On Discrete Investment Rules for Financial Markets

A thesis presented

by

Spyros Skouras

to

The Department of Economics

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Economics

European University Institute

Florence, Italy

July 1999

To my family and friends

Contents

Preface

This collection of essays is the outgrowth of what a few years ago was an interest in the intersection of finance and learning theory. As I thought about how investors do and should learn, the normative issue became increasingly fascinating; this was significant because I found that the literature on forecasting provided a convenient framework in which to nest issues of 'normative learning'. I have now realised that these forecasting and learning problems can both be addressed in a unified, perhaps even elegant, manner within the framework of statistical decision theory.

Most of the thesis was written in Florence, although substantial parts were written during an excessively cold winter in Luxembourg and an excessively hot summer in Athens. Florence was the middle ground in many ways and I feel very lucky to belong in the creative group of people who have roamed Tuscany absorbed in abstract thoughts. Of those in this group who lived at the right time for me to meet, I would like to thank James Dow, Alex Guembel, Grayham Mizon, Mark Salmon, Paul Schure, Johnny Toro and Spyros Vassilakis for helpful discussions. I must also acknowledge useful comments from a number of people who I occasionally imposed my questions on, either by accosting them after their seminars at the Institute or by bombarding them with an inordinate quantity of emails. These include Buz Brock, Neil Ericsson, Blake LeBaron, David Hendry, Charles Manski, Hashem Pesaran and Steve Satchell.

Though I must admit that my feelings towards this have fluctuated in the past, I am now thankful to those who encouraged me to study for a Ph.D.; particularly Giorgia Giovannetti whose suggestion of a cunning substitute for Hawaii had a decisive impact on what followed. Marcia Gastaldo deserves many thanks for help and encouragement throughout the job-hunting period. The Budokan Kenkyukai is thanked for being a constant reminder that there is life after the Institute; Vasilia, Lonkers, Lichnen, LaLa and Lia are thanked for providing all kinds of essential support without which this Ph.D. would *really* not have been possible. Lia is also thanked for providing the name for one of the concepts introduced in the first Chapter of the thesis. While I share her wish that it will make her famous, I am sure she has an easier way of achieving that goal!

Of course, special thanks must go to Ramon Marimon and Søren Jo-

hansen for supervising my work. Ramon Marimon for motivating me to do my best on this project, for a beer in New York and for the important role he played in teaching me to think about economic problems. Søren Johansen for helping me realise that many of the economic problems I was thinking about could be translated into statistical problems and for providing an implicit assurance that I could dive into them as deeply as I liked without fearing I might find myself alone.

Athens,
September 1999

Post Scriptum

Special thanks should be extended to Hashem Pesaran and Allan Timmermann for agreeing to be External Examiners and for providing numerous comments that led to what (I hope they will agree) is a substantial improvement in the content of this Thesis.

Cambridge,
June 2000

Introduction and overview

The starting point for this thesis was the cumulating empirical evidence that discrete investment rules[1] such as technical trading, chartist or market timing rules are very widely used (Allen & Taylor 1992, Frankel & Froot 1990) and also seem to 'perform well' (Levich & Thomas 1993, Brock, Lakonishok & LeBaron 1992, Neftci 1991). Even as I began studying the implications of this evidence, a second wave of papers provided mixed evidence on performance (Brown, Goetzmann & Kumar 1998, Kho 1996, Lee & Mathur 1996), a more careful analysis of data mining considerations (Sullivan, Timmermann & White 1999) as well as microstructure issues (Ready 1997, Knez & Ready 1996). Although empirical evidence is not uniform across studies, it is evident that the empirical properties of these investment rules has attracted substantial interest.[2] Indeed, the controversy surrounding their performance is a consequence of its implications for some important edifices of financial economics. Surprising as this may seem, *ad hoc* investment rules with provocative names (Sylla 1993) which not so long ago were deemed too 'unscientific' to be worth any attention, are now offering insights to issues which lie at the heart of both empirical finance and financial theory.

An implication of this empirical evidence which has been relatively easy to accept is that standard econometric models for financial series must be modified if they are to account for the empirically observed statistical properties of discrete investment rules (Gencay, Ballochi, Dacorogna, Olsen & Pictet 1999, Neely, Weller & Dittmar 1996, Brock et al. 1992). However, whether this is a worthwhile task is a separate issue and as a matter of fact it remains a task which has not been undertaken. It is widely accepted that most econometric models are mis-specified and that it is only worth accounting for any *particular type* of mis-specification if it seems useful to do so for some particular purpose. Hence, the importance that is attached to this line of research is a derivative of each economists' opinion on whether

[1] The definition of discrete investment rules used here is restricted to mappings from a state space (e.g. past prices of shares) to a set of usually binary (long/short) investment decisions; this is broad enough to include many of the most popular technical trading and market timing rules analysed in the literature and used in practice.

[2] An exhaustive bibliography for academic research on technical analysis can be found at http://www.iue.it/Personal/Researchers/Skouras/tabiblio.htm.

the properties of investment rules are of any interest *per se*.

A number of reasons for which this may be the case are scattered in the literature and collected in this thesis. On the positive side, it may be argued that the existence of simple investment rules that perform well contradicts popular interpretations of 'market efficiency' (such as that of Malkiel (1992)) according to which any publicly available investment rule should be 'worthless'.[3] Furthermore, the existence of such rules might also provide grounds on which to build an explanation of why investment rules are so widely used. On the normative side, good performance indicates that the public information used by these rules should also be used by agents attempting to implement optimal investment decisions. This implies that it will be useful to extend empirical analysis of optimal asset allocation decisions to account for the role of conditioning information, which is a challenging road for future research.[4]

The main aim of this thesis is twofold. Firstly, the development of techniques which can deliver improved investment rules and improved estimates of their performance. Secondly, the formalisation of the implications of investment rules' performance for models of financial markets and individual behaviour. To achieve these aims, certain analytical results and computational tools have been developed and are presented, where possible, with a generality that makes them useful in a broad range of applications. Partial answers to a number of questions of somewhat independent interest are provided, including the following:

- Under what conditions is the use of technical trading rules 'rational' in the sense of being consistent with expected utility maximisation'?
- How should the 'performance' of discrete investment rules be measured?
- What are the implications of the empirically observed performance of discrete investment rules for the efficiency with which financial markets use public information?
- How can we account for the widespread use of discrete investment rules?
- What is the relation between a discrete investment rule and a model for forecasting financial series? In particular, how can a good rule be used to improve the design of a forecasting model and how can a good forecasting model be used to design a good rule?

[3] According to Malkiel (1996, p. 161) 'Technical strategies are usually amusing, often comforting but of no real value'.

[4] Direct evidence on the importance of time-varying investment opportunities for investment decisions has now been independently obtained using a variety approaches (Brandt 1999, Campbell & Viceira 1999, Goldbaum forthcoming).

- How should parametric models for discrete investment rules be estimated?

The results are presented as follows. In Chapter 1, we show that an expected utility maximising investor solving a single period cash-asset allocation problem will use technical trading rules if he is risk neutral (and usually not otherwise). This is because unless very restrictive assumptions are imposed on the behaviour of returns, only a risk neutral agent has an optimal decision rule that is a discrete investment rule (as defined here). Furthermore, we prove that ranking technical trading rules according to the preferences of a risk neutral investor is identical to ranking them according to the preferences of any mean-variance investor. On the basis of these observations, we propose the use of this agent's expected utility as a measure of the profitability of investment rules. These considerations provide an expected utility justification for measuring a model's 'economic value' by expected profits (Pesaran & Timmermann 1995, Leitch & Tanner 1991, Breen, Glosten & Jagannathan 1989).

While economic value may be a reasonable theoretical measure of the performance of a single investment rule, we are often interested in considering the *empirical* performance of a *class* of investment rules, such as a class serving as a proxy for 'commonly used technical trading rules'. We show empirically that the standard approach of using an estimate of the performance of a few rules selected from a class as a measure for the performance of the class as a whole entails substantial data mining problems. We therefore propose the use of the *optimal* rule in terms of economic value as a measure of the performance of the class. However, since the optimal rule is itself unknown, it must be estimated. The tools we develop for the estimation of investment rules which 'maximise the economic value of predictability' are presented in Chapters 4 and 5 though they are referred to in all previous Chapters. In this first Chapter they applied (in their most rudimentary guise) to show that an *estimated optimal technical trading rule* can lead to quite powerful empirical conclusions.

In particular, market efficiency can be tested using its implications for the behaviour of investment rules. According to Latham's (1986) definition of market efficiency (Equilibrium-efficiency), a necessary condition for efficiency with respect to past prices is that prices do not affect agents' decisions. Using a long series of returns on the Dow Jones Industrial Average and a parametric model of solutions to agents' objective functions based on technical trading rules, we find that mean-variance agents *will use technical trading rules* that condition on past prices, suggesting that Equilibrium-efficiency is rejected (for sufficiently low transaction costs). While a number of authors have alluded to a link between technical trading rule performance and market efficiency (Bessembinder & Chan 1998, Hudson, Dempsey & Keasey 1996, Cornell & Dietrich 1978), this is the first study which uses an

explicit definition of efficiency that makes a precise link with the substantial theoretical research in this area.

Having established that an appropriate measure of an investment rule's performance is the utility obtained by a risk neutral user, in Chapter 2, we focus on building forecasting models for financial returns which will perform well when used by such an investor. One purpose of this is to develop the link between forecasting and financial decision making; at the same time, we are able to formulate reasonable procedures with which investors might estimate both optimal forecasting and discrete investment rule models. Our analysis takes the perspective of the recent literature on 'forecasting under asymmetric loss' (Christofferson & Diebold 1996, Weiss 1996, Diebold & Mariano 1995, West 1994) in the context of which it has been shown that standard statistical measures of a forecast's performance (such as likelihood or squared errors) do not usually provide a good measure of the forecast's value for solving financial decision problems (Kandel & Stambaugh 1996, West, Edison & Cho 1993, Leitch & Tanner 1991). We also propose an estimation method that performs well as a 'Risk Neutral Forecasting' estimator both analytically and using simulations, providing a route for estimating 'maximally economically significant' forecasting models and investment rules.

In Chapter 3 we return to the issue of 'rationalising' the use of investment rules which we introduced in Chapter 1. We argue that if an agent restricts his investment behaviour to following rules with a binary long/short structure, he may be able to learn which types of rules in this class are good more effectively. If the benefits from easier learning outweigh the costs of this form of bounded rationality, it is easy to see why even a risk averse agent may come to use a rule of this form. While the results we report in this Chapter are very preliminary, they provide a sketch of an argument for how a learning model can be used to explain individual behaviour. They also indicate that certain behavioural rules may not be dominated by standard econometric techniques and therefore that a learning theory with behavioural foundations (e.g. Easley & Rustichini 1999) might have normative value for investors. This is contrasted to the prevalent use of learning theory which is in models of collective decision-making aiming to explain aggregate phenomena (Sargent 1993).

Chapter 4 contains the statistical results that are utilised throughout the thesis. They are presented in the form of consistency proofs for certain estimators of the sign of mean regressions as it turns out that such estimators can also be used to estimate optimal risk neutral forecasting and discrete investment rule models. However, we show that the scope of such estimators extends well beyond this particular problem: there is a large class of interesting decision problems which are solved by the sign of a regression mapping, either because the set of available actions is finite or because, as is the case with the objective functions we work with in previous chapters,

only a finite set of actions can be optimal. Furthermore, certain optimal stopping problems, calibration exercises (e.g. that of Eisenhart (1939)) and inverse regressions (e.g. those of Hendry & Ericsson (1991) and Ericsson, Hendry & Mizon (1998)) can be solved or facilitated by estimating the sign of some regression mapping. We propose and compare a number of estimators that are particularly attractive when robustness to misspecification is of particular importance. The research leading to this paper is still work in progress, the goal of which is to find estimators for the sign of regression mappings with tractable properties. This could be used, for example, to conduct inference about the behaviour of discrete optimal decisions.

Finally, in Chapter 5 we consider the computational obstacles encountered when we attempt to estimate optimal discrete investment rules or related forecasting models. It turns out these are substantial so we design an algorithm that performs this task well and may also be useful in optimising rugged step-function landscapes which arise in other applications.

One of the broad conclusions drawn from this thesis is that it is interesting to formulate parametric models of individual decisions and then estimate them using real data. This is a difficult but worthwhile task that can reveal aspects of the structure of the agent's environment which are most important to the agent. For example, estimators for investment rules may reveal structure in financial series that is relevant for decisions yet remains undetected by standard econometric methods. Observation of this type of empirical regularity should lead to an improved understanding of investment behaviour and - by extension - the properties of financial markets in general.

Interpreted as normative learning models, these estimators provide behavioural rules that fit neatly into the framework of statistical decision functions (see Wald (1971)) when (as is the case here) the learning agent has no control over the stochastic process affecting his objective function.[5] The same applies to most of the recent research on forecasting under asymmetric loss which, if built on expected utility foundations, can be interpreted as precisely such a learning problem. This link has not yet been explored but is a fascinating road for future research which could be nested in the powerful theoretical framework of statistical decision theory. This might pave the road for the empirical study of more general financial and economic decision problems and a body of research on what might be called the econometrics of optimal decision-making, or simply **Decisionmetrics**.

While I hope that this thesis will be read from front to back, it has been written so that each Chapter can stand alone independently as a 'potential journal article'. I apologise to the reader for any repetitiveness and notational inconsistencies that this may cause.

[5] Interestingly, Neyman (1938) referred to the use of such rules as 'inductive behaviour' before the development of the theory of what we now call 'rational behaviour'.

Chapter 1

Financial Returns and Efficiency as seen by an Artificial Technical Analyst

SUMMARY

We introduce trading rules which are selected by an artificially intelligent agent who learns from experience - an Artificial Technical Analyst. These rules restrict the data-mining concerns associated with the use of 'simple' technical trading rules as model evaluation devices and are good at recognising subtle regularities in return processes. The relationship between the efficiency of financial markets and the efficacy of technical analysis is investigated and it is shown that the Artificial Technical Analyst can be used to provide a quantitative measure of market efficiency. We estimate this measure on the DJIA daily index from 1962 to 1986 and draw implications for the optimal behaviour of certain classes of investors. It is also shown that the structure of technical trading rules commonly used is consistent with utility maximisation for risk neutral agents and in a myopic sense even for risk-averse agents.

1.1 Introduction

In the last few years, increasing evidence that technical trading rules can detect non-linearities in financial time series has renewed interest in technical analysis (see e.g. LeBaron 1998*a*, Levich & Thomas 1993, Brock et al. 1992, Neftci 1991). Based on this evidence, much research effort has also been devoted to examining whether trading rules can be used to evaluate and

[0]This paper is forthcoming in the *Journal of Economic Dynamics and Control* and has received the 1998 Grad Student Paper Prize from the *Society of Computational Economics*.

create improved time-series and theory driven returns models (Hudson et al. 1996, Kho 1996, Gencay 1996).

In this type of empirical studies, the term technical analysis is used to refer to the practice of investing according to well-known technical trading rules. However, in other areas of financial theory technical analysis is sometimes defined to be any conditioning of expectations on past prices (Brown & Jennings 1989, Treynor & Ferguson 1985). Indeed, noisy asymmetric information models in which rational agents condition on past prices because they reflect (a noisy signal of) private information provide one explanation of why technical analysis is observed. Unfortunately, theoretical models which lead to conditioning consistent with the *precise form* of observed technical trading rules are currently unavailable: there is as yet no positive model of investment behaviour consistent with the behaviour of real Technical Analysts.

This paper models Technical Analysts as agents whose actions are *de facto* consistent with observed technical trading rules. Our terminology therefore will be consistent with that of researchers examining empirical aspects of technical analysis and as such will be more narrow than that of theoretical models. The objective however is not the modelling of Technical Analysts *per se*; rather, it is to use our model of a Technical Analyst to derive a more sophisticated approach to examining the statistical properties of trading rules. It is somewhat surprising that some studies have found a single arbitrarily selected rule to be 'effective' over long periods (e.g. Brock *et al.*) given that real Technical Analysts use different rules in different times and in different markets. In order to truly evaluate the effectiveness of technical analysis *as implemented* we need a model of how analysts adapt to the market environment.

We provide such a model by introducing Technical Analysts who are artificially intelligent agents (see e.g. Marimon et al. 1990). In Section 2 technical analysis is introduced in the simple case where agents are fully informed and circumstances in which it may be a rational activity are derived. This is a necessary step for the development of a model of a Technical Analyst who learns from his environment - an Artificial Technical Analyst. This agent chooses amongst technical trading rules and his actions are the outcome of an explicit decision problem which formalises the loose notion of what it means for a rule to be 'good' or optimal (examples of informal uses of this term are Allen & Karjalainen (1999), Neely et al. (1996), Pictet et al. (1996), Taylor (1994), Allen & Phang (1994), Chiang (1992) and Pau (1991)). As we discuss in Section 3, this formalisation is important because it indicates that an explicit measure of rule optimality can and should be derived from a specific utility maximisation problem and that a rule which is optimal for investors with different objective functions will not typically exist.

