating, until the bubble ruptures in October 2000 accompanied by a strong majority signal.

As shown in Figure 6.6, the first PC of the six anomalous models during the financial crisis explains 40.8% of the variance over a 15 month time frame. There as well, the minority and majority game models significantly share the first PC, but in more distinct time periods. The financial crisis visible starts in late 2007 with a minority anomaly, while the majority model closely follows the buy-and-hold strategy. In April 2008, the minority and majority models start to synchronize, until a majority takes over in September 2008 and bursts the bubble.

The interplay between anomalous minority and majority patterns can be interpreted as the presence of two major groups of investors, confirming a study of individual traders that identified two investment styles (Goetzmann and Massa, 2002): contrarians profit takers; and positive feedback momentum traders.

6.4 Heterogeneous Two Agent Models

The strong first principal component shared by the anomalous minority and majority models are evidence that the market dynamics emerge from the interplay between multiple agents. Usually, the heterogeneity in multi-agent models is introduced by endowing the individual agents with different strategies, games, and delays. However, the present results lead us to drop the heterogeneity in strategies, and combine two optimal single agent models with different calibration lengths, games, lags, and delays. The heterogeneity in investment horizon is known to be crucial in ABMs to reproduce long-term memory observed in the stock market (Feng et al., 2012).

Combining the class probabilities of two individual agents, as defined in (6.4), is not as straightforward as taking their mean, because the distinct calibration lengths imply different accuracy on the estimated probabilities. We define the accuracy on the estimated class probabilities as the average number of samples per history, which is given by $\frac{L}{2^{\varrho}}$. Hence, the class probability P_{2A}^{-} of a down move in a two agent model, weighting by

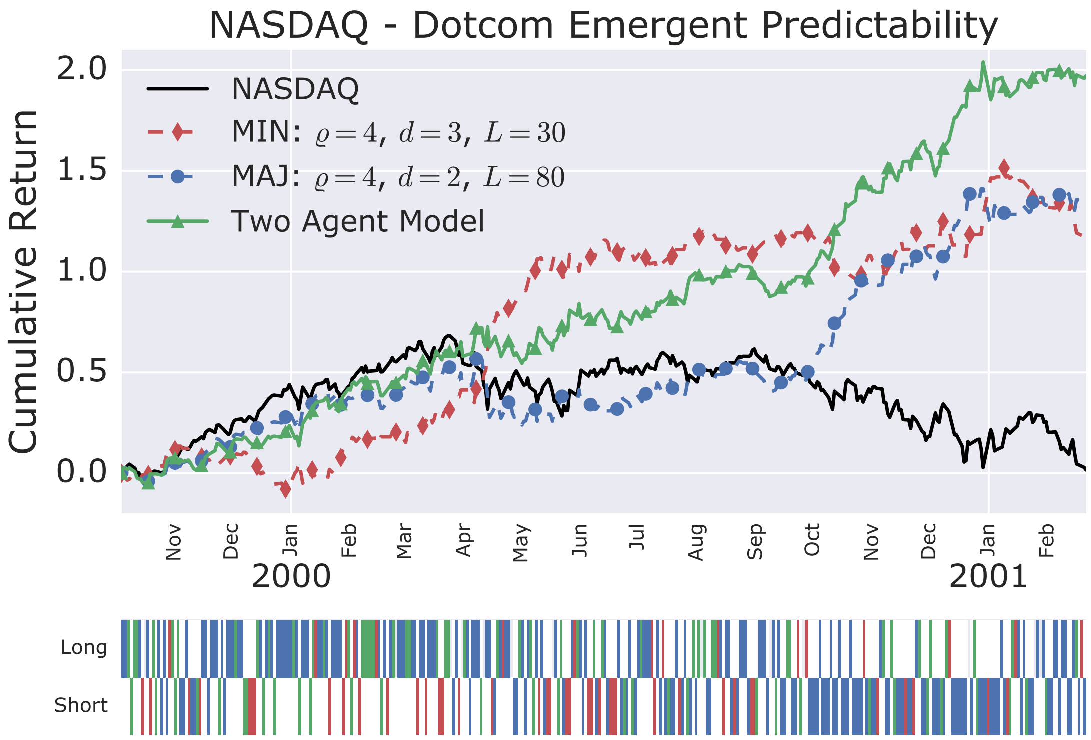


Figure 6.7: **Two agent model on the NASDAQ during the crash of the dotcom bubble.** The two agent model combines the forecast of the shown minority and majority agents as defined in (6.9). The lower panel indicates the long an short position taken by the two agent model, including the following color coding: red, minority agent dominated; blue, majority agent dominated; purple, both agents agreed.

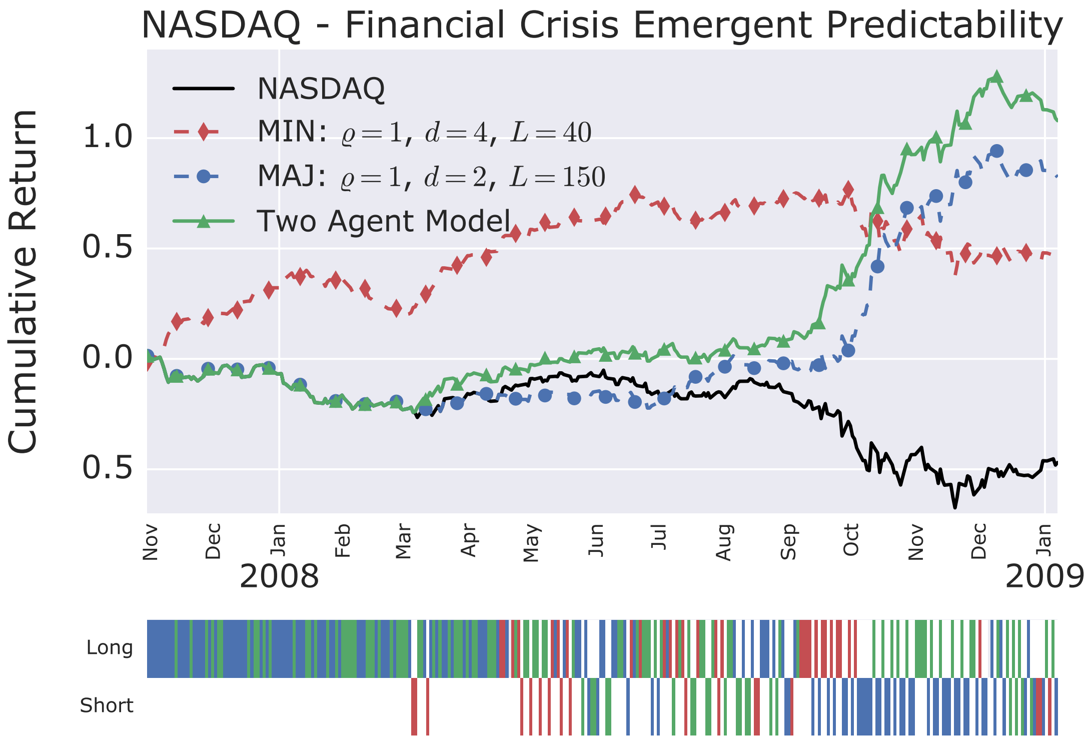


Figure 6.8: **Two agent model on the NASDAQ during the financial crisis.** The two agent model combines the forecast of the shown minority and majority models as defined in (6.9). The lower panel indicates the long an short position taken by the two agent model, with the following color coding: red, minority agent dominated; blue, majority agent dominated; purple, both agents agreed.

	Annualised Sharpe Ratio		Break-even cost (bps)	
Model	Dotcom	F. C.	Dotcom	F. C.
One Agent (MIN)	1.6	1.0	49	72
One Agent (MAJ)	1.9	1.7	65	137
Two Agents Mix	2.6	2.3	93	138

Table 6.1: **Sharpe ratio and buy-and-hold break-even transaction costs of the best single agent minority and majority model, and the two agent model, during the dotcom bubble and financial crisis (F.C.).** The individual model's cumulative return over time is shown in Figure 6.7, respectively Figure 6.8. The two agents models out-perform the single agent models in daily Sharpe ratio above the 95% confidence level.

accuracy the individual class probabilities of the two agents, is given by

$$P_{2A}^{-} = \frac{2^{\varrho_2} \cdot L_1 \cdot P_{A_1}^{-} + 2^{\varrho_1} L_2 \cdot P_{A_2}^{-}}{2^{\varrho_2} \cdot L_1 + 2^{\varrho_1} \cdot L_2}, \tag{6.9}$$

where $P_{A_1}^{-}$ is the class probability of agent one, and $P_{A_2}^{-}$ of agent two. This weighting ensures that the class probabilities of the two single agents are weighted by their respective confidence interval.

The cumulative returns over time of the two agent model, combining the two best single agent models, during the dotcom bubble and the financial crisis is shown in Figure 6.7, respectively Figure 6.8. Table 6.1 shows the improvement in daily Sharpe ratios of the two agent model in comparison to the underlying single agent models. In both cases, the two agent model out-performs above the 95% confidence level the single agent models.

The out-performance is a non-trivial result, because there is no guarantee that the strength and timing of the anomalous predictability in the individual models allows for an improved combined forecast. In a scenario where correct class probabilities are weak, and wrong class probabilities are strong, the individual anomalous predictability could equally well cancel out when combined. The out-performance of the two agent model is proof that the market returns are the result of emerging dynamics from the interplay between contrarian and momentum traders.

6.5 Implementation

The single agent and two agent prediction models are implemented using the **scikit-learn** package in python. The multiple-testing adjusted p-values for the Sharpe ratio can be computed out of the box with the **arch** package in python. The implementation of the single agent model, as a wrapper of the DecisionTreeClassifier class, is provided below.Add

```
import numpy as np
```

```python
from sklearn.tree import DecisionTreeClassifier


class OneAgent(object):
  def __init__(self, game, lags, delay, l):
    # Game is majority (MAJ) or
    # minority (MIN)
    self.game = game

    self.lags = lags
    self.delay = delay

    # Calibration window length L
    self.l = l

  def predict_proba(self, returns):
    # Local shortcuts
    p = self.lags
    d = self.delay
    l = self.l

    # Compute binary returns of
    # up and down moves.
    # Take only l (L) last returns
    b_returns = [
      1 if r >= 0 else -1
      for r in returns[-l:]
    ]

    # <- ϱ -><- d -><- output ->

    # Compute inputs as an array
    # with shape (L − ϱ − d, ϱ)
    inputs = np.array([
      returns[i:-(p + d - i)]
      for i in range(p)
    ], dtype=float).transpose()

    # Compute outputs matching to the inputs
    outputs = b_returns[p+d:]
    # Create a classification tree instance
    tree = DecisionTreeClassifier(
      criterion="gini",
      splitter="best",
      max_depth=None,
      min_samples_split=2,
      min_samples_leaf=1,
      min_weight_fraction_leaf=0.,
      max_features=None,
      random_state=None,
      max_leaf_nodes=None,
      min_impurity_split=0,
      class_weight=None,
      presort=False
    )

    # Train the tree
    tree.fit(inputs, outputs)

    if d == 0:
      prediction_input = [b_returns[-p:]]
    else
      prediction_input = [b_returns[-(p+d):-d]]

    ps = tree.predict_proba(prediction_input)[0]

    # Threshold determination
    dp = ps[1] - ps[0] if len(ps) > 1 else ps[0]

    # Game determination
    return dp if self.game == "MAJ" else -dp

  def predict(self, returns):
    dp = self.predict_proba(returns)
    return np.sign(dp) if np.abs(dp) >= 0 else 0


# Predict a sequence of returns with a model
def strategy_returns(returns, model):
  max_l = 500
  return [
    model.predict(returns[(i-max_l):i])
    for i in range(max_l, len(returns))
  ] * returns[max_l:]
```

The two agent model combines two single agents as implemented below.

```
import numpy as np


class TwoAgent(object):
  def __init__(self, agent1, agent2):
    self.agent1 = agent1
    self.agent2 = agent2

  def predict(self, returns):
    dp1 = self.agent1.predict_proba(returns)
    dp2 = self.agent2.predict_proba(returns)

    w1 = self.agent1.l * 2**self.agent2.lags
    w2 = self.agent2.l * 2**self.agent1.lags
    dp = (w1 * dp1 + w2 * dp2)/(w1 + w2)

    return np.sign(dp) if np.abs(dp) >= 0 else 0
```