A standard application of artificially intelligent agents is to design them

so that their actions can reveal interesting aspects of the environment in which they are placed (see e.g. Sargent 1993, pp.152-160). In this vein, we will use our Artificial Technical Analysts to reveal certain regularities in financial data. In particular, in Section 4 they will be used to characterise financial series as in Brock *et al.*, and we will show that they can provide sharper characterisations than those based on simple technical trading rules.

In Section 5 we propose a quantitative notion of weak market efficiency which admits measurement of the *degree* to which a market is weak form efficient. Furthermore, we show that such measurements can be based on the returns obtained by the Artificial Technical Analyst and illustrate with a measurement of the efficiency of the Dow Jones Industrial Average index, interpreted as a proxy for the market portfolio. This is a step in addressing the relationship between market efficiency and the profitability of technical analysis, an issue that has appeared in some of the theoretical literature (e.g. Brown & Jennings 1989) but is absent from many empirical investigations of technical analysis.

We conclude this paper with a synopsis of our results. The main contribution is the introduction of the Artificial Technical Analyst, a theoretical tool which is shown to be useful in providing corroborating empirical evidence for the view many Technical Analysts hold of econometric returns models and market efficiency: that the models are inadequate for the purpose of making investment decisions and that markets are not always efficient.

1.2 Technical analysis with full information

At some level of abstraction, technical analysis can be viewed as a methodology for selecting decision rules which determine (conditional on certain events) whether a position in a financial asset will be taken and whether this position should be positive or negative. One important difference between an analyst and a utility maximising investor is that the decision rules the analyst uses do not specify the magnitude of the positions he should take.

These observations lead us to the following description of technical analysis:

Definition 1.1 ***Technical Analysis*** *is the selection of a function* $d : \mathbf{I}_t \rightarrow \Omega$ *which maps the information set* $\mathbf{I}_t$ *at time* t *to a space of investment decisions* $\Omega \equiv \{-s, 0, l\}$, $l, s \geq 0$ *specifying the size of short, neutral and long positions respectively.*

This assumption captures the main structure of technical analysis as traditionally practiced[1] (see e.g. Kaufman (1978) and Murphy (1986) - the

[1] In recent years investors have used increasingly sophisticated ways of conditioning

so called 'classics') in that the magnitude of these positions is not specified as a function of $\mathbf{I}_t$. As mentioned in the introduction, this is the definition of technical analysis used in the economics literature on empirical properties of trading rules. However, a different definition is used in the literature on financial equilibrium with asymmetric information where any agent who conditions on past prices is a technical analyst (e.g. Brown & Jennings (1989), Treynor & Ferguson (1985)). We are not apologetic about this: the reason we do not allow a more general definition is because we wish to examine the properties of *observed* trading rules which at first sight seem very different to the investment behaviour we would expect from utility maximising agents. Despite their restrictive nature, these rules include the 'market timing' rules which have been extensively studied in the literature, particularly since Merton (1981) developed an equilibrium theory applicable to their evaluation.

The information sets on which these rules condition on are usually the realisations of some random variable such as prices, volatility measures, or the volume of trading (Blume, Easley & O'Hara 1994) of an asset. Our focus here will be on rules which are defined on the realisation of a truncated history of past prices $\mathbf{P}_t \equiv \{P_t, P_{t-1}, P_{t-N+1}\} \in \mathbb{R}_+^N$ (this is expedient only in that it simplifies the exposition - our proofs do not depend on it).

Assumption 1.1 *The information set $\mathbf{I}_t$ on which Technical Analysis is based is restricted to a truncated history of past prices $\mathbf{P}_t$. For notational convenience, we use $E_t(\cdot)$ to refer to $E(\cdot|\mathbf{P}_t)$.*

Technical Trading Rules and Rule Classes

Technical analysts change the mappings $\mathbf{P}_t \rightarrow \Omega$ (or *technical trading rules*) they use and not all of them use the same rules. Nevertheless, rules used are often very similar and seem to belong to certain families of closely related rules, such as the 'moving average' or 'range-break' family (see Brock *et al.*). These families belong to even larger families, such as those of 'trend-following' or 'contrarian' rules (see for example Chan, Jegadeesh & Lakonishok (1996)). Whilst it is difficult to argue that use of any particular rule is widespread, certain families are certainly very widely used. The distinction between a rule and a family is formalised as a distinction between technical trading rules and technical trading rule classes.

Definition 1.2 *A **Technical Trading Rule Class** is a parametric function*

$$D : \mathbf{P}_t \times B \rightarrow \Omega$$

their decisions on past prices. Such conditioning is technical analysis in the sense of, for example, Treynor & Ferguson (1985) but not in the 'traditional' sense which is the focus of this study.

which for each parameter c in a parameter space $B \subseteq \mathbb{R}^k$ maps past prices into investment positions.

Definition 1.3 *A **Technical Trading Rule** is an element of a technical trading rule class (indexed by a parameter $c \in B$)*

$$d_t \equiv D(\mathbf{P}_t, c) : \mathbf{P}_t \rightarrow \Omega$$

which determines a unique investment position as a function of past prices.[2]

Notice that the trivial technical trading rule which always has a position as large as possible in the asset ($d_t = l$), is identical to the 'Buy-and-Hold' strategy which specifies that once some quantity of an asset is bought, this quantity should not be altered and the value of the investment should be allowed to evolve without any form of intervention.

1.2.1 Technical analysis and rationality

Having defined the main concepts required to describe technical analysis, we now attempt to identify investors who would choose to undertake this activity. In particular, we find restrictions on preferences of utility maximising investors which guarantee they behave *as if* they were Technical Analysts. The purpose of this is to clarify the meaning of 'optimal technical analysis' in a full information setting. This concept can then be applied to the more interesting case where optimal behaviour must be *learned* from experience.

Consider the following simple but classic investment problem. An investor with utility function U can invest in two assets: A risky asset paying interest $R_{t+1} = \frac{P_{t+1} - P_t}{P_t}$ (random at time t) and a riskless asset (cash) which pays no interest. He owns wealth W_t and his objective is to maximise expected utility of wealth at the end of the next period by choosing the proportion of wealth a invested in the risky asset. We will assume that institutional constraints permit borrowing and short-selling at the riskless interest rate (assumed to be zero) but only to a finite extent determined as a multiple of current wealth, so that a is constrained to lie in the compact set $[-s, l]$. Expectations E_t are formed on the basis of past prices $\mathbf{P}_t$, as dictated by $A1$.

Formally, when there are no transaction costs, the problem is:

$$\begin{aligned} & \max_{a \in [-s,l]} E_t U(W_{t+1}) \\ s.t.\ W_{t+1} &= aW_t(1 + R_{t+1}) + (1 - a)W_t, \end{aligned}$$

[2] Notice that any set of rule classes $\{D_i\}$ can be seen as a meta-class itself, where the parameter vector $[i, c]$ determines a specific technical trading rule. Also, note that the way we have defined D allows us to accommodate the case where the parameter vector is $[c(\mathbf{P}_t), c]$ i.e. contains estimated parameters.

or equivalently,

$$\max_{a\in[-s,l]} E_t U(W_t(1+aR_{t+1})), \tag{1.1}$$

the solution a^* of which can be denoted as a function:

$$a^* : \mathbf{P}_t \to [-s, l]\,, \tag{1.2}$$

satisfying:

$$a^* \equiv \arg\max_{a\in[-s,l]} E_t U(W_t(1+aR_{t+1})). \tag{1.3}$$

In a general utility maximisation setting therefore, the proportion of wealth invested is a function that depends on the conditional distribution of returns on past prices, utility and wealth. Since by definition a technical trading rule may only take three distinct values and cannot be a function of wealth, rational investment behaviour differs from technical analysis because it cannot be described by a trading rule unless special assumptions are made about the nature of the objective function. The inclusion of transaction costs in the objective function cannot easily reverse this fact. Whilst it can make a^* a discontinuous function, a^* will be limited to take values in Ω only under unrealistic joint assumptions on utility, the conditional distribution of returns and the nature of the transaction costs.[3]

In the special case that investors are risk-neutral however, their optimisation problem has a bang-bang solution which depends on $sign\left[E_t\left(R_{t+1}\right)\right]$. Denoting a^*_{rn} the solution to the risk-neutral investor's problem and assuming that when expected returns are zero the investor stays out of the market ($E_t\left(R_{t+1}\right) = 0 \Leftrightarrow a^*_{rn} = 0$), then:

$$a^*_{rn} : \mathbf{P}_t \to \{-s, 0, l\}\,,$$

which *is* consistent with the definition of a technical trading rule.

We have therefore shown that a risk-neutral investor conditioning on past prices will choose technical trading rules of the form typically encountered and as defined here. Other agents may also do so, but only under very specific assumptions on the underlying series that are unlikely to be realistic. As we shall see in Section 5, technical trading rules are behavioural rules that have attractive properties for certain classes of *boundedly rational* risk averse agents. These properties suggest an explanation of why technical trading rules are used by risk-averse agents.

Our current result is summarised in the following proposition:

[3] Except in the trivial case where large transaction costs make the buy-and-hold strategy an optimal decision for a rational investor.

Proposition 1.1 *For all past prices, wealth and short-selling or borrowing constraints, the solution of (1.1) for a risk-neutral investor is a technical trading rule.*

A direct corollary of this proposition is that expected utility maximisation and technical analysis *are compatible*. This allows us to define a technical analyst as an expected utility maximising investor:

Definition 1.4 *A **Technical Analyst** is a risk-neutral investor who solves:*

$$\max_{d \in D} d(\mathbf{P}_t) \cdot E_t(R_{t+1}), \tag{1.4}$$

where D is a technical trading rule class.

We will let $R^d_{t+1} \equiv d(\mathbf{P}_t) \cdot R_{t+1}$ denote the returns obtained by a technical analyst who uses a rule d. Clearly, different trading rules lead to different expected returns.

1.3 Artificial Technical Analysts

Having specified what is meant by technical analysis in the case of full information, let us assume henceforth that the technical analyst does not know $E_t(R_{t+1})$ but has a history of observations of $\mathbf{P}_t$ on the basis of which he must decide his optimal action at time t. This is a similar amount of information to that possessed by econometricians and hence a technical analyst learning his optimal actions in this environment can be modeled as an *artificial intelligent agent* in the sense of Sargent (1993) or Marimon et al. (1990). In this section, we will propose a 'reasonable' model for how a Technical Analyst might try to learn his optimal actions. The term *Artificial Technical Analyst* will refer to an agent who attempts to make the decisions of a Technical Analyst but who is equipped with an explicit 'reasonable' mechanism for learning optimal actions rather than rational expectations.

1.3.1 Parametrising Analysts' learning

Typically, the learning technique of an artificial agent is similar to that of an econometrician. In the context of this paper, the agent might learn the solution to (1.4) by selecting a forecasting model for $E_t(R_{t+1})$ from some parametric class (e.g. GARCH-M). This selection is typically made according to some standard statistical estimation method such as least squares or quasi-maximum likelihood. The artificial adaptive agent then chooses an action which would be optimal if $E_t(R_{t+1})$ were in fact what the forecasting model predicts.

Whilst for some applications this may be a useful approach, we are forced to depart from this methodology somewhat due to the fact that there is significant empirical evidence that statistical fitness criteria can be misleading when applied to decision problems such as that of the Technical Analyst. For example, Kandel & Stambaugh (1996) show that statistical fitness criteria are not necessarily good guides for whether a regression model is useful to a rational (Bayesian) investor. Taylor (1994) finds that trading based on a channel trading rule outperforms a trading rule based on ARIMA forecasts chosen to minimise in-sample least squares because the former is able to predict sign changes more effectively than the latter.[4] More generally, Leitch & Tanner (1991) show that standard measures of predictor performance are bad guides for the ability of a predictor to discern sign changes of the underlying variable.[5]

These empirical considerations suggest that any reasonable model of analysts' learning must take his loss function into account. One way to achieve this would be to create a Bayesian Artificial Technical Analyst but this would require specification of a prior on the conditional distribution of returns which might be very difficult. Instead, we confine the Artificial Technical Analyst to a frequentist perspective and calculate an estimator $\hat{d}$ for the trading rule solving (2.3) given a specified technical trading rule class D. This can be viewed as a *decision theoretic* approach in that learning about the underlying stochastic environment is replaced with the task of learning about the optimal decision.[6]

One convenient estimator of the solution to (1.4) is given by the solution to the following simple in-sample analogue to (1.4):

[4] Under some assumptions on the underlying processes, the technical analyst is interested primarily in the sign of R_{t+1} rather than in its actual value. In particular, the value of R_{t+1} is irrelevant for his decision problem if $sign(R_{t+1})$ is known with certainty. More generally, we will show in Chapter 2 that what matters is the prediction of a certain quantile of the distribution of $sign(R_{t+1})$. Satchell & Timmermann (1995) show that, without severe restrictions on the underlying series, least square metrics are not directly related to sign-based metrics.

[5] A number of studies of technical trading implicitly or explicitly assume away the possibility that there exists a non-monotonic relationship between the accuracy of a prediction in terms of a metric based on least squares and a metric based on the profit maximisation. Examples are Allen & Taylor (1990), Curcio & Goodhart (1991) and Arthur et al. (1996) who reward agents in an artificial stockmarket according to traditional measures of predictive accuracy. When the assumption is made explicit its significance is usually relegated to a footnote, as in Allen & Taylor (1990), fn. p.58, *"our analysis has been conducted entirely in terms of the accuracy of chartist forecasts and not in terms of their profitability or 'economic value' although one would expect a close correlation between the two"*. As we have argued however, the preceding statement is unfounded and results of such studies should be interpreted with caution.

[6] Brandt (1999) recently proposed a related approach for estimating optimal investment decisions but which is applicable only to risk-averse agents since it relies on non-parametric estimation of smooth first order conditions.

$$\max_{\mathbf{c}\in B} \Sigma_{i=t-m}^{t-1} D(\mathbf{P}_i, \mathbf{c}) \cdot R_{i+1}, \tag{1.5}$$

where B is some parameter space determining the choice set of technical trading rules. If

$$\frac{1}{m}\Sigma_{i=t-m}^{t-1} D(\mathbf{P}_i, c) R_{i+1}$$

converges uniformly to

$$E\left\{D(\mathbf{P}_t, c) \cdot R_{t+1}\right\}$$

almost surely as $m \to \infty$, under certain regularity conditions[7] it is also the case that the maximum on B of the former expression converges to the maximum of the latter almost surely as $m \to \infty$, so this estimator is consistent.

Next we choose D so as to impose some restrictions on the solution to (1.5) that allow regularities of the in-sample period to be captured. Having no theory to guide us on how to make this choice it is reasonable to use empirically observed rule classes D^o. We may therefore define an Artificial Technical Analyst as follows:

Definition 1.5 *An **Artificial Technical Analyst** is an agent who solves:*

$$\max_{c\in B} \Sigma_{i=t-m}^{t-1} D^o(\mathbf{P}_i, c) R_{i+1}, \tag{1.6}$$

*where D^o is an **empirically observed** trading rule class.*

Turn now to an example illustrating the mechanics of this agent which will be useful in subsequent sections.

Example: Choices of the Artificial Technical Analyst determine the optimal Moving Average rule.

The moving average rule class is one of the most popular rule classes used by technical analysts and has appeared in most published studies of technical analysis. For these reasons, we will use it to illustrate how an Artificial Technical Analyst might operate if this is the set from which he chooses rules. Let us begin with a definition[8] of this class:

[7] Assumptions 4.2-4.3 of Chapter 4 are such conditions.

[8] As defined, the moving average class is a slightly restricted version of what Brock *et al.* (1991, 1992) refer to as the "variable length moving average class" (in particular, the restriction arises from the fact that the short moving average is restricted to have length 1).