The p-value for the directional accuracy can be computed efficiently as an independence test on the confusion matrix, as implemented below.

```
from sklearn.metrics import confusion_matrix
from scipy.stats import chi2_contingency


def directional_p_value(true_returns, predicted_returns):
    # Ensure returns are binary sequences of up and down moves
    true_returns = [0 if r <= 0 else 1 for r in true_returns]
    predicted_returns = [0 if r <= 0 else 1 for r in predicted_returns]

    # Compute confusion matrix
    cm = confusion_matrix(true_returns, predicted_returns)

        # Return p-value of Pearson independence test
    return chi2_contingency(cm)[1]
```

6.6 Conclusion

Starting from the common definition of minority and majority agents with binary strategies, we prove their equivalence to optimal decision trees. The low computational cost of computing optimal decision trees allows us to test all meaningful single agent models for anomalous trading performance on the NASDAQ. Periods with anomalous Sharpe ratio and directional accuracy are found in strong correlation with the dotcom bubble and financial crisis.

We disentangle the complex dynamics of multiple co-occurring anomalous models using principal component analysis. The first principal component of the anomalous models during the early stage of the dotcom bubble, and during the financial crisis, explain more than 40% of the variance. The first component reveals significantly shared dynamics between the anomalous minority and majority game. Both bubbles are characterized by a precursory minority signal, likely to origin from an increased number of contrarian traders expecting market anti-persistency. The minority signal is subsequently followed by a majority signal, likely to origin from a market consensus of an imminent draw-down.

A novel mechanism, combining two individual agents into a two agent model, reveals that the collective intelligence of two agents uncovers emerging predictability. The results demonstrate that the heterogeneity in investment horizons among agents is a crucial feature of the stock market dynamics, confirming results from the calibration of stylized facts (Feng et al., 2012).

Single agent models perform a spectroscopy of financial markets, identifying anomalous market dynamics that are precursors to an imminent rupture in current market dynamics. Two agent models have to be further explored to determine their potential in predicting anomalous market dynamics.

Chapter 7

Conclusion & Outlook

The focus of this thesis is on out-of-sample predictability using non-linear ensemble models. In complex systems, the observed variables at the macro level arise from interactions at the micro level. The particularity of such systems is that the behavior emerging at the macro level is typically not deducible from the rules determining the behavior at the micro level. Complex systems typically exhibit non-linear behavior at the macro level, which is often of chaotic nature. Such complex behavior makes it difficult to calibrate models and make reliable forecasts.

The first complex system for which we developed a non-linear ensemble model is that of oil production in the UK and Norway. Past studies typically use the Hubbert model to extrapolate aggregate oil production. However, such an aggregate extrapolation is unlikely to be accurate, because the oil production profiles arising at the country level result from many oil fields with diverse production curves, and a non-trivial process of discovery of new fields. A back-test showed that the Hubbert model is highly unstable over time, and can be improved. Our model, which extrapolates each oil field individually, and models the discovery process of new giant and dwarf fields, revealed to be stable over time and to accurately capture the fat tail in declining oil production. After three years of out-of-sample forecast, the true oil production is still within the one standard deviation band of the forecast.

The second and major application of non-linear ensemble models developed in this thesis is the prediction of daily equity index returns. Assuming that the weak efficient market hypothesis holds true, no model can predict the returns at levels that are profitable after transaction costs. However, existing research on agent based models established that they can reproduce many stylized facts of financial markets. Especially, studies performing out-of-sample forecasts did find significant excess predictability using agent based models. In this thesis, we used state of the art benchmarks to determine if agent based models can challenge the weak efficient market hypothesis.

The space of sensible single agent majority and minority models turned out to be fairly large. A single agent model is determined by the memory length of the agent, the in-sample length used for calibration, and the game played (majority and minority). In total, the single agent model space contains two hundred instances, resulting from the different parameter combinations to test. Our study of trading performance of single agent models found statistically significant Sharpe ratio on the S&P 500, FTSE, NASDAQ, and Dow Jones, after adjusting for multiple testing. However, the mixed-game multi-agent model, calibrated using genetic algorithms, could not outperform the single agent models. The calibration of emergent phenomena remained an open problem, which we subsequently tackled with smaller incremental steps.

Further on, we showed that the agents in majority and minority games can be represented as fixed classification trees. In particular, single agent models with the knowledge of all possible strategies are equivalent to a fixed classification tree calibrated with the CART algorithm. The proof of this equivalence established a formal bridge between agent based modeling and statistical learning. In turns out that a multi-agent minority or majority model is not far from a random forest or gradient boosting model, which combine multiple weak learners (i.e. a single agent) into a strong learner. However, major differences exist, in particular multiple agents cannot be calibrated in isolation, and the non-stationary time series nature of the problem prevents the construction of a test set.

Given the connection between agent based models and classification trees, we studied the predictive power of a broad number of statistical learning models. Besides the decision tree model, we found that the support vector regression, nearest neighbor, and gradient boosting models as well have significant abnormal trading performance on some equity indices. We have proven that decision tree models can capture non-linear return sign dependencies that would go unnoticed with linear autoregressive models. These results show that while returns are typically not linearly correlated, there are number of transient non-linear dependencies missed by conventional linear time series models. The strength of these dependencies does challenge the weak efficient market hypothesis.

Last, we explored two agent models, to determine if two agents can calibrate emergent return dependencies that cannot be described by a single agent model. The result was positive, the two agent model did significantly outperform the single agent model during the dotcom bubble and financial crisis. The key to combining the predictions of the two individual agents is the correct weighting of their class probabilities. The class probabilities need to be weighted by their confidence interval, resulting from the average number of samples per history available for training.