Definition 1.6 *The **Moving Average technical trading rule class** $MA(P_t, c_t)$ is an (empirically observed) technical trading rule class such that:*

$$MA(\mathbf{P}_t, c) \equiv \left[\begin{array}{c} l \text{ if } P_t \geq (1+\phi)\frac{\Sigma_{i=0}^{m} P_{t-i}}{m+1} \\ 0 \text{ if } (1-\phi)\frac{\Sigma_{i=0}^{m} P_{t-i}}{m+1} \leq P_t \leq (1+\phi)\frac{\Sigma_{i=0}^{m} P_{t-i}}{m+1} \\ -s \text{ if } P_t < (1-\phi)\frac{\Sigma_{i=0}^{m} P_{t-i}}{m+1} \end{array} \right], \quad (1.7)$$

$$\text{where } \mathbf{P}_t \equiv [P_t, P_{t-1}, ..., P_{t-N}],$$

$$c \equiv (m, \phi),$$

$$B \equiv \{\mathbf{M}, \Phi\},$$

$$\mathbf{M} \equiv \{1, 2, ...M\} \text{ is the 'memory' of the rule,}$$

$$\Phi \equiv \{\phi : 0 \leq \phi \leq \Phi\} \text{ is the 'filter' (or bandwidth) of the rule.}$$

Now if $D^o = MA(\mathbf{P}_t, c)$, (1.6) becomes:

$$\max_{m \in \mathbf{M}, \phi \in \Phi} \Sigma_{i=t-N}^{t-1} MA(\mathbf{P}_i, m, \phi) R_{i+1}. \quad (1.8)$$

An Artificial Technical Analyst learning technical trading rules by solving (1.8) uses N daily observations of $MA(\mathbf{P}_i, m, \phi)R_{i+1}$ derived from $N + M + 1$ observations of prices P_t. Let us assume $M = 200$ (a 'standard' value for the longest moving average of interest), $N = 250$ (approximately a year's worth of data is used for estimation), that $s = l = 1$ (position size cannot exceed current wealth) and $\Phi = 0.02$ (another 'standard'). How do the trading rules used by this Artificial Technical Analyst behave when $\mathbf{P}_t$ are draws from the Dow Jones Industrial Average index? Figure 1.1 plots the parameters (m, ϕ) indexing trading rules estimated according this method[9], where the parameters were estimated using a daily rolling recursion from t=1/6/1962 to 31/12/1986.[10]

[9] Φ was discretised to $\Phi = \{0, 0.005, 0.01, 0.015, 0.02\}$. This discretisation allowed us to solve (7) by trying all $\dim(\mathbf{N}) \cdot \dim(\Phi) = 1000$ points composing the solution space in each of the 6157 recursions. More sophisticated search methods could lead to more intelligent Artificial Technical Analysts but such niceties do not seem necessary when D^o is as narrow as it is in this example.

[10] This data corresponds to the third subperiod used by Brock *et al.* and to most of the data used by Gencay (1996).

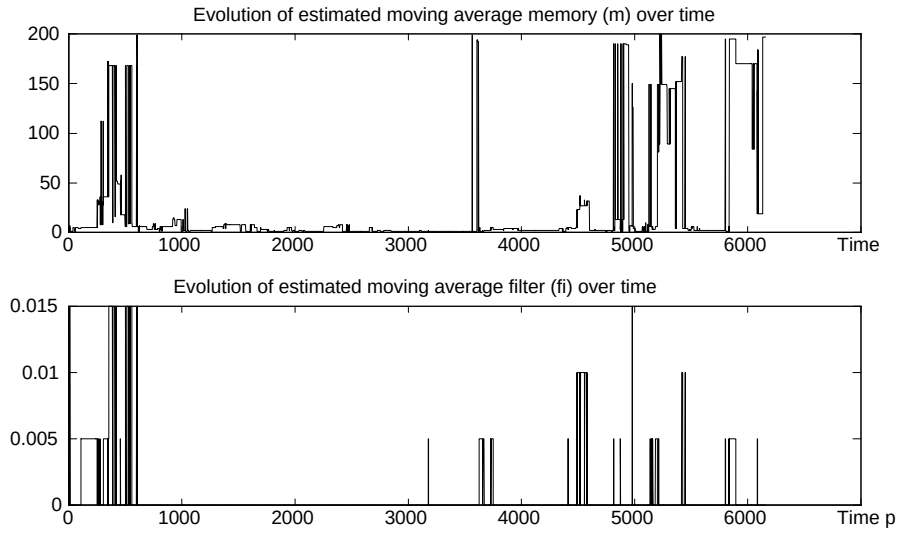

Figure 1.1. Evolution of estimated parameters $\widehat{m}$ and $\widehat{\phi}$ respectively, for 6157 daily observations of the DJIA index.

The sharp discontinuities observed in the sequence of estimated parameters suggests that the very different rules have similar performance; this coupled with sampling uncertainty or non-stationarity in the series might be sufficient to produce the observed behaviour. They also strongly suggest there is additional structure in the series which a 'more intelligent' Artificial Technical Analyst (one with a more sophisticated learning mechanism) might be able to identify. For example, some method (e.g. an averaging rule) for reducing the impact of sampling uncertainty might be helpful

1.4 Artificial Technical Analysts and the distribution of returns in financial markets.

Much of the literature on technical trading rules has asked whether popular types of rules will yield returns in excess of what would be expected under some null hypothesis on the distribution of returns (e.g. Brock et al. 1992, Levich & Thomas 1993, Neftci 1991, etc.). By answering this question in the affirmative, it has been possible to reject these hypotheses; the performance of trading rules thus serves as an interesting type of specification test. The rules considered are typically selections that are meant to be 'representative' for a plausible and widely used rule class. However, the fact that they are chosen according to non-rigorous and often implicit criteria makes results drawn from them subject to standard data-mining criticisms which diminish their forcefulness. This is a problem that is avoided if the rules considered are the choices of an Artificial Technical Analyst which are by construction explicit and can be expected to be robust with respect to

reasonable variations in the agent's design.[11] In Section 4.1 we show with reference to the study of Brock *et al.* that such criticisms are *not trivial.* In Section 4.2 we show that the Artificial Technical Analyst can be used to construct more powerful tests of hypotheses regarding the correct specification of models for returns. This is because the Artificial Technical Analyst is good at recognising regularities in the series that are usually not the primary focus of attention - though they often should be since they affect decisions and hence have an important economic interpretation.

1.4.1 The variation of returns across rules

Clearly, we must rely on empirical evidence to see whether rule returns are correlated closely enough within a class to justify using a few rules as proxies for the behaviour of the class as a whole.[12] Figure 1.2 below shows the returns accruing to *each* rule belonging to the moving average class if it were applied on the DJIA index throughout the period considered (1962 - mid 1986).

[11] Of course, a degree of arbitrariness remains in our selection of the rule class to be tested. However, we have already mentioned that there exists much stronger empirical evidence on the basis of which to choose a rule class than for any specific rule. The arbitrariness involved in the specification of learning schemes may be an additional problem, but overall such choices are generally considered to be robust in other applications (e.g. econometrics) and are certainly more robust than choices of arbitrary rules. Of course, such robustness claims can only be evaluated empirically; in the Appendix we provide 'Auxiliary Results' that indicate the Artificial Technical Analyst we work with in this paper is indeed robust.

[12] That this is the case is suggested by Brock *et al*, who write that *'Recent results in LeBaron (1990)* [now available in LeBaron (1998*b*)] *for foreign exchange markets suggest that the results are not sensitive to the actual lengths of the rules used. We have replicated some of those results for the Dow index'*, p1734, fn. The "recent results" to which Brock *et al.* refer are a plot of a certain statistic of 10 rules. Apart from the fact that 10 rules constitute a small sample, the minimum statistic is almost half the size of the maximum statistic - so it is not entirely clear that these results support the claim made.

On the other hand, it is hard to criticise the conclusions of Brock *et al.* since the rules they chose happened to generate returns which were slightly *lower* than the average of the class they considered.

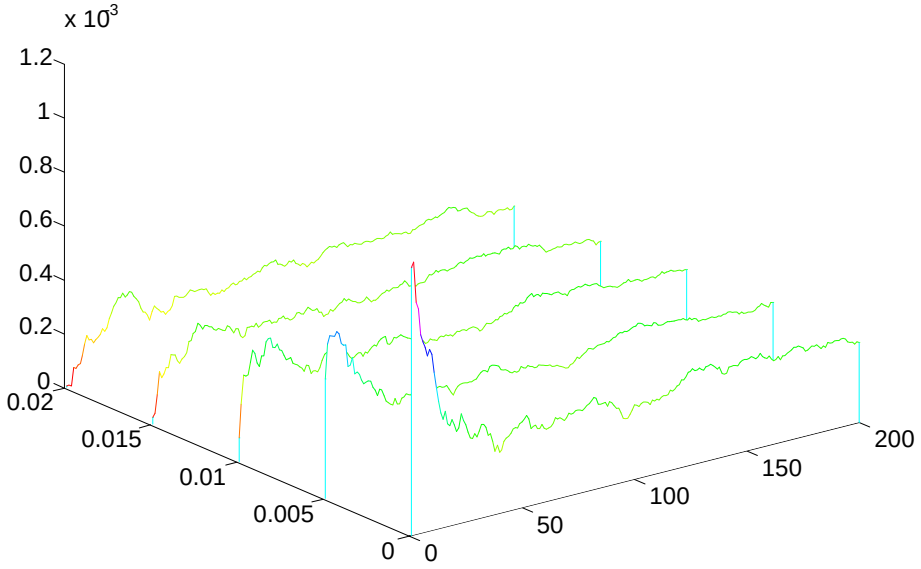

Figure 1.2. Mean returns (over 6157 periods) of each rule as a function of its parameters m and ϕ.

This figure illustrates the inadequacies of the *ad hoc* approach whereby specific rules are used as proxies for the distribution of expected returns of a whole class. Notice in particular two highly prominent facts evident in this figure.

Firstly, while all rules earn positive returns, the mean return of the best rule in the class is **1270** times larger than that of the worst rule. Since the means are taken from samples with more than 6000 observations, it is unlikely that sampling uncertainty can account for these differences.[13] We must conclude that returns accruing to rules *within the same class* vary very significantly.

Secondly, the expected returns of rules display significant variation even *within local regions* of the parameter space. The best rule is the three period moving average with no filter $MA(3,0)$ and the worst is the four period moving average with a 2 percent filter $MA(4,0.02)$. This is important because most researchers choose to calculate returns for a few rules sampled evenly from the space of all rules, reflecting the unfounded implicit assumption that rules are 'locally' representative, i.e. that parameter choices are robust.

Taken together these two observations imply that *ad hoc* rules cannot be the basis for convincing tests of specifications of models for returns, as previous studies have suggested. Rules must be selected according to an explicit procedure (such as that of the Artificial Technical Analyst) which

[13] A bootstrap simulation might be used to evaluate this conjecture.

is justifiable on theoretical grounds.

1.4.2 Artificial Technical Analysts' rules generate 'above average' profits

We now show that 'representative' mean returns are likely to be smaller than those obtained by an Artificial Technical Analyst and hence will be less informative about the structure in returns processes. The reason for this is that a well designed Artificial Technical Analyst who chooses a technical trading rule from some class should have learned to make a better-than-average choice of d. Imposing the use of a 'representative' rule by an Artificial Technical Analyst would be the analogue of estimating a parametric model for a time series by choosing the parameters which have average rather than least squared errors.

The following table serves to empirically confirm this reasoning. Utilising some of the information in Table V of Brock, Lakonishok & LeBaron (1991)[14] it indicates that the results reported there on the basis of various fixed rules are much weaker than those which can be drawn by using the time-varying estimated rule $\widehat{d}_t$ derived in Section 3.1.1. That the rules chosen by Brock *et al.* are 'average' across the space of all rules can be verified by inspection of their position in Figure 1.2.

In this table, t-ratios are reported for the hypotheses of equality between unconditional returns and returns conditional on a buy and a sell position for various types of technical trading rules (if returns are independently distributed, these differences should be zero). The table indicates that all t-ratios are much higher for the estimated rule we have developed than for the simple rules used by Brock *et al.* This implies that the Artificial Technical Analyst is 'better' at detecting the dependencies in the DJIA series since it can be used to reject the hypothesis that the returns are independent draws from some distribution with much greater confidence than that offered by Brock *et al.*'s analysis.[15] We expect that the Artificial Technical Analyst's rule will be equally powerful as a specification test for particular models for the time dependency in returns, such as those considered by Brock *et al.* (AR, GARCH-M, EGARCH). However, we must leave confirmation of this conjecture for future research.

As we have already mentioned, it is inevitable that the creation of an Artificial Technical Analyst involves certain design decisions which themselves raise data-mining issues. We have noted however that such an agent is likely to be robust with respect to such decisions. Furthermore, there is often a very natural way to make these decisions in the context of a specific

[14] Note that Brock et al. (1992) reproduce only a part of this table.

[15] The table also contains information which is sufficient to show that the Cumby & Modest (1987) test for market timing would, if the riskless interest rate were zero, confirm the ability of a technical analyst learning optimal rules to conduct market timing.

Rule	N(buy)	N(sell)	Buy	Sell	Buy>0	Sell>0	buy-sell
Always long	6157	0	0.00023	-	-	-	-
Brock et al (m_t, ϕ_t)							
(50,0)	3468	2636	0.00036	-0.0004	0.5167	0.4879	0.00041
			(0.90076)	(-1.16108)			(1.78607)
(50,0.01)	2782	1985	0.00053	0.00003	0.5230	0.4861	0.00049
			(1.64014)	(-0.70959)			(1.89872)
(150,0)	3581	2424	0.00037	-0.00012	0.5205	0.4777	0.00049
			(0.94029)	(-1.49333)			(2.11283)
(150,0.01)	3292	2147	0.00035	-0.00018	0.5216	0.4742	0.00052
			(0.80174)	(-1.67583)			(2.13824)
(200,0)	3704	2251	0.00037	-0.00016	0.5173	0.4780	0.00053
			(0.92753)	(-1.64056)			(2.23379)
(200,0.01)	3469	2049	0.00038	-0.00018	0.5189	0.4763	0.00056
			(0.96907)	(-1.66579)			(2.26328)
Average			0.00037	-0.00011			0.00048
ATA $\{d(m_t^*, \phi_t^*)\}_{t=1}^{6157}$	3313	2650	0.00095	-0.00067	0.5337	0.4675	0.00162
			(3.95033)	(-4.57949)			(7.34848)

Table 1.1: The first column of this table indicates which rule is being used. The parameters in parentheses constitute specifications of the memory and filter of the moving average rule used. The second and third columns indicate the number of days in which the rule was long or short respectively and the fourth and fifth the mean return on those days. The sixth column reports the proportion of days on which the rules made profits conditional being long and the seventh column provides a similar statistic for days when they were short. T-statistics applying a test used in Brock et al. to check significance are provided for some estimates.

application. Here for example we have avoided most problems of this form by anchoring our selection of the trading rule class and the data set to the set of Brock *et al.*[16] Whilst the selection of the length of the optimisation period N is still under our control we set $N = 250$ on the *a priori* basis that it is the standard rounded approximation to the number of trading days in a year. In the Appendix we replicate some of our results for other values of N and thereby provide evidence that this choice has only a marginal effect on our results.[17]

In summary, the Artificial Technical Analyst provides a powerful and robust tool for examining the empirical properties of technical trading rules;

[16] Of course, that also means we inherit whatever data-mining criticisms may be levied on Brock *et al.* The results of Sullivan et al. (1999) suggest these may be substantial.

[17] This is hardly surprising given that this parameter operates exactly like the choice of sample size in a recursively estimated econometric model (which also does not usually affect results if - as in this case - it is reasonably large).

we therefore propose the use of Artificial Technical Analysts' rules as the basis for model specification tests. This decreases the probability of obtaining misleading results whilst at the same time delivering more powerful conclusions relative to those obtainable from an analysis of 'representative' rules (which are also difficult to identify).

1.5 Market efficiency and technical trading

It is often stated that 'If markets are efficient, then (technical) analysis of past price patterns to predict the future will be useless', (Malkiel 1992). In this section, we attempt to analyse the relationship between the efficiency of markets and the efficacy of technical analysis, with a view to a formal assessment of this statement.

Currently, only a model-specific notion of efficiency is available, deriving from successive refinements on the definition of Fama (1970). The latest element in this sequence of definitions is that of Latham (1986) according to which a market is *E-Efficient* ('E' for equilibrium) with respect to an information set if and only if its revelation to all market participants would leave both equilibrium prices and investment decisions unchanged. However, there seems to be little consensus as to what empirical properties an efficient market should display, partly due to the lack of an accepted equilibrium model for financial markets (LeRoy 1989, Fama 1991). Furthermore, market efficiency is only testable in the context of such a model.

We now propose a definition of efficiency which has the advantage that its testability does not hinge on the assumption of a specific equilibrium model, but rather on a model for the behaviour of at least *some* agents in the market. Our definition is a *necessary condition* for the market to be E-efficient if we accept any model in which some agents our modeled in this way. It is also consistent with the weak requirement implicit in almost all efficiency notions according to which conditioning on publicly available information does not increase utility. In its very weakest forms, this is interpreted as meaning that once transaction costs are included, no risk-averse agent can increase his utility by attempting to 'time' (i.e. forecast the direction of) the market. This statement is so weak that some authors (for example LeRoy 1989, p.1613 fn.) consider it to be non-testable. However, if we assume that the time series of prices is the market clearing equilibrium of an economy with a single risky asset or that we know the market portfolio and have a time series of prices for it, we show that the performance of technical trading rules can be used to construct precisely such a test. For simplicity, the analysis is restricted to the case where the risk-free interest rate is zero and past prices are the information set with respect to which we evaluate efficiency (weak-form).