The empirical results obtained in this thesis revealed abnormal non-linear dependencies in stock market returns. While we analyzed the decision tree models in detail, the abnormal predictability found with support vector regression, nearest neighbor, and gradient boosting models needs to be further explored. The nearest neighbor model should be tested for all sensible number of neighbors, and the fixed radius nearest neighbor variant needs to be tested as well. The gradient boosting model could be a powerful generalization of decision trees that works in a broader number of scenarios. However, an algorithm to tune the learning rate of the gradient boosting in a time series context needs to be devised.

We have proven that the decision tree type models can capture non-linear return dependencies that would go unnoticed with linear autoregressive models. Similarly, all statistical learning models should be analyzed in a times series context to determine what type of dependencies they can capture, and how they overlap or complement conventional autoregressive models. A unification of statistical learning and time series analysis into "time series learning" is only a matter of time. Current research in time series learning has already studied numerous non-parametric and non-linear models, which complement or overlap with the non-parametric models in statistical learning.

The study of trading performance in this thesis only used daily equity index returns. Further studies should test for predictability using more data, such as trading volume, daily high and low return, implied volatility, interest rates, and other publicly available data with possible causal relations. To avoid overfitting, a careful feature selection needs to be performed. An alternative to feature selection could be the extraction of principal components, which are then used as inputs. Besides using a broader set of input data, the predictability needs as well to be studied at different time scales. Intraday returns, or even complete order book information, provide a much larger dataset, enabling the calibration of more complex models.

The search for abnormal market predictability was performed over a large model space, resulting from different parameter combinations. The study of two agent models showed that the predictions can be aggregated so as to outperform the performance of single agent models. However, it remains open how to systematically aggregate the predictions of all models into a single forecast. Multiple possibilities have to be explored to determine what works best in the non-stationary context of financial markets. A first possibility is to aggregate the model forecasts weighted by the posterior probability of their in-sample performance. A second possibility could be to weight the models by the p-value of their out-of-sample performance, which however would introduce a new parameter for the window size on which to evaluate the performance. Eventually, some non-linear weighting by these probabilities could turn out to be the optimal aggregation mechanism.

The theoretical results derived in this thesis established strong connections between agent based modeling and statistical learning. Further research should be dedicated to describe agent based models in the formalism of statistical learning, and ultimately integrate agent based models as a new set of algorithms to construct ensemble predictor from multiple weak learners (i.e. agents). Similar to the random forest or gradient boosting algorithms, but calibrating synchronously the weak learners. Given that neural networks with long short term memory are Turing complete (i.e. universal computers), any agent based model can necessarily be described as a neural network. Hence, the integration of agent based models as a subbranch of deep learning is only a matter of time. Key to this integration is the development of interpretation algorithms, which describe a calibrated neural networks in human understandable terms (e.g. different agents).

Bibliography

Alfarano, Simone, Thomas Lux, and Friedrich Wagner, 2005, Estimation of agent-based models: The case of an asymmetric herding model, *Computational Economics* 26, 19–49.

Alfi, V., M. Cristelli, L. Pietronero, and A. Zaccaria, 2009a, Minimal agent based model for financial markets I: Origin and self-organization of stylized facts, *The European Physical Journal B* 67, 385–397.

Alfi, V., M. Cristelli, L. Pietronero, and A. Zaccaria, 2009b, Minimal agent based model for financial markets II: Statistical properties of the linear and multiplicative dynamics, *The European Physical Journal B* 67, 399–417.

Almgren, R F, 2003, Optimal execution with nonlinear impact functions and trading-enhanced risk, *Applied Mathematical Finance* 10, 1–18.

Andersen, J. V., and D. Sornette, 2005, A Mechanism for Pockets of Predictability in Complex Adaptive Systems, *Europhys. Lett.* 70, 697–703.

Andersen, Jorgen Vitting, and Didier Sornette, 2002, The $-game, *European Physical Journal B* 31, 141–145.

Andrews, Donald, 1991, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation, *Econometrica* 59, 817–858.

Ardila, D. A., Z. Forro, and D. Sornette, 2015, The Acceleration Effect and Gamma Factor in Asset Pricing, *Swiss Finance Institute Research Paper No. 15-30. Available at SSRN: http://ssrn.com/abstract=2645882* .

Arthur, W. B. Brian, 1994, Inductive Reasoning and Bounded Rationality, *The American Economic Review* 84, 406–411.

Atsalakis, George S., and Kimon P. Valavanis, 2009, Surveying stock market forecasting techniques - Part II: Soft computing methods, *Expert Systems with Applications* 36.

Atsalakis, George S., and Kimon P. Valavanis, 2013, Surveying stock market forecasting techniques, Part I: Conventional methods.

Bak, P, M Paczuski, and M Shubik, 1997, Price Variation in a Stock Market with Many Agents, *Physica A* 246, 430–453.

Baker, Monya, 2016, 1,500 Scientists lift the lid on reproducibility, *Nature* 533, 452–454.

Barberis, Nicholas, Andrei Shleifer, and Robert Vishny, 1998, A model of investor sentiment, *Journal of Financial Economics* 49, 307–343.

Barras, Laurent, Olivier Scaillet, and Russ Wermers, 2005, False Discoveries in Mutual Fund Performance : Measuring Luck in Estimated Alphas, *The Journal of Finance* 65, 179–216.

Bishop, Christopher M, 2006, *Pattern Recognition and Machine Learning* (Springer-Verlag New York, Inc.).

Bonabeau, Eric, 2002, Agent-based modeling: Methods and techniques for simulating human systems, *Proceedings of the National Academy of Sciences* 99, 7280–7287.

Bouchaud, Jean-Philippe J-P., Julien Kockelkoren, and Marc Potters, 2006, Random walks, liquidity molasses and critical response in financial markets, *Quantitative Finance* 6, 115–123.

Brandt, Adam R, 2007, Testing Hubbert, *Energy Policy* 35, 3074–3088.

Brecha, Robert J, 2012, Logistic curves, extraction costs and effective peak oil, *Energy Policy* 51, 586–597.

Brock, William a., and Cars H. Hommes, 1997, A Rational Route to Randomness, *Econometrica* 65, 1059–1095.

Cao, Lijuan, and E H Francis Tay, 2001, Financial Forecasting Using Support Vector Machines, *Neural Computing & Applications* 10, 184–192.

Carhart, Mark M., 1997, On Persistence in Mutual Fund Performance, *The Journal of Finance* 52, 57–82.