The evaluation of financial market efficiency has an interesting role to

play in building useful dynamic asset pricing models. There is now a technology for determining broad conditions such a model must satisfy in order to be consistent with a given sample of data (Hansen & Jagannathan 1991, Hansen & Richard 1987). For some purposes it may be crucial that the model used also replicates the efficiency properties of actual markets (e.g. when the model will be used to address issues relating to the role of financial markets in allocating resources or to evaluate the investment performance of various types of agents). We can use the results of this section to judge whether a market is efficient and evaluate whether it is necessary to design models that replicate this feature. Whilst the relationship between technical analysis, market efficiency and dynamic asset pricing models is interesting, it lies beyond the scope of this paper and we must leave it for future research.

We will refer to the version of the efficient market hypothesis that we have alluded to as the Lack of Intertemporal Arbitrage (LIA) Hypothesis and discuss its implications for technical trading rules. It will become evident that this efficiency notion is formulated so that it is consistent with the idea that if technical analysis 'works', loosely speaking, then markets must be inefficient. In this sense it formalises the efficiency notion to which many empirical analyses of trading rule returns allude, yet typically leave undefined. A desired property efficiency notions have failed to deliver is a way of quantifying near efficiency. The following definition of LIA allows precisely such a quantification.

Definition 1.7 *The **Lack of Intertemporal Arbitrage (LIA) Hypothesis** holds for all investors with objective functions in some space $\mathcal{U}$ if their optimal decisions do not depend on past prices.*

The nature of the space $\mathcal{U}^e$ for which LIA holds can be viewed as a *measure of the **efficiency*** of a market. For example, as it increases, fewer agents (distinguished by their utility functions) find past prices useful and hence the market becomes more efficient (with respect to past prices). In particular if for two markets A and B we know that $\mathcal{U}_A^e \subset \mathcal{U}_B^e$, LIA provides a well defined sense in which market A is *less efficient* than B. Such comparisons are relevant if there exist markets which may be treated as separate on *a priori* grounds or if we wish to compare the efficiency of a single market during different time periods. Of course when neither $\mathcal{U}_A^e$ nor $\mathcal{U}_B^e$ is contained in the other, efficiency rankings may be harder, but nevertheless this definition gives us a language in which we can formalise the concept of near-efficiency.

Here we focus on the simple allocation decision (1.1) and therefore treat the space $\mathcal{U}$ as a way of imposing restrictions on the form U can take. LIA is confirmed for a class of agents solving (1.1) if knowledge of past prices does not affect their optimal actions. This is consistent with the idea that current prices reflect all information in past prices that might be of relevance. In

this case, LIA requires that for $U \in \mathcal{U}$, all $\mathbf{P}_t$ and all W_t:

$$\arg \max_{a \in [-s,l]} E\left\{U\left[W_t(1+aR_{t+1})|\mathbf{P}_t\right]\right\} = \arg \max_{a \in [-s,l]} E\left\{U\left[W_t(1+aR_{t+1})\right]\right\}. \tag{1.9}$$

This is implied by LIA when the only source of available information are past prices. It may be interesting to consider whether LIA holds when agents have access to auxiliary sources of information. While this would be beyond the scope of this paper it would be interesting since it would allow us to account for sources of predictability such as any publicly known time-varying risk-premia.

Clearly, LIA imposes restrictions on the joint distribution of returns and past prices. While it does not require these to be independent, it requires that knowledge of past prices does not affect an agent's optimal investment decision. For example, suppose only the third and higher order moments of the conditional distribution of returns depend on past prices; then in a market with mean-variance agents, actions will not be affected by knowledge of past prices (LIA holds) even though in a market without agents of this type, they will (LIA does not hold). Our definition implies that market efficiency is defined with respect to a class of agents and that the size of this class can therefore be interpreted as a measure of the *degree* of the market's efficiency. Formally, the degree of efficiency is determined by the form of the space:

$$\mathcal{U}^e \equiv \left\{ U \in \mathcal{U} : \begin{array}{l} \arg\max_{a \in [-s,l]} E\left\{U\left[W_t(1+aR_{t+1})\right]|\mathbf{P}_t\right\} \\ = \arg\max_{a \in [-s,l]} E\left\{U\left[W_t(1+aR_{t+1})\right]\right\} \end{array} \forall\, \mathbf{P}_t \right\}. \tag{1.10}$$

For a general equilibrium model of returns to be LIA efficient with respect to $\mathcal{U}^e$ it must be the case that agents with objective functions in $\mathcal{U}^e$ behave identically whether or not they know $\mathbf{P}_t$. A dual way of describing the degree of LIA efficiency is in terms of the *set of distributions $\mathcal{F}_\mathcal{U}$ for which LIA efficiency with respect to $\mathcal{U}$ holds.* Denoting a joint distribution of $(R_{t+1}, \mathbf{P}_t)$ as F, a marginal of R_{t+1} as F_R and a conditional of returns on prices as $F_R|\mathbf{P}$, this set is given by:

$$\mathcal{F}_\mathcal{U} \equiv \left\{ F : \begin{array}{l} \arg\max_{a \in [-s,l]} \int U\left[W_t(1+aR_{t+1})dF_R|\mathbf{P}\right] \\ = \arg\max_{a \in [-s,l]} \int U\left[W_t(1+aR_{t+1})\right] dF \end{array} \forall\, \mathbf{P}_t, U \in \mathcal{U} \right\}. \tag{1.11}$$

To decide *empirically* whether a market is efficient with respect to a particular set of objective functions $\mathcal{U}$ of interest, we can conduct the following hypothesis test:

$$H_0 : F \in \mathcal{F}_\mathcal{U} \quad \textbf{(LIA)}, \tag{1.12}$$

versus,

$$H_1 : F \notin \mathcal{F}_{\mathcal{U}} \text{ (Not LIA)}. \tag{1.13}$$

As we have already mentioned, we do not interpret a rejection of the null as an indication that the market cannot be modelled as an equilibrium of a standard model. Such a rejection would simply provide evidence against the particular notion of efficiency according to which prices should not be 'useful' to any agent (implicit, for example, in Malkiel's statement with which we opened this section). As we have discussed, this is a very weak notion of efficiency which in a well-defined sense is a necessary condition for E-efficiency.

1.5.1 Technical Trading Rules and LIA

The Artificial Technical Analyst provides a condition on rule returns for testing LIA

If for some market the true joint distribution of returns and the past prices F^* is known, we can check whether LIA holds for utility functions in $\mathcal{U}$, in some cases even *analytically*. If F^* is unknown, one test of the hypothesis described by (1.12-1.13) could be based on checking whether an estimated model for F^* is in $\mathcal{F}_{\mathcal{U}}$. Here we propose an alternative test based on the implications of the null hypothesis for the returns obtained by an Artificial Technical Analyst. This approach is based on the fact that a *sufficient* condition for LIA to be rejected is that technical trading is preferred over a position that is optimal when the agent does not condition on the history of past prices[18]. That this is the case is almost trivial, but is shown formally below:

Proposition 1.2 *Let a^* be an investment (as a fraction of wealth) which maximises an investor's (unconditional) expected utility when the (known) joint distribution of returns and past prices is given by F:*

$$a^* = \arg \max_{a \in [-s,l]} \int U\left[W_t(1 + aR_{t+1})\right] dF.$$

Assume this investor can increase his expected utility by investing this fraction a^* according to a trading rule d :

$$\max_{a \in [-s,l]} \int U\left[W_t(1 + a^* d\left(\mathbf{P}_t\right) R_{t+1})\right] dF > \max_{a \in [-s,l]} \int U\left[W_t(1 + a^* R_{t+1})\right] dF.$$

[18] This is stronger than the condition that the technical trading rule is preferred over the buy-and-hold strategy because this strategy in addition to not conditioning on prices does not condition on wealth either.

Then LIA does not hold, i.e.:

$$F \notin \mathcal{F}_{\mathcal{U}}.$$

Proof. See Appendix A. ■

We can use our proposition to reject LIA for a class $\mathcal{U}$ in which U belongs by showing that there exists a technical trading rule d with associated returns R^d_{t+1} such that *for all* a:

$$\int U\left[W_t(1+aR^d_{t+1})\right] dF > \int U\left[W_t(1+aR_{t+1})\right] dF, \tag{1.14}$$

Clearly, markets may be inefficient even when technical trading rules *do not* perform well, but we are not concerned with this here.

The Risk-Neutral Case

In order to implement a test of the null hypothesis (1.11), we must specify $\mathcal{U}$. Let us begin with the very narrow specification requiring $\mathcal{U}$ to contain only linear functions. Then (1.14) is equivalent to:

$$E(R^d_{t+1}) > E(R_{t+1}). \tag{1.15}$$

We can use the rules $\{d_t\}_{t=1}^{6157}$ which were optimal for the Artificial Technical Analyst in Section 3.1.1 and their corresponding returns to test the null hypothesis for the Dow Jones Industrial Average. Referring to the table below, we find that the probability LIA is not rejected is extremely low.

Returns	Mean	St. Dev.	Probability under Null
Always long	0.0002334	0.008459	-
ATA	0.000801	0.008335	8.887e-5

Table 1.2: Note that the last column was calculated conditional on the (false) assumption that returns of rules were normal i.i.d.

Hence we can conclude with great confidence that conditioning on past prices is desirable (LIA is rejected) for risk-neutral agents investing in the market for the DJIA index. The set $\mathcal{U}^e$ of utility functions with respect to which the market is efficient (in the LIA) sense cannot include any linear functions.

The Mean-Variance Case

Let us now consider whether the market is efficient with respect to quadratic utility functions. This may be the case if rules satisfying (1.15) involve a sufficiently larger variance than market returns to make them less desirable to a mean-variance agent. For example, LeRoy (1989) argues that:

> *...even though the existence of serial dependence in conditional expected returns implies that different formulas for trading bonds and stock will generate different expected returns, because of risk, these alternative trading rules are utility-decreasing relative to the optimal buy-and-hold strategies.*

To check whether this is the case, consider the class of quadratic utility functions

$$\left\{U : U(W) = aW^2 + bW + c, a \leq 0, \frac{-b}{2a} \geq W\right\}.$$

For all U in this class, if wealth W is a random variable, it is the case that $EU(W)$ is increasing w.r.t. $E(W)$, and decreasing w.r.t $Var(W)$.

The following proposition shows the somewhat surprising result that if there exists a rule that mean-dominates a long position, then it will also variance dominate it and hence LeRoy's statement is a logical impossibility in rather general circumstances. The proposition is crucial because it establishes broad circumstances in which the mean-variance case collapses to the risk-neutral case.

Proposition 1.3 *If the trading rule d_1 leads to a larger expected return than another rule d_2 with positive expected returns, i.e.*

$$E(d_1(\mathbf{P}_t) R_{t+1}) > E(d_2(\mathbf{P}_t) R_{t+1}) \geq 0,$$

and:

(a) The second rule is always long (i.e. is the 'Buy and Hold' strategy) and long positions are not smaller in absolute value than short positions

$$\begin{aligned} d_2(\mathbf{P}_t) &= l \text{ all } \mathbf{P}_t, \\ l &\geq s, \end{aligned}$$

or

(b) Trading rules have a binary structure and position sizes are symmetric

$$\begin{aligned} d_{1,}d_2 &\in \{-s, l\}, \\ l &= s, \end{aligned}$$

Then the returns from rule 1 have a smaller variance than the returns from rule 2

$$V(d_1(\mathbf{P}_t) R_{t+1}) < V(d_2(\mathbf{P}_t) R_{t+1}).$$

Proof. See Appendix. ■

This Proposition is useful for establishing the following corollary, but also because it provides a shortcut to ranking rules by performance in terms of *Sharpe Ratios* - an exercise attempted by many researchers and practitioners, supposedly as an alternative to ranking by mean returns. In the circumstances indicated however, Sharpe Ratios are inversely related to mean returns and such exercises are often redundant.

Corollary 1.1 *Consider a market in which the unconditional expected returns are positive* ($E(R_{t+1}) \geq 0$). *Suppose agents solving (1.1) have a tighter constraint on short positions than long positions* ($l \geq s$). *Suppose also that (1.15) holds so the market is* ***not*** *LIA-efficient for risk neutral agents. Then it will also* ***not*** *be LIA-efficient for mean-variance agents.*

Proof. See Appendix. ■

Under the very plausible conditions[19] of this corollary the risk-neutral case implies the mean-variance case and therefore mean-variance investors in the DJIA would find knowledge of past prices useful. Indeed, note that Table 1.2 confirms the implications of this Corollary.

The Risk-Averse Case

It is substantially more complicated to use technical trading rules to provide evidence that LIA does not hold when $\mathcal{U}$ is a concave class of functions. An exception arises when market and trading rule returns are normally distributed in which case second order stochastic dominance can be expressed as a function of means and variances.

Proposition 1.4 *Consider a market in which the unconditional returns are normally distributed with positive mean* ($E(R_{t+1}) \geq 0$). *Suppose agents solving (1.1) have a tighter constraint on short positions than long positions* ($l \geq s$). *Suppose also that (1.15) holds, so the market is* ***not*** *LIA-efficient for risk neutral agents and furthermore that this is the case for a trading rule with normally distributed returns. Then it will also* ***not*** *be LIA-efficient for risk-averse agents.*

Proof. See Appendix. ■

If the assumptions of Proposition 4 are not satisfied, we can reformulate the test for LIA in terms of a second order stochastic domination criterion of trading rule returns over market returns. This is shown in Proposition 5 below:

[19] But note that these results are only useful in a single risky asset/portfolio setting.

Proposition 1.5 *Let $f_{R_{t+1}}$, $f_{R^d_{t+1}}$ be the marginal densities of market and trading rule returns respectively. Define the function:*

$$M(\gamma) \equiv \int_{-\infty}^{\gamma} [f_{R_{t+1}}(x) - f_{R^d_{t+1}}(x)]dx. \tag{1.16}$$

Suppose agents solving (1.1) have a tighter constraint on short positions than long positions ($l \geq s$). If unconditional market returns are positive ($E(R_{t+1}) \geq 0$) and there exists a trading rule such the returns of which are distributed so that:

$$M(\gamma) \geq 0 \ \forall \ \gamma,$$

and $M(\gamma) > 0$ for at least one γ then the market is ***not*** *LIA-efficient for any risk-averse agent.*

Proof. See Appendix. ■

Clearly, if we reject the hypothesis that $M(\gamma) \leq 0$ which is an implication of LIA, we can also reject LIA. However, the available tests for this hypothesis are not generally applicable or involve huge computational costs (Tolley & Pope 1988).[20] This obstacle forces us to offer only an informal evaluation of whether LIA can be rejected. Such an evaluation can be conducted by inspecting a plot of the sample version $\hat{M}(\gamma)$ of $M(\gamma)$ for the returns of the Artificial Technical Analyst and the Dow Jones Industrial Average.

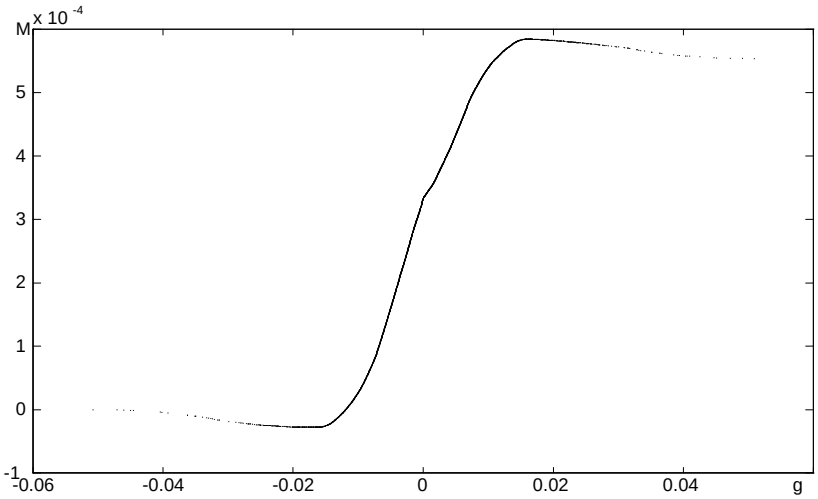

Figure 1.3: This is $\widehat{M}(\gamma)$ the sample version of $M(\gamma)$.

Observing Figure 1.3, we notice that for small γ, $M(\gamma) < 0$, indicating that the minimum returns from the Artificial Technical Analysts' rule

[20] After the completion of this Chapter a new technique appeared (Fisher, Wilson & Xu 1998) that may in fact resolve this problem.

resulted in smaller returns than the minimum market return. This implies that an agent with a utility function which greatly penalizes very low returns would prefer **not to use** the trading rule. Hence, even without taking account of sample uncertainty we are unable to reject LIA in the risk-averse case. Taking sample uncertainty into account using a formal statistical procedure cannot reverse this result, since it could only weaken the case against LIA. We conclude that there *are* risk-averse agents who may be indifferent to information contained in past prices so Malkiel's statement can be justified albeit in a narrow sense: The set of utility functions $\mathcal{U}^e$ with respect to which the DJIA is efficient *does not* include any mean-variance utility functions but includes at least *some* strictly concave utility functions.

1.5.2 Efficiency with Transaction Costs

We now define LIA for the case where an investor faces proportional symmetric transaction costs. In this situation, the set of distributions consistent with LIA is:

$$\mathcal{F}_{\mathcal{U}}^{\kappa} \equiv \left\{ F : \begin{array}{c} \arg\max_{a\in[-s,l]} \int U\left[W_t(1+a_tR_{t+1}) - \kappa_t\left|a_t - a_{t-1}\right|\right] dF_R|\mathbf{P}_t \\ = \arg\max_{a\in[-s,l]} \int U\left[W_t(1+a_tR_{t+1}) - \kappa_t\left|a_t - a_{t-1}\right|\right] dF \\ \forall\ \mathbf{P}_t, U \in \mathcal{U} \end{array} \right\} \tag{1.17}$$

where a_{-1} is the size of yesterday's position.

Consider an Artificial Technical Analyst choosing rules to maximise the type of objective function in (1.17). As before, we let $s = l = 1$ and the Artificial Technical Analyst uses rules given by (compare to (1.8)):

$$\max_{m\in\mathbf{M},\phi\in\Phi} \Sigma_{i=t-m}^{t-1}\left[MA(\mathbf{P}_i, m, \phi)R_{i+1} - \kappa\left|MA(\mathbf{P}_i, m, \phi) - a_{i-1}\right|\right] \tag{1.18}$$

where κ are proportional transaction costs and $a_t \in \{-1, 0, 1\}$ is the position at time t. It is important to note that the trading rules used are now a function of κ.

Beginning with a neutral position ($a_{t-m-1} = 0$), we can obtain the estimated parameters, estimated trading rules and out-of-sample trading rule returns $R_{t+1}^{d^*}$ at each date for various levels of proportional transaction costs κ. Our objective will be to find a test statistic which can be used to reject LIA. With this, we can then determine the level of transaction costs[21] κ for which LIA cannot be rejected at the 5% significance level.[22]

This exercise is conceptually similar to that of Cooper & Kaplanis (1994) who try to estimate the level of 'deadweight costs' that would explain the

[21] Note that as defined, the cost of switching from a long to a short position and vice versa is 2κ.

[22] It is important to note that Proposition 1.3 can be extended to the case of transaction costs if these are small enough. The same is not true for Proposition 1.1 if transaction costs are proportional.

home bias in international equity portfolios.[23] Notice that allowing for transaction costs changes our near-equilibrium notion in that we now seek pairs $\{\mathcal{U}^e, \kappa^e\}$ rather than just $\mathcal{U}^e$ for which LIA holds.

In Table 1.3 below we have tabulated the returns from a rule used by the Artificial Technical Analyst solving (1.18). The level of costs κ^e below which we can reject the hypothesis that $F^* \in \mathcal{F}^\kappa_\mathcal{U}$ for $\mathcal{U}$ including risk-neutral or mean-variance utility functions is represented by the line dividing Table 1.3. Notice that this κ incorporates the special case of zero transaction costs which was reported in Table 1.2.

Rules	Mean	St. Dev.	Prob. under Null	$\Pi_{t=1}^{6157}(1+R_t^d)$
Always long	0.0002334	0.008459	-	2.378
ATA				
κ=0	0.000801	0.008335	8.887e-05	110.7
κ=0.0001	0.0007304	0.008328	0.0005109	71.41
κ=0.0002	0.0006911	0.008321	0.001238	55.88
κ=0.0003	0.0006196	0.008318	0.00533	35.63
κ=0.0004	0.0005508	0.008311	0.01788	22.99
κ=0.0005	0.0004893	0.008305	0.0452	15.44
κ=0.0006	0.0004707	0.00829	0.058	13.67
κ=0.0007	0.0003741	0.00828	0.1755	7.101
κ=0.0008	0.0003099	0.008273	0.3061	4.456
κ=0.0009	0.0002763	0.008205	0.3876	3.453
κ=0.001	0.0002191	0.008215	0.5379	2.13

Table 1.3: The first column indicates which rule isunder consideration. The next two columns indicate the empirical mean and the standard deviation of the rules' returns. The fourth column shows the probability (under the assumption of normal distributions) that the mean returns from a specific rule were smaller or equal to the mean market returns. The final column shows the cumulative returns from each strategy during the whole time period.

The κ indicates that at the 5% level of significance, LIA will not be rejected for $\kappa \geq \kappa^e = 0.06\%$. The mean return of the optimal rule remains larger for $\kappa \leq 0.09\%$ (but not for the usual confidence margin).[24] These levels of costs make it tempting to argue that with today's cost conditions[25] LIA is rejected for a broad class of investors. However, costs were certainly larger at the beginning of the sample we have considered. How large the

[23] More recetly, this approach has been used by Luttmer (1999)

[24] Note that Proposition 3a can be extended to the case with transaction costs if these are "small enough". Table 3 indicates a mean-variance investor should probably prefer the ATA's rule.

[25] An investor with access to a discount broker, e.g. via email, can purchase 1000 shares of a company listed on the NYSE for a $9.95 fee. However, micro-structure frictions such as bid-ask spreads may be more important.

decrease in transaction costs has been and how it has affected different types of investors is a question which is beyond the scope of this paper and which we do not attempt to answer. It is probably fair to say that few agents would face transaction costs that were so low during most of the sample.

Bessembinder & Chan (1998) have independently performed a similar exercise on the same data to find the maximum level of costs for which a certain investment strategy beats Buy-and-Hold and arrive at a much higher estimate. This is due to the fact that they allow their strategy to take long positions *double the size* of that allowed for what they call the Buy-and-Hold strategy. Since the strategy is usually long and mean returns in the market are positive, clearly the strategy does better than Buy-and-Hold but this is merely a reflection of the larger positions their strategy takes. Their results are likely to be a lot weaker if their Buy-and-Hold strategy were defined in a more standard way. They would be even weaker if, as we do here, they reported results that are statistically significant.

1.5.3 Qualifications and further comments on the results

In this section, we have developed a test of our version of the weak efficient market hypothesis based on the profits of an Artificial Technical Analyst. We have used this test to characterise a broad class of investors (defined by their preferences and transaction costs) who would find it desirable to condition on past prices; as long as we believe such investors exist, we should reject our version of the efficient market hypothesis and hence E-efficiency for which our version is a necessary condition. Note that there are numerous asymmetric information models which generate equilibria for which LIA is rejected (e.g. Hussman (1992), Brown & Jennings (1989), Treynor & Ferguson (1985)) so some of the available models may describe the data generating process accurately (at least in this respect).

The methodology proposed is useful because it *relates efficiency to a set of objective functions rather than an equilibrium of some model.* It allows efficiency *comparisons* across *time* and *markets* by comparison of the generality of $\{\mathcal{U}^e,\kappa^e\}$ for which LIA is not empirically rejected. It thus provides a quantifiable measure of near efficiency. Equally importantly, it formalises a sense in which markets can be characterised as inefficient when empirical studies find trading rules to be profitable. It therefore formalises the popular notion of efficiency according to which agents solving simple investment problems such as (2.1) should not need to condition on past prices (or use technical trading rules).

The empirical exercise we conduct using the DJIA is intended to illustrate the implications of rule efficacy for market efficiency. A more accurate quantification of market efficiency using the Artificial Technical Analyst would require data on something closer to the market portfolio than the DJIA and a relaxation of the assumption that the riskless interest rate is

zero. We chose to focus on the DJIA data in order to make our results comparable to those of other researchers who have used it (Brock *et al.*, Gencay (1996), Sullivan et al. (1999), Bessembinder & Chan (1998) among others) and to diminish the susceptibility of our analysis to data-mining criticisms. However, the use of this data by us and them is unfortunate because the DJIA is not a traded asset and as such suffers from the complications raised by non-synchronous trading (Scholes & Williams 1977). Some authors attempt to correct for this (for example by requiring positions to be taken one day after a signal for them is generated) but there is no corrective procedure that does not raise other equally serious complications arising from the microstructure of markets. Further complications also arise from the fact that transaction costs for the DJIA during 1962-1986 are likely to have varied and may not even be proportional as we have assumed. Finally, the series in question is not adjusted for dividends which may bias our results against LIA.

Use of recent ultra high frequency data for options on market indexes would probably mitigate many of these problems and should be the subject of future investigations. Ultimately, the only indisputable empirical analysis of these issues would require the *Artificial* Technical Analyst to become *Real*: if actual trades based on this agents' decisions turned out to be profitable, this would constitute *prima facie* evidence against the belief that 'technical analysis is useless'.

1.6 Conclusions

Utility maximising best responses derived from investment decisions depend on utility functions, wealth, transaction costs and positions in the market. Technical trading rules are a class of behavioural rules which impose restrictions on the functional form these best response functions may take. However, these restrictions are only binding for strictly risk-averse investors and therefore utility maximising investors will typically 'be' Technical Analysts if they are risk-neutral (Proposition 1). Nevertheless, if for some reason risk-averse agents restrict themselves to the use of technical analysis (e.g. because more sophisticated investment rules are costly to derive, learn and implement) we can provide circumstances in which the choices of a Technical Analyst are also optimal for risk-averse agents in this constrained sense (Propositions 3, 4 and 5).

Artificial Technical Analysts use past data to choose rules from classes known to be used in practice. We show in Section 4.2 that the chosen rules will be more profitable than *ad hoc* rules used in previous studies (e.g. Brock *et al.*) and since they are chosen in a reasonable and robust way, they are also subject to less serious data-mining criticisms (Section 4.1). Taking these facts into account, we suggest that bootstrap based model specification

tests based on rule returns as pioneered by Brock et al. (1992) should be augmented with artificially intelligent agents in the spirit of Sargent (1993).

In Section 5 we attempt to formalise the idea that the efficacy of technical analysis and the efficiency of financial markets must be inversely linked. To do this, we begin with the observation that if the equilibrium of a model is such that agents would be better off using technical trading rules, then the equilibrium is not efficient in Latham's (1986) E-efficiency sense. Using daily data on the DJIA we characterise a class of agents (defined by preferences and transaction costs) who find the choices of an Artificial Technical Analyst valuable. Under appropriate assumptions, the size of this class (which includes mean-variance agents facing transaction costs lower than 0.06%) may be interpreted as a measure of the degree of efficiency of the NYSE (Section 5).

There are many natural extensions of this work so we restrict ourselves to some indicative suggestions. Firstly, the Artificial Technical Analyst could be made 'more intelligent' by making his learning more sophisticated and by widening the space of trading rules from which he chooses. For example, it is possible to extend the information set so that rules can condition on variables other than past prices (indicating that the Artificial Technical Analyst is capable even of fundamental analysis!). Secondly, it would be important to try and find statistical processes describing returns that are consistent with the Artificial Technical Analyst's profits. Finally, for applications where market efficiency is of central relevance, it would be valuable to develop a way of constructing dynamic asset pricing models which are consistent both with the data (e.g. Hansen & Richard (1987)) *and* with the degree of market inefficiency as quantified by the efficacy of Technical Analysis.

1.7 Appendixes

Proofs of propositions and Corollaries

Proof of Proposition 1.1. In text. ■

Proof of Proposition 1.2. The assumption may be rewritten as:

$$E\{E[U(W_t(1+a^*d(\mathbf{P}_t)R_{t+1}))|\mathbf{P}_t]\} > E\{E[U(W_t(1+a^*R_{t+1}))|\mathbf{P}_t]\},$$

which implies that there exists a set $\mathbb{Q}$ in $\mathbb{R}^k_+$ such that $\Pr(\mathbf{P}_t \in \mathbb{Q}) > 0$ and for all $\mathbf{P}_t \in \mathbb{Q}$:

$$E[U(W_t(1+a^*d(\mathbf{P}_t)R_{t+1}))|\mathbf{P}_t] > E[U(W_t(1+a^*R_{t+1}))|\mathbf{P}_t].$$

Now define:

$$a^{**} = \left\{ \begin{array}{c} a^*d(\mathbf{P}_t) \text{ if } \mathbf{P}_t \in \mathbb{Q} \\ a^* \text{ otherwise} \end{array} \right\}.$$

Clearly,

$$E[U(W_t(1+a^{**}R_{t+1}))|\mathbf{P}_t] \geqslant E[U(W_t(1+a^*R_{t+1}))|\mathbf{P}_t] \ \forall \ \mathbf{P}_t,$$

and the inequality is strict if $\mathbf{P}_t \in \mathbb{Q}$ which means a^* cannot be optimal for the conditional investment decision. ■

Proof of Proposition 1.3. Part (a)

Define $\overline{d}_i \equiv \frac{1}{l}d_i$, $i = 1, 2$

The assumption implies $E(d_1R_{t+1}) > E(lR_{t+1}) \geq 0$ so

$$E(\overline{d}_1R_{t+1}) > E(R_{t+1}) \geq 0$$

Notice that $\overline{d}_i = \left\{-\frac{s}{l}, 0, 1\right\}$ and by assumption $l \geq s$ so $\left(\overline{d}_i\right)^2 \leq 1$.

Hence $(\overline{d}_iR_{t+1})^2 \leq (R_{t+1})^2$

Using the fact that $V(x) = E\left(x^2\right) - E(x)^2$ it follows that:

$$V\left(\overline{d}_1R_{t+1}\right) < V(R_{t+1})$$

From which it also follows directly that

$$V(d_1R_{t+1}) < V(R_{t+1})$$

Part (b)

Define $\overline{d}_i \equiv \left(\frac{1}{l}\right)d_i$, $i = 1, 2$

Then the assumption $E(d_1R_{t+1}) > E(d_2R_{t+1}) \geq 0$ implies

$$E(\overline{d}_1R_{t+1}) > E(\overline{d}_2R_{t+1}) \geq 0$$

Notice that $\overline{d}_i = \{-1, 1\}$ so $\left(\overline{d}_i\right)^2 = 1$.

Hence $(\overline{d}_i R_{t+1})^2 = (R_{t+1})^2$, $i = 1, 2$

Using the fact that $V(x) = E\left(x^2\right) - E(x)^2$ it follows that:

$$V\left(\overline{d}_1 R_{t+1}\right) < V\left(\overline{d}_2 R_{t+1}\right)$$

Hence also:

$$V(d_1 R_{t+1}) < V(d_2 R_{t+1})$$

■

Proof of Corrolary 1.1. Ex hypothesi the returns of the trading rule are larger than the positive market returns:

$$E(R^d_{t+1}) > E(R_{t+1}) \geq 0$$

So by Proposition 3:

$$V(R^d_{t+1}) < V(R_{t+1})$$

These two inequalities imply also that for any a:

$$\begin{aligned} E(W_t(1 + aR^d_{t+1})) &> E(W_t(1 + aR_{t+1})) \\ Var(W_t(1 + aR^d_{t+1})) &< Var(W_t(1 + aR_{t+1})) \end{aligned}$$

which implies that for all mean-variance utility functions U :

$$EU(W_t(1 + a^* R^d_{t+1})) > EU(W_t(1 + a^* R_{t+1}))$$

■

Proof of Proposition 1.4. In normal environments second order stochastic domination is equivalent to mean-variance domination (Hanoch & Levy 1969). This together with Proposition 3a yield the desired conclusion.

■

Proof of Proposition 1.5. As is well known, the assumption on $M(\gamma)$ is a sufficient condition for:

$$EU(R^d_{t+1}) > EU(R_{t+1}) \; \forall \text{ concave } U$$

Notice now that when $E(R_{t+1}) \geq 0$ then the optimal proportion of invested wealth a^* is positive for all wealth levels and so $U(W_t(1 + a^* x))$ is concave in x, since:

$$\begin{aligned} \frac{\partial}{\partial x} U(W_t(1 + a^* x)) &= W_t a^* U' \geq 0 \\ \frac{\partial^2}{\partial x^2} U(W_t(1 + a^* x)) &= (W_t a^*)^2 U'' < 0 \end{aligned}$$

Therefore it must also be that

$$EU(W_t(1 + a^* R^d_{t+1})) > EU(W_t(1 + a^* R_{t+1}))$$

■

Auxiliary Results

Rule	N(buy)	N(sell)	Buy	Sell	Buy>0	Sell>0	buy-sell
Always long	6157	0	0.00023	-	-	-	-
BLL Average			0.00037	-0.00011			0.00048
ATA (N=250)	3313	2650	0.00095	-0.00067	0.5337	0.4675	0.00162
			(3.95033)	(-4.57949)			(7.34848)
ATA restricted	3463	2401	0.0005144	-0.00008416	0.5224	0.4794	0.0003221
ATA (N=225)	3372	2697	0.0009558	-0.0007419	0.538	0.4598	0.0008416
ATA (N=200)	3374	2703	0.000928	-0.0006966	0.537	0.4632	0.00081

Table 1.4: Compare to Table 1 in the text. We have estimated statistics for the Artificial Technical Analyst using varying sample sizes (N) and for a 'restricted' space of rules where the Analyst is constrained to choose only among the six rules of Brock *et al.* described in Table 1.