Chakraborti, A., I.M. Toke, M. Patriarca, and F. Abergel, 2011, Econophysics review: II. Agent-based models, *Quant. Finance* 11, 1013.

Challet, D., A. Chessa, M. Marsili, and Y.-C. Zhang, 2001, From Minority Games to real markets, *Quantitative Finance* 1, 168–176.

Challet, D., and Y.-C. Zhang, 1997, Emergence of cooperation and organization in an evolutionary game, *Physica A: Statistical Mechanics and its Applications* 246, 407–418.

Challet, Damien, Matteo Marsili, and Riccardo Zecchina, 2000, Statistical Mechanics of Systems with Heterogeneous Agents: Minority Games, *Physical Review Letters* 84, 1824–1827.

Challet, Damien, Matteo Marsili, and Yi-Cheng Zhang, 1999, Modeling market mechanism with minority game, *Physica A: Statistical Mechanics and its Applications* 276, 284–315.

Challet, Damien, and Yi-Cheng Zhang, 1998, On the minority game: Analytical and numerical studies, *Physica A: Statistical Mechanics and its Applications* 256, 514–532.

Chen, Fang, Chengling Gou, Xiaoqian Guo, and Jieping Gao, 2008, Prediction of stock markets by the evolutionary mix-game model, *Physica A: Statistical Mechanics and its Applications* 387, 3594–3604.

Chen, Shu-Heng, and Chia-Hsuan Yeh, 2002, On the emergent properties of artificial stock markets: the efficient market hypothesis and the rational expectations hypothesis, *Journal of Economic Behavior & Organization* 49, 217–239.

Chen, Wun-Hua, Jen-Ying Shih, and Soushan Wu, 2006, Comparison of support-vector machines and back propagation neural networks in forecasting the six major Asian stock markets, *IJEF* 1, 49.

Chiang, W.-C., T. L. Urban, and G. W. Baldridge, 1996, A neural network approach to mutual fund net asset value forecasting, *Omega* 24, 205–215.

Chiarella, C., R. Dieci, and X.-Z. He, 2009, Heterogeneity, Market Mechanisms, and Asset Price Dynamics, *Handbook of Financial Markets: Dynamics and Evolution (North-Holland, Elsevier)* Chapter 5, 277–344.

Christoffersen, Peter F., and Francis X. Diebold, 2003, Financial Asset Returns, Direction-of-Change Forecasting, and Volatility Dynamics, *Management Science* 52, 1273–1287.

Christoffersen, Peter F., Francis X. Diebold, Roberto S. Mariano, Anthony S. Tay, and Yiu Kuen Tse, 2006, Direction-of-Change Forecasts Based on Conditional Variance, Skewness and Kurtosis Dynamics: International Evidence, *Journal of Financial Forecasting* 1, 1–22.

Cont, R, 2001, Empirical properties of asset returns: stylized facts and statistical issues, *Quantitative Finance* 1, 223–236.

Cont, Rama, 2005, Volatility Clustering in Financial Markets : Empirical Facts and Agent-Based Models, *A Kirman & G Teyssiere (eds.): Long memory in economics* 21.

Creamer, Germán, and Yoav Freund, 2010, Automated trading with boosting and expert weighting, *Quantitative Finance* 10, 401–420.

De Grauwe, P., 2010, The Scientific Foundation of Dynamic Stochastic General Equilibrium (DSGE) Models, *Public Choice* 144, 413–443.

De Martino, A., I. Giardina, A. Tedeschi, and M. Marsili, 2004, Generalized minority games with adaptive trend-followers and contrarians, *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics* 70, 2–5.

Del Degan, M. A., 2012, Analysis of peak oil with focus on Norwegian oil production .

Eklund, Anders, Thomas E. Nichols, and Hans Knutsson, 2016, Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates, *Proceedings of the National Academy of Sciences* 113, 7900–7905.

Elsenbroich, Corinna, 2011, Explanation in Agent-Based Modelling: Functions, Causality or Mechanisms?, *Journal of Artificial Societies and Social Simulation* 15, 1.

European Commission, 2014, EU Crude Oil Imports.

Evstigneev, I. V., T. Hens, and K. R. Schenk-Hoppé, 2009, Emergence of cooperation and organization in an evolutionary game, *Handbook of Financial Markets: Dynamics and Evolution (North-Holland, Elsevier)* chapter 9, 507–566.

Fama, Eugene F., 1970, Efficient Capital Markets: A Review of Theory and Empirical Work, *The Journal of Finance* 25, 383–417.

Fama, Eugene F., 1991, Efficient Capital Markets: II, *The Journal of Finance* 46, 1575–1617.

Fama, Eugene F., 1998, Market efficiency, long-term returns, and behavioral finance, *Journal of Financial Economics* 49, 283–306.

Fama, Eugene F., and Kenneth R. French, 1993, Common risk factors in the returns on stocks and bonds, *Journal of Financial Economics* 33, 3–56.

Fama, Eugene F., and Kenneth R. French, 2009, Luck Versus Skill in the Cross Section of Mutual Fund Return, *Journal of Finance* 65, 1915–1947.

Farmer, J D, A Gerig, F Lillo, and H Waelbroeck, 2012, How efficiency shapes market impact, *Quantitative Finance* (http://arxiv.org/abs/1102.5457).

Farmer, J Doyne, 2002, Market force, ecology and evolution, *Industrial and Corporate Change* 11, 895–953.

Farmer, J Doyne, and Duncan Foley, 2009, The economy needs agent-based modelling, *Nature* 460, 685–686.

Feng, Ling, Baowen Li, Boris Podobnik, Tobias Preis, and H. Eugene Stanley, 2012, Linking agent-based models and stochastic models of financial markets, *Proceedings of the National Academy of Sciences* 109, 8388–8393.

Fernández-Delgado, Manuel, Eva Cernadas, Senén Barro, and Dinani Amorim, 2014, Do We Need Hundreds of Classifiers to Solve Real World Classification Problems?, *J. Mach. Learn. Res.* 15, 3133–3181.