We conclude that a wide space of rules from which to choose is important for the performance of the Artificial Technical Analyst but that the precise values of N are not (as long as N is large enough for some stability to exist). These results are unsurprising but we provide them as empirical confirmation of what we suspected was the case.

Chapter 2

Risk Neutral Forecasting

SUMMARY

This chapter analyses optimal point forecasting in a simple investment context. The relation between the conditional distribution of returns and optimal point forecasts for a risk neutral investor is characterised in detail and it is shown that any mapping that has the same sign as the conditional mean of returns is a risk neutral investor's best predictor. It may therefore be difficult to model the conditional mean yet easy to model a 'risk neutral best predictor', for which some new methods are proposed. Using a simple simulation we show that the proposed approach may be better than standard competitors even in seemingly unfavourable situations. The forecasts thus obtained have maximal economic value (according to the usual definition) and are sufficient for the construction of mean-variance optimal binary investment rules.

2.1 Introduction

The desire to predict returns of financial series is to some extent responsible for the genesis of Economic Science: John Law, Richard Cantillon, Henry Thornton and David Ricardo developed their interest for economic systems through their activities as financial speculators.[1] For reasons that are obvious, interest in this topic has not waned.[2]

The objective of this paper is to study *Risk Neutral Forecasting* - the forecasting problem of a risk neutral investor in financial markets. Focus on this agent's decision problem is motivated by a number of related considerations which are worth examining in some detail.

[0] An overview of the results in this paper appear in 'An Introduction to Risk Neutral Forecasting', Abu-Mostafa, Y.S., B. LeBaron, A.W. Lo and A.S. Weigend, 1999, *Computational Finance*, MIT Press

[1] See, for example, Tvede (1997).

[2] See Campbell, Lo & MacKinlay (1997) for an overview of the state of the art.

The importance of the risk neutral investor's forecasting problem

The first reason for which the risk neutral investor's forecasting problem is of central importance in analysing optimal forecasts in investment contexts is that risk-neutral investors are probably the only expected utility maximising agents who can act on the basis of a *point* forecast. For example, a mean-variance agent will be interested in a joint forecast of both the mean and variance of a financial asset. More generally, risk averse agents will be interested in distribution forecasts for returns. We will return to this formally and in more detail in Section 2.2 but for now it should be appreciated that to the extent that risk neutrality offers a convenient starting point for understanding how optimal point forecasts for investors differ from 'generic' predictors.

Perhaps for related reasons, the loss function corresponding to the risk neutral decision problem we study has become the standard metric by which to quantify what is often referred to as the '*economic value*' of financial forecasts. For example, Henriksson & Merton (1981) develop a formal market timing test for measuring economic value and Leitch & Tanner (1991) evaluate the forecast performance of professional forecasters according to this metric. There have also been attempts to quantify the economic value of financial time-series models (Pesaran & Timmermann 1995, Pesaran & Timmerman 1998), structural models (Breen et al. 1989), technical trading rules (Sullivan et al. 1999) and agnostic 'money-machines' such as neural nets (LeBaron 1998*a*). Risk Neutral Forecasting can therefore also be seen as the task of formulating forecasts which are *optimal in terms of economic value*.

Furthermore, such forecasts can be used to construct binary *investment rules* (i.e. decision rules restricting an investor's choice set to two positions - usually of opposite sign) that are optimal for a wide class of agents. Investment rules with this structure, including technical trading rules (Brock et al. 1992) and market timing rules (Henriksson & Merton 1981) have received extensive attention in the literature, perhaps because they represent a boundedly rational mode of behaviour which not only results in a considerable simplification of investment decisions but is also empirically observed. These rules *simplify decisions* by allowing a broad class of agents to reduce their decision problem to a problem of Risk Neutral Forecasting. This occurs because of a remarkable property of binary investment rules ensuring that the rule which maximizes expected profits also minimizes the variance of profits (see Proposition 3, Chapter 1). Hence, the same rule is optimal for all mean-variance agents and they will take positions identical to those of a risk neutral investor. That binary investment rules can provide useful information for risk averse agents is a robust fact that has been shown to hold in general settings (Merton 1981).

More generally, good Risk Neutral Forecasting models can serve as nor-

mative models of how a restricted class of agents *should* estimate forecasts for returns. Indeed, there is some evidence suggesting that certain institutional investors and some individuals who invest a small proportion of their wealth behave in a fashion well approximated by risk neutrality. Furthermore, these techniques also serve as positive models of how agents *do* learn to forecast which in certain cases may be very useful.[3] However, since funds invested according to an approximately risk neutral objective function constitute a fairly small proportion of total invested funds, a positive model of this form is probably of limited interest.

Finally, an important, complicated and often neglected issue in producing good parametric model based forecasts for expected utility maximisers is accounting for parameter uncertainty (e.g. Barberis (1999)). This is particularly challenging in non-Bayesian frameworks which are the norm for modelling relatively high frequency financial series (for a review of available methods, see Bjornstad (1990)). While it may be unsurprising, it is certainly convenient that in some cases the investors' risk neutrality renders parameter uncertainty issues irrelevant.

In sum, the risk neutral forecasting problem should be viewed as an extremely convenient benchmark that is nevertheless interesting *per se*.

Organisation of chapter

In Section 2 we introduce the risk neutral investment decision and formalise the associated optimal predictors, which are the target of models for Risk Neutral Forecasting. In Section 3 we thoroughly characterise the relationship between optimal risk neutral forecasts and the conditional distribution of returns. In particular, we develop relations with the conditional mean of returns and with conditional quantiles of the sign of returns. Mappings the sign of which is the same as the sign of the conditional mean of returns are optimal forecasts for the risk neutral investor.

This is exploited in Section 4 where we discuss how Risk Neutral Forecasting models might be developed. Given that in applied financial forecasting misspecified models are a fact of life, we would like to adopt a modelling strategy that is robust with respect to misspecification. We therefore adopt a pseudo optimal modelling strategy (Skouras 2000) since pseudo optimal models have a number of desirable properties for expected utility maximisers using potentially misspecified models. In particular, they may be optimal even when the model is substantially misspecified and when this is not the case, they will be the best model in the class in an average sense. This paper is the first to apply a pseudo optimal modelling approach to a substantive problem.

[3] See Marimon (1996) and Sargent (1993) for a discussion of the scope of learning models in Economics.

It is also shown that a consistent estimator for a pseudo optimal Risk Neutral Forecasting model is provided by an estimator (discussed in detail in Chapter 4) for the sign of a conditional mean. This estimator is a sample analog of an unconditional version of the investors' objective function - hence is closely related to the idea of estimating models according to the 'relevant loss function' (Weiss 1996). While it has been argued that such estimators are 'good',[4] neither theoretical nor empirical justifications have not been forthcoming. In this paper, we provide reasons for which such estimators may be good in the context of the Risk Neutral Forecasting problem. Furthermore, simulation evidence is provided in a study that compares the relative performance of the proposed estimator to a maximum likelihood estimator.

Section 5 concludes; an Appendix follows, containing proofs of certain propositions.

2.2 Prediction and Investments

2.2.1 Best prediction of returns

Returns R_{t+1} to holding a financial asset from the current period t to $t+1$ may be described as:

$$\begin{aligned} R_{t+1} &= \mu_R(X_t) + U_{t+1} \\ E\left(U_{t+1}|x\right) &= 0, \end{aligned} \tag{2.1}$$

where X_t is a random variable with realisations x in $\mathcal{X} \subseteq \mathbb{R}^k$, $\mu_R : \mathcal{X} \to \mathbb{R}$ is a possibly non-linear function, U_{t+1} is a scalar disturbance term and $E\left(*|X_t\right)$ is the expectation function of a random variable conditional on realisations of X_t.

The objective of any forecaster of R_{t+1} observing (only) events in $\mathcal{X}$ can be formalised as the determination of the functional form of a 'best predictor' of R_{t+1} conditional on X_t, i.e. of a function which minimises the forecasters' expected loss when used in her decision problem.

Definition 2.1 *A **best predictor** is a mapping $p : \mathcal{X} \to \mathbb{R}$ that satisfies for all $x \in \mathcal{X}$:*

$$p\left(x\right) \in \arg\min_{\widehat{r}_{t+1} \in \mathbb{R}} \int L(R_{t+1}, \widehat{r}_{t+1}) dF_{R_{t+1}}|x \tag{2.2}$$

where $F_{R_{t+1}}|x$ is the cumulative distribution function of R_{t+1} conditional on $X_t = x$ and $L : \mathbb{R}^2 \to \mathbb{R}$ is a loss function which for each realisation of R_{t+1} assigns a loss at $t+1$ to a prediction $\widehat{r}_{t+1}$ (at time t) for R_{t+1}.

[4] To quote Granger (1993) '*if we believe that a particular criterion... should be used to evaluate forecasts then it should also be used at the estimation stage of the modelling process*'.

Obviously the behaviour of the set of solutions to (2.2) will depend strongly on the choice of loss function (for examples illustrating this dependence, see Christofferson & Diebold (1996)). When the context in which forecasts will be used is unknown, the convention is to allow certain 'standard' measures of location such as the conditional mean or median to be interpreted as forecasts. Standard forecasts may be 'good' predictors for some class of loss functions (if they satisfy some robustness property[5]), but will be *best* only for a very restricted class of loss functions.[6]

2.2.2 Prediction of returns in the context of a simple investment decision

Consider the simple *one period cash-single asset allocation decision* conditional on $X_t = x$:

$$\begin{aligned} & \max_{a \in [-s,l]} E\left\{U(W_{t+1})|x\right\} \qquad (2.3) \\ s.t.\ W_{t+1} &= aW_t(1+R_{t+1}) + (1-a)W_t \end{aligned}$$

where a (the fraction of wealth W_t invested in the asset) is constrained to be finite valued to capture (binding) borrowing and short-selling constraints. It is worth noting that while we abstract from these issues here, proportionate transaction costs and a riskless asset can easily be accommodated in the Risk Neutral Forecasting problem (see Chapter 4, Section 4.2).

A necessary condition for the *existence* of a best predictor corresponding to this investment decision problem is that the solution of (2.3) may be described as an action rule $a : \mathbb{R} \to [-s, l]$ mapping predictions $\widehat{r}_{t+1}$ for R_{t+1} to optimal investments, expressed as fractions of wealth. The reason for this is that $a(*)$ must be used to construct a loss function L - without which a best predictor cannot be defined. However, it is well known that unless restrictive assumptions are imposed on expected utility, (point) forecasts for returns do not provide sufficient information for utility maximising investment behaviour (the solution to (2.3) cannot be described by a mapping such as $a(*)$) so such a loss function does not exist and therefore (point) forecasting is not well defined. Indeed, this observation is at least in part responsible for the direction of recent research effort towards conditional distribution forecasting (Diebold, Hahn & Tay 1999, Diebold, Gunther & Tay 1998, Hansen 1994).

Appropriate assumptions on $\{U, F_{R_{t+1}}|x\}$ and on the investor's knowledge of $F_{R_{t+1}}|x$ can ensure that a forecast $\widehat{r}_{t+1}$ summarises the information

[5] But Geisel & Zellner (1968) and Zellner (1973)argue that such robustness properties are unlikely.

[6] These measures of location are best predictors for appropriately chosen loss functions. In particular, the aforementioned examples are optimal for squared error $L(r_{t+1}, \widehat{r}_{t+1}) = (r_{t+1} - \widehat{r}_{t+1})^2$ and absolute error loss functions $L(r_{t+1}, \widehat{r}_{t+1}) = |r_{t+1} - \widehat{r}_{t+1}|$ respectively.

necessary for utility maximising investors to solve (2.3). In particular, if investors know certain carefully chosen properties of $F_{R_{t+1}}|x$ but not its mean[7] point forecasts *are* sufficient for utility maximisation and the optimal investment *can* be described by a function $a(*)$ (which in some cases can even be determined analytically[8]). However, such assumptions are highly artificial not only because they impose very arbitrary restrictions on what the investor knows, but also because extreme assumptions on $F_{R_{t+1}}|x$ are required. On the other hand, if an investor is risk neutral such artificial assumptions are unnecessary because her decision will depend only on highly restricted features of the behaviour of $F_{R_{t+1}}|x$. By restricting the class of utility functions we are interested in to this narrow case, we are able to determine $a(*)$ without making any assumptions on $F_{R_{t+1}}|x$ or the investor's knowledge of it.[9] We will focus on this case and therefore maintain the following assumption throughout:

Assumption 2.1 *Investors are risk-neutral, i.e.* $U(W) = \beta_1 + \beta_2 W$.

All obvious time subscripts will henceforth be ignored.

2.2.3 Best prediction of returns by a risk neutral investor

Risk Neutral Best Predictors

A necessary and sufficient condition for $a(*)$ to be a function solving (2.3) when $U(W) = \beta_1 + \beta_2 W$ is that for all x :

$$a(x) = (l + s) \cdot I[E(R|x)] - s \tag{2.4}$$

where I is a sign indicator function taking the value 1 if the value in brackets is positive and 0 otherwise. We make the natural assumption that the risk neutral investor uses Bayes' rules with respect to her forecasts. If the risk neutral investor's forecast for R is $\widehat{r}$ she will take a position given by:

$$a(\widehat{r}) = (l + s) \cdot I[\widehat{r}] - s$$

[7] For example, under standard assumptions making utility a function of the conditional mean and variance, it is sufficient that the conditional variance is known. Analogously, West et al. (1993) derive a loss function for predictions of the conditional variance of returns by assuming mean-variance utility and that the conditional mean is known.

[8] See Campbell & Viceira (1999) for example.

[9] An alternative would be to depart from the objective of predicting returns and to focus instead on direct prediction of the optimal decision mapping using a non-parametric estimation technique (see e.g. Brandt (1999)). Brandt's (1999) approach which relies on estimation of solutions to Euler equations is not applicable for a risk neutral investor because the solution to the decision problem of interest does not satisfy a first order condition.

In this case, her loss is the difference between the utility (at time $t+1$) of a correct forecast and that obtained given that the forecast is $\widehat{r}$. This loss is a random function given by:

$$\begin{aligned} L(R,\widehat{r}) &= ((l+s)\cdot I[R]-s)R \\ &\quad -((l+s)\cdot I[\widehat{r}]-s)R \end{aligned} \tag{2.5}$$

As is evident from the definition of a best predictor (2.2), a loss function can be replaced with any increasing linear transformation without affecting the set of best predictors. We may multiply and subtract positive constants to obtain a simpler equivalent loss function:

$$L(R,\widehat{r}) = -R\cdot I[\widehat{r}] \tag{2.6}$$

Evidently, the borrowing and short-selling constraints of a risk neutral agent solving (2.3) do not influence her loss function in a way that affects optimal prediction (as long as they are finite). We may therefore use the more succinct loss function associated with an agent completely unable to borrow ($l=1$) or short-sell ($s=0$) to characterise best predictors of all risk neutral agents solving (2.3). Such predictors are called *risk neutral best predictors*.

Definition 2.2 *A* ***risk neutral best predictor***[10] *(RNBP) is a mapping* $p:\mathcal{X}\to\mathbb{R}$ *satisfying:*

$$p(x)\in\arg\min_{\widehat{r}\in\mathbb{R}^1} -\int R\cdot I[\widehat{r}]\,dF_R|x,\ x\in\mathcal{X} \tag{2.7}$$

2.3 Relation of risk neutral best predictors to the conditional distribution of returns

In this section we derive some properties of RNBPs that can be used to relate them to more familiar predictors.

2.3.1 RNBPs and the conditional mean of returns.

Simple manipulations of a risk neutral best predictor's definition formalise its relation to the conditional mean of returns. The following proposition shows that any predictor of returns with the same sign as the conditional mean will be best for a risk neutral investor. This is intuitive since this agent's action only depends on the sign of the conditional mean and hence such a prediction will always lead to optimal actions.

[10] For reasons that will become obvious, we will sometimes refer to this as an *ex post* risk neutral best predictor.

Proposition 2.1 *A mapping $p : \mathcal{X} \to \mathbb{R}$ is a risk neutral best predictor if and only if:*

(a) It satisfies:

$$p(x) \in \arg\min_{\widehat{r}\in\mathbb{R}} -\mu_R(x) \cdot I[\widehat{r}], \ x \in \mathcal{X} \tag{2.8}$$

Equivalently,

(b) It is a ***sign-preserving transformation*** *of the conditional mean, i.e. it satisfies:*

$$p(x) = \tau(\mu_R(x)), x \in \mathcal{X}$$

for some $\tau \in \mathbb{T} \equiv \{\tau : \tau(y) > 0 \Leftrightarrow y > 0\}$

Proof. (a) The definition of a risk neutral best predictor (2.7) is equivalent to (2.8).

(b) The set of solutions to (2.8) is the set of functions $\{\tau(\mu_R) : \tau \in \mathbb{T}\}$. Since, by definition, p is in this solution set, our conclusion follows immediately. ■

Proposition 1(a) states the trivial fact that the conditional mean $\mu_R(*)$ is a RNBP, which is not surprising since it is obvious from (2.4) that risk neutral investors can make optimal decisions on the basis of conditional means. A little less obvious is part (b) according to which any sign-preserving transformation of $\mu_R(*)$ is also a RNBP. The importance of this derives from the fact that the space of functions included in $\mathbb{T}$ is large,[11] as illustrated in Figure 2.1. Hence it may be that while $\mu_R(*)$ displays irregular features which make it difficult to estimate (for example, it may have discontinuities or may be changing over time), there is a $\tau \in \mathbb{T}$ such that $\tau(\mu_R(*))$ is regular (e.g. continuous and constant over time) and therefore easy to estimate accurately. In this case, the risk neutral investor can potentially *estimate a best predictor* ***even*** *when it is impossible in practice to estimate the conditional mean.*

[11] Manski (1988*b*, p.737) provides a characterisation of a small but important subset of $\mathbb{T}$.

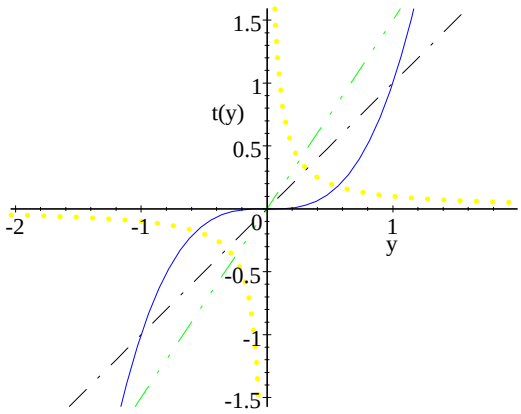

Figure 2.1 Some sign preserving transforms

2.3.2 RNBPs and the conditional distribution of the sign of returns.

Rather than thinking of $\widehat{r}$ as a prediction for R one may also think of (an indicator of) the sign of the prediction $I_{\widehat{r}} = I[\widehat{r}]$ as a prediction for (an indicator of) the sign of returns $I_R \equiv I[R]$. This interpretation can be useful as long as the relationship between the sign of the RNBP and the distribution of the sign of returns is understood.

In the first instance one might think that the sign of the RNBP is a forecast which maximises the probability of correctly forecasting the sign of returns (for example, the market timing test of Henriksson & Merton (1981) measures investors' forecasting ability according to this metric). However, this is not the case: A predictor that does not maximise the probability of correctly forecasting the sign of returns but which is good at predicting this sign when 'the stakes are high' will be preferred by a risk neutral investor. It is easy to show that the condition that the mean and median of $F_R|x$ have different signs is necessary and sufficient for a forecast maximising the probability of a correct sign to lead to wrong decisions.

Example 2.1 *Suppose R is a discrete random variable such that* $\Pr(R = 0.1) = 0.2$, $\Pr(R = -0.001) = 0.8$. *A positive constant $\widehat{r}_1$ is a risk neutral best predictor by Proposition 1, since $E(R) = 0.0192 > 0$. Consider now the prediction $\widehat{r}_2 < 0$. This has an 80% probability of correctly forecasting the sign of returns but it induces a risk neutral investor to be short and hence incur expected losses. On the other hand, a prediction $\widehat{r}_1 > 0$ has only a 20% chance of getting the sign right but results in a long position and expected profits.*

It is therefore clear that the risk neutral best predictor makes a compromise between correctly predicting the sign of returns and maximising the relative magnitude of returns when they are right compared to when they are wrong. The following proposition formalises this trade-off.

Proposition 2.2 *Define* $A : \mathcal{X} \to [0,1]$ *as:*

$$A(x) \equiv \frac{E(|R| \, | I_R = 1, x)}{E(|R| \, | I_R = 1, x) + E(|R| \, | I_R = 0, x)}, \; x \in \mathcal{X} \tag{2.9}$$

A mapping $p : \mathcal{X} \to \mathbb{R}$ *is a risk neutral best predictor if and only if:*

$$p(x) \in \arg\min_{\widehat{r} \in \mathbb{R}^1} -I[\widehat{r}] \cdot (A(x) - \Pr(I_R = 0|x)), \; x \in \mathcal{X} \tag{2.10}$$

Proof. See Appendix A. ■

Proposition 2.2 reveals the relationship between the best predictor and the conditional distribution of the sign of returns (the binary random variable I_R). One may think of A as a measure of the magnitude of returns when they are positive ($|R| \, | I_R = 1$) in relation to their magnitude when they are negative ($|R| \, | I_R = 0$). The proposition implies that if the distribution of R is sufficiently skewed towards the right, then the risk neutral investor should be long even if $\Pr(I_R = 0|x) > 0.5$ and A quantifies what is sufficient in relation to $\Pr(I_R = 0|x)$. Another way of expressing the relation between the c.d.f. of the sign of returns and RNBPs is provided by Proposition 3, showing that $I_{\widehat{r}}$ must be a particular (determined by $A(*)$) non-constant conditional quantile of I_R.

Proposition 2.3 *Let* $Q_\alpha(x)$ *be the* α*'th quantile of* $I_R|x$ *so that:*

$$Q_\alpha(x) \equiv \min_{\zeta \in [0,1]} \zeta : \Pr(I_R \leq \zeta | x) \geq \alpha, \; x \in \mathcal{X} \tag{2.11}$$

Let $Q_A(x)$ *be the* $A(x)$*'th quantile of* $I_R|x$ *so that:*

$$Q_A(x) \equiv \min_{\zeta \in [0,1]} \zeta : \Pr(I_R \leq \zeta | x) \geq A(x), \; x \in \mathcal{X} \tag{2.12}$$

A mapping $p : \mathcal{X} \to R$ *is a RNBP if and only if:*

$$Q_A(x) = I[p(x)], x \in \mathcal{X}$$

i.e. it is positive if and only if the $A(X)$*'th quantile of* $I_R|x$ *is* 1.

Proof. From the c.d.f. of I_R it is easy to verify that :

$$Q_A(x) = \left\{ \begin{array}{l} 1 \Leftrightarrow A(x) > \Pr(I_R = 0) \\ 0 \Leftrightarrow A(x) \leq \Pr(I_R = 0) \end{array} \right\}, \; x \in \mathcal{X}$$

It follows from Proposition 2 that:

$$I[p(x)] = \left\{ \begin{array}{l} 1 \Leftrightarrow A(x) > \Pr(I_R = 0) \\ 0 \Leftrightarrow A(x) \leq \Pr(I_R = 0) \end{array} \right\}, x \in \mathcal{X}$$

Combining these facts we obtain the desired result. ■

As simple as this proposition may be, it is somewhat surprising that the risk neutral best prediction problem is equivalent to determining a non-constant quantile of the sign of returns (depending on the realisation of X). Another way of thinking about this is that the sign of the RNBP minimises a lin-lin loss function with slopes determined by $A(x)$. This fact can considerably simplify the problem of best prediction when it is believed that the DGP displays certain properties. Here is an example of such a case.

Example 2.2 *Suppose it has been established that $F_{I_R}|x$ is such that for all x*

$$E(|R| \, | I_R = 1, x) = E(|R| \, | I_R = 0, x).$$

Then $A(x) = \frac{1}{2}$ for all x and hence $Q_A(x)$ is the median of $F_{I_R}|x$. Knowledge of the median of the conditional distribution of the sign of returns is sufficient for a risk neutral investor to make her optimal decisions.[12]

It may be instructive to think intuitively about the degree to which the behaviour of $F_R|x$ must be known in order to obtain the risk neutral best predictor. Assuming that $A(*)$ is known, a lower bound is provided by the conditional quantiles $Q_a(*)$ of the sign of returns over the range of $A(*)$: without this, the optimal decision will be unknown, at least for some x. An obvious upper bound is the conditional mean, but a tighter upper bound is given by I_R the conditional distribution of the sign of returns.

2.4 Modelling Risk Neutral Forecasts

Having characterised risk neutral best predictors and their relation to the conditional distribution of returns $F_R|x$, we now turn to the problem of modelling them when $F_R|x$ is unknown, but an empirical distribution function F_N consisting of draws $\{r_n, x_n\}_{n=1}^N$ from F is observed. The objective is now to define a 'good' prediction rule $q : \mathbb{R}^N \times \mathcal{X}^N \times \mathcal{X} \rightarrow \mathbb{R}$ mapping the observed sample and conditioning variable into a forecast $\widehat{r}$, i.e. $\widehat{r} = q(F_N, x)$.

To proceed, some further structure is usually imposed on the form of potential prediction rules. Here we will focus on the situation where a potentially mis-specified parametric model $\mathcal{G}$ for $F_Y|x$ is postulated, taking the

[12] More generally, when A is equal to a constant a, Risk Neutral Best Predictors can be found by determining the a'th quantile of the binary response I_R. Quantile regressions of binary responses have been studied by Manski & Thompson (1989).

form:

$$\mathcal{G} = \left\{ G_R|\left[x, c\right], c \in B \subseteq \mathbb{R}^k \right\}$$

It is 'natural' to use F_N to choose a parameter $\widehat{c}_N$ (and hence a model) and assume that a RNBP with respect to $G_R|\left[x, \widehat{c}_N\right]$ will be a good approximation to the RNBP with respect to $F_R|x$ - particularly if parameter uncertainty and the appropriate loss function are accounted for in estimation of c (see e.g. Zellner (1978)). In this case forecasts are given by decision rules of the form $q : B \times \mathcal{X} \to \mathbb{R}$ with $\widehat{r} = q\left(\widehat{c}_N, x\right)$. Of course, there are other ways of constructing prediction rules (see e.g. Geisser (1993)) but we will restrict our attention to this type of rules in what follows.

In the rest of this section we will propose a novel way of obtaining $\widehat{c}_N$ and discuss the reasons for which it may be preferable to standard estimators of the pseudo-maximum likelihood, least squares and Bayesian variety.

2.4.1 Pseudo-optimal risk neutral forecasting models

If the model is mis-specified - as is likely to be the case in most applications - we have no guarantee that standard estimators lead to good forecasting rules (let alone best predictors) even asymptotically. In particular, standard estimators generally converge to *pseudo-true* parameters (minimising the Kullback-Leibler discrepancy from F) and it is not clear that the investor will do well using a predictor based on a pseudo-true model. In particular, there may be a model in $\mathcal{G}$ that leads to predictions of greater utility to the investor.

Skouras (2000) examines this issue in the context of general expected utility maximisation problems that nest the Risk Neutral Forecasting problem considered here. It is assumed that agents use Bayes' rules with respect to models (as we do here) and models are compared in terms of the payoffs of the Bayes' rules they generate. A *pseudo-optimal model* in $\mathcal{G}$ is defined as the model in $\mathcal{G}$ that generates the largest (unconditional) expected utility (evaluated with respect to F) for its users. In the context of the risk neutral investment decision of interest, the *pseudo-optimal risk neutral forecasting model in* $\mathcal{G}$ may be defined by a *pseudo-optimal risk neutral forecasting parameter* b such that:

$$b \in \arg\min_{c \in B} - \int R \cdot I\left[m(X, c)\right] dF, \tag{2.13}$$

where F is the joint c.d.f. of (R, X) and

$$m\left(x, c\right) = \int R \cdot dG_R|\left[x, c\right]$$

is the conditional mean of the model with parameter c.

Pseudo-optimal models are preferable to pseudo-true models for two reasons. The first arises from the observation that pseudo-optimal models may be as good as the true conditional distribution $F_R|x$ for the purposes of a particular decision-maker even when $\mathcal{G}$ is misspecified. In the context of the risk neutral investment decision, this will be the case when $\mathcal{G}$ is misspecified yet there is a $b \in B$ such that $m(*, b)$ is a sign-preserving transform of the conditional mean $\mu_R(*)$. In order to characterise all fully optimal models, we show that a necessary and sufficient condition for $m(*, b)$ to be a RNBP is that there is an optimal model in $\mathcal{G}$ and that b satisfies (2.13).

Proposition 2.4 *Let $m : \mathcal{X} \times B \to \mathbb{R}$ be the conditional mean of a model $G_R|[*, *]$ for $F_R|x$ such that $m(x, b)$, $b \in B$ is a RNBP. Then $G_R|[*, b]$ is an optimal risk neutral forecasting model almost everywhere on $\mathcal{X}$ if and only if $G_R|[*, b]$ is pseudo-optimal (i.e. b satisfies (2.13)).*

Proof. See Appendix. ■

The second advantage of pseudo-optimal models is that even when it is not the case that pseudo-optimal models are in fact optimal, Bayes' rules with respect to the pseudo-optimal model will be better in indefinite repetitions of the utility maximisation problem than any other Bayes' rules with respect to models in $\mathcal{G}$. This is because these Bayes rules optimise an unconditional version of the expected utility maximisation problem of interest. Equivalently, while there may exist a better model with respect to which to form Bayes' rules for some $x \in \mathcal{X}$, it will nevertheless be optimal *on average* (where each realisation of X is weighted by its density).

In our context, a risk neutral best predictor is the risk neutral investors' Bayes' rule with respect to the truth $F_R|x$. It solves an expression equivalent to:

$$\max_{a \in [-s, l]} E\{W_t \cdot (1 + a \cdot R_{t+1}) | x\}, x \in \mathcal{X}$$

On the other hand, the risk neutral best predictor with respect to the pseudo-optimal risk neutral forecasting model in $\mathcal{G}$ is a function $m(*, b)$ such that b solves:

$$\max_{c \in B} E\{W_t \cdot [1 + ((l + s) \cdot I[m(X, c)] - s) \cdot R_{t+1}]\}$$

By contrast, Bayes' rules with respect to pseudo-true models do not satisfy any property that makes them suitable for use by an expected utility maximiser and hence there is *no ground on which to expect the standard modelling approach to be 'good' for a risk neutral investor.*[13] We will therefore apply the Pseudo-optimal modelling approach of Skouras (2000). It is

[13] Pseudo-true models often minimise the least-squares distance of a model from the conditional mean. The properties of such approximations are often misundertood as discussed by White (1980).

worth noting that within this approach $\mathcal{G}$ need only specify models for the conditional mean $\mu_R|x$ of R; if this is the case, $\mathcal{G}$ should be interpreted as a semi-parametric model for $F_R|x$.

2.4.2 Estimating Risk Neutral Forecasting models

We now turn to the issue of estimating a pseudo-optimal risk neutral forecasting model. In Chapter 4 we propose general conditions under which various estimators asymptotically converge to b given by (2.13). The most general conditions are for the estimator b_N given by:

$$b_N \in \arg\min_{c\in B} - \int R \cdot I\left[m(X,c)\right] dF_N, \tag{2.14}$$

since it does not require correct specification of $m(*,*)$ for the sign of the conditional mean (Assumption 4.1). Assuming these conditions hold, use of forecasts given by $m(*,b_N)$ will be asymptotically pseudo-optimal. The reason for this difference is the fact that b_N is a sample analogue of (2.13); we will refer to the above estimator as the SAME estimator of a RNBP as it is constructed from the Sample Analogue Moment Extremum condition satisfied by the RNBP. In Chapter 4, we study the statistical and computational properties of this estimator. In this Chapter we consider its properties as a way of estimating a risk neutral forecasting model.[14]

In order to compare estimated forecasting models in finite samples, we will require that they perform well in terms of *risk* which is a measure of their performance when they use a partuclar forecasting rule $q\left(\widehat{c}_N, x\right)$ and estimator $\widehat{c}_N$. For the risk neutral investor, the *conditional risk* with respect to x when F is the true distribution is given by:

$$\mathcal{R}\left[\left(F,\widehat{c}_N\right)|x\right] = -\int R \cdot I\left[m(x,\widehat{c}_N)\right] dF_{R,\widehat{c}_N}|x \tag{2.15}$$

where $F_{R,\widehat{c}_N}|x$ is the joint c.d.f. of R and $\widehat{c}_N$ conditional on x (derivable from F, the joint c.d.f. of R and X). An unconditional measure of risk is given by integrating over $\mathcal{X}$ in (2.15) to obtain:

$$\mathcal{R}\left[\left(F,\widehat{c}_N\right)\right] = -\int R \cdot I\left[m(X,\widehat{c}_N)\right] dF_{R,X,\widehat{c}_N}. \tag{2.16}$$

While we know that the SAME estimator will asymptotically minimise the unconditional measure of risk under general conditions on F, it is unreasonable to expect any estimator to minimise either measure of risk for all possible F in finite samples. Rather what we are after are estimators the performance of which is fairly *robust* with respect to F (Berger 1985).

[14] This estimator is an application of the idea of estimating a model using the relevant loss function (Weiss 1996, Granger 1993, Manski 1991).

Verification of such a property is a difficult task and there are few general results, particularly when estimators are based on misspecified models.