Fievet, L., Z. Forro, P. Cauwels, and D. Sornette, 2015, A general improved methodology to forecasting future oil production: Application to the UK and Norway, *Energy* 79, 288–297.

Fievet, Lucas, and Didier Sornette, 2017, Decision Trees Unearth Return Sign Correlation in the S&P 500, *Quantitative Finance* .

Forró, Zalán, Peter Cauwels, and Didier Sornette, 2012, When games meet reality: is Zynga overvalued?, *The Journal of Investment Strategies* 1, 119–145.

French, Ken, 2012, Data Library.

Galla, Tobias, and J. Doyne Farmer, 2013, Complex dynamics in learning complicated games, *Proceedings of the National Academy of Sciences* 110, 1232–1236.

Gary, Michael Shayne, and Robert E Wood, 2010, Mental models, decision rules, and performance heterogeneity, *Strategic Management Journal* 32, 569–594.

Ghonghadze, Jaba, and Thomas Lux, 2016, Bringing an elementary agent-based model to the data: Estimation via GMM and an application to forecasting of asset price volatility, *Journal of Empirical Finance* 37, 1–19.

Goetzmann, William N., and Massimo Massa, 2002, Daily Momentum and Contrarian Behavior of Index Fund Investors, *The Journal of Financial and Quantitative Analysis* 37, 375–389.

GOV.UK, 2014, Oil and gas: field data.

Greiner, Alfred, Willi Semmler, and Tobias Mette, 2011, An Economic Model of Oil Exploration and Extraction, *Computational Economics* 40, 387–399.

Grossman, S. J., and J. E. Stiglitz, 1980, On the impossibility of informationally efficient markets, *The American Economic Review* 70, 293–408.

Guresen, Erkam, Gulgun Kayakutlu, and Tugrul U. Daim, 2011, Using artificial neural network models in stock market index prediction, *Expert Systems with Applications* 38, 10389–10397.

Hamilton, James Douglas, 1994, *Time Series Analysis*, first edition (Princeton University Press).

Hansen, Peter Reinhard, 2005, A Test for Superior Predictive Ability, *Journal of Business & Economic Statistics* 23, 365–380.

Harras, G., and D. Sornette, 2011, How to grow a bubble: A model of myopic adapting agents, *Journal of Economic Behavior and Organization* 80, 137–152.

Harrison, M. J., and D. M. Kreps, 1978, Speculative Investor Behavior in a Stock Market with Heterogeneous Expectations, *The Quarterly Journal of Economics* 92, 323–336.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman, 2001, *The Elements of Statistical Learning*, volume 1 of *Springer Series in Statistics* (Springer New York Inc., New York, NY, USA).

Hommes, C H, 2002, Modeling the stylized facts in finance through simple nonlinear adaptive systems, *Proceedings of the National Academy of Sciences* 99, 7221–7228.

Hommes, C .H., 2006, Heterogeneous agent models in economics and finance, in *Handbook of Computational Economics (Elsevier B.V.), Edited by Leigh Tesfatsion and Kenneth L. Judd*, volume 2, 1109–1186.

Hommes, C. H., and F. Wagener, 2009, Complex Evolutionary Systems in Behavioral Finance, *Handbook of Financial Markets: Dynamics and Evolution (North-Holland, Elsevier)* chapter 4, 217–276.

Höök, Mikael, and Kjell Aleklett, 2008, A decline rate study of Norwegian oil production, *Energy Policy* 36, 4262–4271.

Hsu, Po-Hsuan, Yu-Chin Hsu, and Chung-Ming Kuan, 2010, Testing the predictive ability of technical analysis using a new stepwise test without data snooping bias, *Journal of Empirical Finance* 17, 471–484.

Hsu, Po-Hsuan, and Chung-Ming Kuan, 2005, Reexamining the Profitability of Technical Analysis with Data Snooping Checks, *Journal of Financial Econometrics* 3, 606–628.

Hsu, Po-Hsuan, Mark P Taylor, and Zigan Wang, 2016, Technical trading: Is it still beating the foreign exchange market?, *Journal of International Economics* 102, 188–208.

Huang, C., D. Yang, and Y. Chuang, 2008, Application of wrapper approach and composite classifier to the stock trend prediction, *Expert Systems with Applications* 34, 2870–2878.

Hubbert, Marion King, 1956, Nuclear energy and the fossil fuels, *Drilling and Production Practice* 36.

Hyafil, Laurent, and Ronald L Rivest, 1976, Constructing optimal binary decision trees is NP-complete, *Information Processing Letters* 5, 15–17.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani, 2014, *An Introduction to Statistical Learning: With Applications in R* (Springer Publishing Company, Incorporated).

Jefferies, P., M. L. Hart, P. M. Hui, and N. F. Johnson, 2001, From market games to real-world markets, *The European Physical Journal B - Condensed Matter and Complex Systems* 20, 493–501.

Jegadeesh, Narasimhan, and Sheridan Titman, 1993, Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency, *The Journal of Finance* 48, 65–91.

Jiang, Zhi Qiang, Wei-Xing Zhou, Didier Sornette, Ryan Woodard, Ken Bastiaensen, and Peter Cauwels, 2010, Bubble diagnosis and prediction of the 2005-2007 and 2008-2009

Chinese stock market bubbles, *Journal of Economic Behavior and Organization* 74, 149–162.

Johansen, A., and D. Sornette, 2000, The Nasdaq crash of April 2000: Yet another example of log-periodicity in a speculative bubble ending in a crash, *The European Physical Journal B-Condensed …* 328, 319–328.

Johansen, Anders, Olivier Ledoit, and Didier Sornette, 2000, Crashes as critical points, *International Journal of Theoretical and Applied Finance* 3, 219–255.

Johnson, N.F., D. Lamper, P. Jefferies, M. L. Hart, and S. Howison, 2001, Application of multi-agent games to the prediction of financial time series, *Physica A* 299, 222–227.

Kaizoji, T., M. Leiss, A. Saichev, and D. Sornette, 2015, Super-exponential endogenous bubbles in an equilibrium model of rational and noise traders, *Journal of Economic Behavior and Organization* 112, 289–310.

Kyle, A S, 1985, Continuous Auctions and Insider Trading, *Econometrica* 53, 1315–1336.

Laherrère, Jean, 2002, Forecasting future production from past discovery, *International Journal of Global Energy Issues* 18, 218–238.

Laherrère, Jean, and Didier Sornette, 1998, Stretched Exponential Distributions in Nature and Economy: “Fat Tails” with Characteristic Scales, *Physical Journal B-Condensed Matter and Complexe Systems* 2, 525–539.

Lamper, D, S D Howison, and N F Johnson, 2002, Predictability of large future changes in a competitive evolving population, *Phys Rev Lett* 88, 17902.

Lebaron, Blake, 2006, Agent-based Financial Markets: Matching Stylized Facts With Style, *Post Walrasian Macroeconomics Beyond the Dynamics Stochatic General Equilibrium Model* 221–238.

Ledoit, Oliver, and Michael Wolf, 2008, Robust Performance Hypothesis Testing with the Sharpe Ratio, *Journal of Empirical Finance* 15, 850–859.

Leung, Mark T, Hazem Daouk, and An-Sing Chen, 2000, Forecasting stock indices: a comparison of classification and level estimation models, *International Journal of Forecasting* 16, 173–190.

Lillo, F, D Farmer, and R Mantegna, 2003, Master curve for price impact function, *Nature* 421, 129–130.

Lo, Andrew W, 2002, The Statistics of {Sharpe} Ratios, *Financial Analysts Journal* 58, 36–52.

Lo, Andrew W, 2012, Adaptive Markets and the New World Order, *Financial Analysts Journal* 68, 18–29.

Loken, Eric, and Andrew Gelman, 2017, Measurement error and the replication crisis, *Science* 355, 584–585.

Lux, Thomas, 1995, Heard Behaviour, Bubbles and Crashes, *The Economic Journal* 105, 881–896.

Lux, Thomas, 2012, Estimation of an agent-based model of investor sentiment formation in financial markets, *Journal of Economic Dynamics and Control* 36, 1284–1302.

Lux, Thomas, and Michele Marchesi, 1999, Scaling and criticality in a stochastic multi-agent model of a financial market, *Nature* 397, 498–500.

Lux, Thomas, and Michele Marchesi, 2000, Volatility Clustering in Financial Markets: a Microsimulation of Interacting Agents, *International Journal of Theoretical and Applied Finance* 03, 675–702.

Lynch, Michael C, 2002, Forecasting oil supply : theory and practice, *Quarterly Review of Economics and Finance* 42, 373–389.

Malevergne, Yannick, Vladilen F Pisarenko, and Didier Sornette, 2005, Empirical Distributions of Log-Returns : between the Stretched Exponential and the Power Law?, *Quantitative Finance* 5, 379–401.

Malkiel, Burton G, 2003a, The Efficient Market Hypothesis and Its Critics, *Journal of Economic Perspectives* 17, 59–82.

Malkiel, Burton Gordon, 2003b, *A Random Walk down Wall Street: The Time-tested Strategy for Successful Investing* (New York: W.W. Norton).

Maymin, Philip Z., 2011, Markets are efficient if and only if P = NP, *Algorithmic Finance* 1, 1–11.

Milgrom, Paul, and Nancy Stokey, 1982, Information, trade and common knowledge, *Journal of Economic Theory* 26, 17–27.

Miller, Edward M., 1977, Risk, uncertainty, and divergence of opinion, *The Journal of Finance* 32, 1151–1168.

Mills, Terence C., and Raphael N. Markellos, 2008, *The Econometric Modelling of Financial Time Series* (Cambridge University Press).

Mills, Terence C., and Raphael N. Markellos, 2009, Financial Economics, Non-linear Time Series in, in *Encyclopedia of Complexity and Systems Science*, 3435–3448.

Murray, James W, and Jim Hansen, 2013, Peak Oil and Energy Independence : Myth and Reality, *Eos, Transactions American Geophysical Union* 94, 245–246.

Newey, Whitney K, and Kenneth West, 1994, Automatic Lag Selection in Covariance Matrix Estimation, *Review of Economic Studies* 61, 631–653.

Niederhoffer, Victor, and M F M Osborne, 1966, Market Making and Reversal on the Stock Exchange, *Journal of the American Statistical Association* 61, 897–916.

Norwegian Petroleum Directorate, 2014, Factpages.

Patzelt, Felix, and Klaus Pawelzik, 2013, An Inherent Instability of Efficient Markets, *Scientific Reports* 3.

Pearson, Karl, 1900, On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling, *Philosophical Magazine Series 5* 50, 157–175.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, 2011, Scikit-learn: Machine Learning in Python, *Journal of Machine Learning Research* 12, 2825–2830.

Pesaran, M. Hashem, and Allan Timmermann, 1992, A Simple Nonparametric Test of Predictive Performance, *Journal of Business & Economic Statistics* 10, 461–465.

Phillips, Peter C B, Yangru Wu, and Jun Yu, 2011, Explosive behavior in the 1990s Nasdaq: When did exuberance escalate asset values?, *International Economic Review* 52, 201–226.

Plackett, R L, 1983, Karl {Pearson} and the Chi-Squared Test, *International Statistical Review* 51, 59.

Politis, Dimitris N, and Joseph P Romano, 1992, A circular block-resampling procedure for stationary data, *Exploring the limits of bootstrap* 263–270.

Raberto, Marco, Silvano Cincotti, Sergio M. Focardi, and Michele Marchesi, 2001, Agent-based simulation of a financial market, *Physica A: Statistical Mechanics and its Applications* 299, 319–327.

Romano, Joseph P., and Michael Wolf, 2005a, Exact and Approximate Stepdown Methods for Multiple Hypothesis Testing, *Source Journal of the American Statistical Association* 100, 94–108.

Romano, Joseph P., and Michael Wolf, 2005b, Stepwise multiple testing as formalized data snooping, *Econometrica* 73, 1237–1282.

Romano, Joseph P, and Michael Wolf, 2016, Efficient computation of adjusted p-values for resampling-based stepdown multiple testing, *Statistics & Probability Letters* 113, 38–40.

Rubinstein, Ariel, 1997, *Modeling Bounded Rationality* (The MIT Press).

Samadiou, E, E Zschischang, D Stauffer, and Thomas Lux, 2007, Agent-based models of financial markets, *Reports on progress in physics* 70, 409–450.

Satinover, J. B., and D. Sornette, 2007, "Illusion of control" in time-horizon minority and Parrondo Games, *European Physical Journal B* 60, 369–384.

Satinover, J B, and D Sornette, 2009, Illusory versus genuine control in agent-based games, *Eur. Phys. J. B* 67, 357–367.

Satinover, J. B., and D. Sornette, 2012a, Cycles, determinism and persistence in agent-based games and financial time-series I, *Quantitative Finance* 12, 1051–1064.

Satinover, J. B., and D. Sornette, 2012b, Cycles, determinism and persistence in agent-based games and financial time-series II, *Quantitative Finance* 12, 1064–1078.

Scheinkman, J. A., and W. Xiong, 2003, Overconfidence and Speculative Bubbles, *Journal of Political Economy* 111, 1183–1220.

Senneret, Marc, Yannick Malevergne, Patrice Abry, Gerald Perrin, and Laurent Jaffres, 2016, Covariance versus Precision Matrix Estimation for Efficient Asset Allocation, *IEEE Journal of Selected Topics in Signal Processing* .

Sharpe, William F., 1964, Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk, *The Journal of Finance* 19, 425.

Sharpe, William F, 1994, The Sharpe Ratio, *Portfolio Management* 21, 49–58.

Shiller, Robert J, 1981, Do Stock Prices Move Too Much to be Justified by Subsequent Changes in Dividends?, *The American Economic Review* 71, 421–436.

Simon, Herbert A., 1955, A Behavioral Model of Rational Choice, *The Quarterly Journal of Economics* 69, 99–118.

Smith, James L, 1980, A Probabilistic Model of Oil Discovery, *The Review of Economics and Statistics* 62, 587–594.

Sornette, Didier, 2003, *Why Stock Markets Crash: Critical Events in Complex Financial Systems* (Princeton University Press).

Sornette, Didier, 2004, *Critical Phenomena in Natural Sciences, Chaos, Fractals, Self-organization and Disorder: Concepts and Tools*, second edition (Heidelberg).

Sornette, Didier, 2014, Physics and Financial Economics (1776-2014): Puzzles, Ising and agent-based models, *Reports on progress in physics* 77.

Sornette, Didier, and Peter Cauwels, 2015, Financial bubbles: mechanisms and diagnostics, *Review of Behavioral Economics* 2, 279–305.

Sornette, Didier, and Anders Johansen, 1997, Large financial crashes, *Physica A* 245, 14.

Sornette, Didier, Anders Johansen, and Jean-Philippe Bouchaud, 1995, Stock market crashes, Precursors and Replicas, *J.Phys.I France 6* 6, 167–175.

Sornette, Didier, and Wei-Xing Zhou, 2006, Importance of Positive Feedbacks and Overconfidence in a Self-Fulfilling Ising Model of Financial Markets, *Physica A* 370, 704–726.

Subrahmanyam, Avanidhar, 2008, Behavioural Finance: A Review and Synthesis, *European Financial Management* 14, 12–29.

Sullivan, Ryan, Allan Timmermann, and Halbert White, 1999, Data-Snooping, Technical Trading Rule Performance, and the Bootstrap, *The Journal of Finance* 54, 1647–1691.

Tay, Francis E. H., and Lijuan Cao, 2001, Application of support vector machines in financial time series forecasting, *Omega* 29, 309–317.

Timmermann, Allan, and Clive W. J. Granger, 2004, Efficient market hypothesis and forecasting, *International Journal of Forecasting* 20, 15–27.

Tversky, Amos, and Daniel Kahneman, 1974, Judgment under Uncertainty: Heuristics and Biases, *Science* 185, 1124–1131.

Villa, S., and F. Stella, 2014, A continuous time Bayesian network classifier for intraday FX prediction, *Quantitative Finance* 14, 2079–2092.

White, Halbert, 2000, A Reality Check for Data Snooping, *Econometrica* 68, 1097–1126.

Wiesinger, J., D. Sornette, and J. Satinover, 2012, Reverse Engineering Financial Markets with Majority and Minority Games Using Genetic Algorithms, *Computational Economics* 41, 475–492.

Xue-shen, Sui, Qi Zhong-ying, Da-ren Yu, Hu Qing-hua, and Zhao Hui, 2007, A Novel Feature Selection Approach Using Classification Complexity for SVM of Stock Market Trend Prediction, *International Conference on Management Science and Engineering* 1654–1659.

Yen, Stephane Meng Feng, Ying Lin Hsu, and Yi Long Hsiao, 2015, Can hedge fund elites consistently beat the benchmark? A study of portfolio optimization, *Asia Pacific Management Review* 20, 275–284.

Zhang, Qunzhi, 2013, *Disentangling Financial Markets and Social Networks: Models and Empirical Tests*, Ph.D. thesis, ETH Zurich.

Zhang, Qunzhi, D. Sornette, and J. Satinover, 2013, Mixed-game virtual stock markets combining minority, delayed minority, majority and dollar agent-based models, *working paper ETH Zurich* .

Zhang, Yi-cheng, 1999, Toward a theory of marginally efficient markets, *Physica A: Statistical Mechanics and its Applications* 269, 30–44.

Zunino, Luciano, Massimiliano Zanin, Benjamin M Tabak, Darío G Pérez, and Osvaldo A Rosso, 2009, Forbidden patterns, permutation entropy and stock market inefficiency, *Physica A: Statistical Mechanics and its Applications* 388, 2854–2864.

Curriculum Vitae

2013-2017	Doctoral student at the Chair of Entrepreneurial Risks, ETH Zurich
2012-2013	Software Engineer at ELCA, Zurich
2010-2012	Master student in Physics, ETH Zurich
2009-2010	McGill University: Exchange year
2007-2010	Bachelor student in Physics, EPF Lausanne
2002-2007	Ecole Européenne de Luxembourg, „Baccalauréat Européen“
2000-2002	Lycée Français de Conakry (Guinea)
1996-2000	Ecole Française d’Accra (Ghana)
1989	Born in Bruxelles, Belgium