In the following Section we conduct a simulation experiment of limited but important scope. The goal is to compare the performance of the SAME estimator and a standard maximum likelihood estimator in a circumstance that would seem to favour the latter.

2.4.3 Simulated Risk of estimators for Risk Neutral Forecasting

Consider the following DGP:

$$\begin{aligned} R_{t+1} &= b_0 + b_1 R_t + U_{t+1} \\ U &\sim N\left(0, \sigma^2\right) \text{ iid} \\ R_0 &\sim N\left(\frac{b_0}{1-b_1}, \sigma^2 \left(1-b_1\right)^{-1}\right) \\ b_0 &= 0.00015;\ b_1 = 0.0330;\ \sigma = 0.0108 \end{aligned}$$

The parameters of this DGP were determined using OLS to estimate an AR(1) model on a series of returns drawn from the empirical distribution of IBM daily closing prices[15] from 1st January 1990 through to 6th November 1997 (2049 observations).

Suppose the parametric form of the DGP is known, but that the values of the parameters are not. For reasons discussed in more detail in Chapter 4, our estimator can only provide a consistent estimate of the scale of the parameters of a linear model. Hence, the model to be estimated takes the form:[16]

$$m(X, c) = c + X$$

An estimate for $\frac{b_0}{b_1}$ is b_N satisfying:[17]

$$b_N \in \arg\min_c - \int R_{t+1} \cdot I\left[c + R_t\right] dF_N \tag{2.17}$$

We took $T = 10^4$ draws from the c.d.f. F_{b_N} of b_N to obtain the simulated distribution $\widehat{F}_{b_N}$ and repeated for various sample sizes. In particular, we set: $N = \left\{100 \cdot 2^{l-1}\right\}_{l=1}^{7}$. The size of T was chosen so that the standard

[15] Obtained from DATASTREAM on the last day in the dataset.

[16] Strictly speaking, we should also estimate the model $s\left(R, c\right) = c - R$ unless we know the sign of b_1, but (for simplicity) we will assume here (and whenever we estimate linear models) that this is indeed known. Often either theory or related empirical evidence suggest this sign. Relaxing this assumption would involve significant computational effort and is unlikely to be important for our conclusions.

[17] To fix b_N at a particular point, we took (both here and later) the value closest to zero satisfying (4.9).

deviation $\widehat{\sigma}_{b_N}$ of $\widehat{F}_{b_N}$ did not change 'much' between $\frac{T}{2}$ and T for any N. The distribution $\widehat{F}_{b_N}$ was then used to study the risk associated with this estimator.

In this setting, maximum likelihood suggests itself as a strong competitor to the parametrisation obtained using the SAME estimator, since statistical theory suggests it will have a number of desirable properties (asymptotic normality, efficiency, quadratic convergence rates etc.). Maximum likelihood estimates can be obtained as:

$$\left(b_0^{ml}, b_1^{ml}\right) \equiv \arg \min_{(c_0,c_1)} \int \left(R_{t+1} - c_0 - c_1 R_t\right)^2 dF_N \tag{2.18}$$

so it is readily seen that they are also least squares estimates. We might expect the attractive tractable properties of ML estimation as well as the additional assumptions we conveniently impose to provide a 'better' forecasting model. However, the relevant comparison is in terms of the Risk associated with using each estimator.

Conditional Risk comparison of estimators

The risk conditional on r_t for the forecasts based on estimates $\widehat{c}_0, \widehat{c}_1$ is a function of the form $\mathcal{R}\left[(b_0, b_1, \sigma, \widehat{c}_0, \widehat{c}_1)\,|r_t\right]$. Given that (b_0, b_1, σ) are fixed in this simulation, we may suppress them in our notation for the risk function so that it takes the form:

$$\begin{aligned} \mathcal{R}\left(\widehat{c}_0, \widehat{c}_1 | r_t\right) &= -\int R_{t+1} \cdot I\left[\widehat{c}_0 + \widehat{c}_1 r_t\right] dF_{R_{t+1},\widehat{c}_0,\widehat{c}_1} | r_t & (2.19) \\ &= -\int E\left(R_{t+1} | r_t\right) \cdot I\left[\widehat{c}_0 + \widehat{c}_1 r_t\right] dF_{\widehat{c}_0,\widehat{c}_1} | r_t & (2.20) \end{aligned}$$

We will make the assumption that for the estimators we consider $\widehat{c} \equiv \frac{\widehat{c}_0}{\widehat{c}_1}$ and R_t are independent[18] and furthermore that $\widehat{c}_1 > 0$, so that:

$$\mathcal{R}\left(\widehat{c} | r_t\right) = -\left(b_0 + b_1 r_t\right) \cdot \left(1 - F_{\widehat{c}}\left(-r_t\right)\right) \tag{2.21}$$

Let $\left(b_0^{ml}, b_1^{ml}\right)$ and be parameter estimates obtained by maximum likelihood. The associated forecasts are given by:

$$\widehat{r}^{ml} = b_0^{ml} + b_1^{ml} r_t. \tag{2.22}$$

It may be worth noting that the maximum likelihood forecasts are also optimal point Bayes forecasts for this decision problem when priors are diffuse. Furthermore, it may be verified that taking parameter uncertainty into

[18] This assumption is reasonable when the series does not exhibit dependence, or when the sample used for estimation is sufficiently distant to t with respect to the dependence in the series.

account will not affect the form of these forecasts, because risk neutrality implies optimal estimates are a function only of $E\left(R_{t+1}|r_t\right)$ with uncertainty about this being irrelevant.

The risk conditional on r_t for the maximum likelihood forecasts is:

$$\mathcal{R}\left(b^{ml}|r_t\right) = -\left(b_0 + b_1 r_t\right) \cdot \left(1 - F_{b^{ml}}\left(-r_t\right)\right) \tag{2.23}$$

where $F_{b^{ml}}$ is the c.d.f. of $b^{ml} \equiv \frac{b_0}{b_1}$.

Analogously, the SAME forecasts are given by:

$$\widehat{r}_N = b_N + r_t. \tag{2.24}$$

with risk:

$$\mathcal{R}\left(b_N|r_t\right) = -\left(b_0 + b_1 r_t\right) \cdot \left(1 - F_{b_N}\left(-r_t\right)\right) \tag{2.25}$$

where F_{b_N} is the c.d.f. of b_N.

We can get an idea of how these risk measures compare by using the simulated distributions $\widehat{F}_N$ and $\widehat{F}_{b^{ml}}$ obtained in our simulation to plot the conditional risk functions of each forecasting model for each N (Figure 2.2).

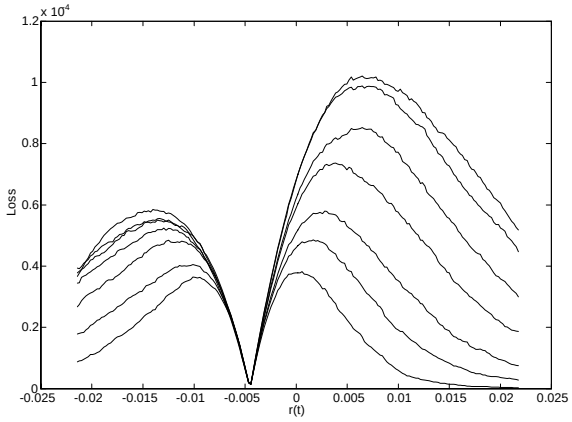

Figure 2.1(a). This is a plot of (normalised) conditional risk for each sample size N of the SAME estimator forecasts. As the sample size increases and estimates become more accurate, risk decreases and the risk function becomes smoother.

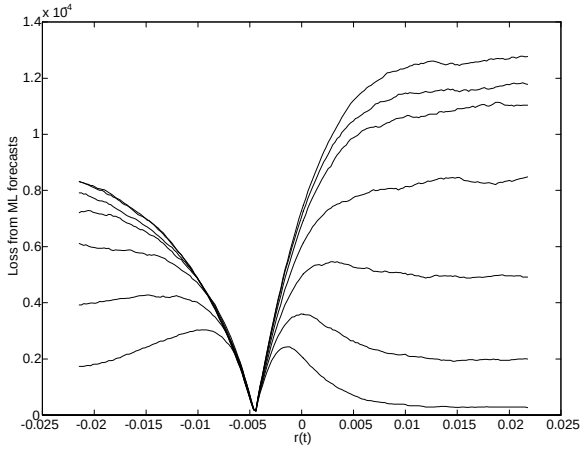

Figure 2.1(b). This is a plot of (normalised) risk for each sample size N of the ML estimator forecasts.

It is clear from (2.21) that the performance of an estimator depends on how closely the distributions $F_{b^{ml}}$ and F_{b_N} are concentrated around the ratio $\frac{b_0}{b_1}$. It turns out that the ML estimator provides very bad estimates for this ratio in medium-sized samples because when b_1^{ml} takes on small values, the maximum likelihood estimate $\frac{b_0^{ml}}{b_1^{ml}}$ explodes[19] and as a consequence, the SAME forecasts are better *for all* r_t in realistic sample sizes. This problem is probably due to the fact that in financial series, b_0 and b_1 are likely to be small relative to σ suggesting the sign of $\frac{b_0}{b_1}$ will often be wrong. The close connection between the maximum likelihood estimates and Bayes' estimates suggests that they are admissible, meaning that for some (b_0, b_1, σ) they will not be dominated by the SAME estimator. However, this might be for an irrelevant region in the parameter space.

For sufficiently large samples, the fact that the b^{ml} converges at a faster rate than b_N (see Chapter 4) leads the ML forecasts to become better for most r_t (particularly those close to the mean of the marginal distribution F_{R_t} of R_t).

Unconditional Risk comparison of estimators

When neither estimator performs better for all r_t, a ranking which is *not* conditional on r_t can be provided by the measure in (2.16). For this simulation, it may be expressed as:

$$\mathcal{R}\left(b^{ml}\right) = -\int \left(b_0 + b_1 R\right) \cdot \left(1 - F_{b^{ml}}\left(-R\right)\right) dF_R \tag{2.26}$$

[19] This is related to problems addressed in the statistical literature on 'inverse regression' - see for example Kruthckoff (1967).

and

$$\mathcal{R}(b_N) = -\int (b_0 + b_1 R) \cdot (1 - F_{b_N}(-R))\, dF_R \tag{2.27}$$

Analytical expressions for these Risk functions conditional on particular values of b_N or b^{ml} can be evaluated using formulae provided in Appendix B. Using the simulated distributions $\widehat{F}_{b^{ml}}$ and $\widehat{F}_{b_N}$ as estimates for $F_{b^{ml}}$ and F_{b_N} in the expressions above, we obtain simulated estimates of the unconditional Risk[20] for each estimator and each N. These appear in Table 2.1 and Figure 2.2 below.

N	100	200	400	800	1600	3200	6400	∞
$-10^4 \cdot \mathcal{R}(b^{ml})$	0.99471	1.1420	1.3233	1.5401	1.8026	2.0400	2.1994	2.3428
$-10^4 \cdot \mathcal{R}(b_N)$	1.7105	1.7328	1.8048	1.8772	1.9784	2.0592	2.1408	2.3428

Table 2.1 Each column corresponds to a sample size N for which we compute the unconditional Risk of each estimator. The column corresponding to ∞ gives the profits obtained when the true model is used.

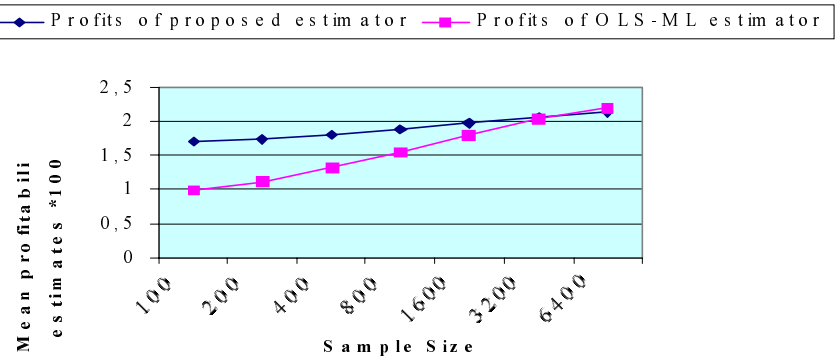

Figure 2.2. Plot of estimators' unconditional Risk as a function of sample size.

We conclude that the SAME estimator performs better up to very large sample sizes according to this metric. This suggests that for series that are believed to be sufficiently stable over time to make it reasonable to use large samples in estimation, it may be a good idea to switch to the maximum likelihood estimator in large samples. However, this is conditional on the model being correctly specified, since otherwise we know that the SAME estimator is likely to be do better in large samples because it converges to the pseudo-optimal parameter.

[20] In Chapter 3 we also report mean squared errors for the estimators in this simulation.

More generally, caution is warranted in drawing generalisations from results obtained within this narrow environment. It is worth noting that the performance of standard estimators in terms of *un*conditional risk can be improved by using modifications that take into account parameter uncertainty and the nature of the relevant risk function (see e.g. Zellner (1978) for one approach to this problem) even though the same is not true for the conditional problem.

2.5 Conclusion

We have characterised the forecasting problem of a risk neutral investor and proposed an estimator for models of her optimal forecasts. Restricting attention to this type of investor limits the scope of results obtained but allows us to develop: (1) A convenient starting point for formulating a *decision theoretic foundation* for financial forecasting (but see also Granger & Pesaran (forthcoming) and Granger & Pesaran (2000)), (2) An attempt to provide a rationalisation of 'economic value' as a performance metric and estimation techniques that can be used to improve forecast performance according to this metric, (3) A set of results that may be used for a formal treatment of 'investment rule' estimation, applicable for example to the estimation of technical trading rules and market timing rules.

The 'risk neutral best predictor' is a formalisation of the notion of an 'optimal point forecast for a risk neutral investor'. Since any function which has the same sign as the conditional mean is a risk neutral best predictor, it follows that Risk Neutral Forecasting is 'easier' than conditional mean forecasting in the sense that less stringent assumptions on model specification accuracy are required for optimal predictors to be tenable. This occurs because the conditional mean is only one of an infinity of risk neutral best predictors any of which solves the forecasting problem of the risk neutral investor. Thus the estimator proposed for learning risk neutral best predictors will lead to models that are optimal for this investor under weak assumptions on model specification.

More generally, the estimation method proposed will lead to a *pseudo optimal model*. That is, it will lead to the model in the chosen class that maximises the unconditional expected utility of a risk neutral investor when used as the basis on which beliefs and forecasts are formed. Furthermore, our simulation has shown that even in conditions that seemingly favour standard estimators this estimator may be better for the risk neutral investor. It is not entirely clear what drives this interesting result.

A number of trivial but interesting extensions have been discussed such as including transaction costs in the decision problem of the risk neutral agent; others are feasible no doubt. The exploration of complementaries with other lines of research in financial forecasting may also be an avenue

for fruitful work. For example, it would be interesting to use our results to modify Lo & MacKinlay's (1997) methods for constructing maximally predictable portfolios to obtain portfolios the predictability of which would be of *maximal economic significance*.

It would be useful to consider how other stages of the modelling process (in addition to estimation) should be modified to accommodate the specific problem of Risk Neutral Forecasting. For example, deriving the distribution of *out of sample* performance of estimated Risk Neutral Forecasting models could be used for model selection and evaluation (see West (1994)). Hopefully our characterisation of optimal forecasts in this context will be useful in doing so. The need for such results is accentuated by the fact that the properties of the proposed estimator are understood only imperfectly (see Chapter 4).

All these directions are important but the main priority for future research is *empirical*. The results developed permit estimation of a Risk Neutral Forecasting model that can combine the structure of econometric models for returns with the profitability of the most successful investment rules. A hybrid model of this form should provide new insights into the structure of financial series in general and may be particularly useful in detecting regularities which exert a strong influence on investment behaviour.

2.6 Appendixes

Appendix A: Proofs of lengthy propositions

Proof of Proposition 2. Since for $x \in \mathcal{X}$:

$$\mu_R(x) = \Pr(I_R = 0|x)E(R|I_R = 0, x) + \Pr(I_R = 1|x)E(R|I_R = 1, x)$$

It follows that for $x \in \mathcal{X}$:

$$\begin{aligned}\mu_R(x) &= -\Pr(I_R = 0|x)E(|R||I_R = 0, x) + \Pr(I_R = 1|x)E(|R||I_R = 1, x) \\ &= E(|R||I_R = 1, x) - \Pr(I_R = 0|x)\left[E(|R||I_R = 0, x) + E(|R||I_R = 1, x)\right]\end{aligned} \tag{2.28}$$

Define $h : \mathcal{X} \to \mathbb{R}_+$ such that:

$$h(x) \equiv \frac{1}{E(|R||I_R = 0, x) + E(|R||I_R = 1, x)} \tag{2.29}$$

If $p(x)$ is a risk neutral best predictor, by Proposition (1a) it satisfies:

$$p(x) \in \arg\min_{\widehat{r}\in\mathbb{R}^1} -I[\widehat{r}] \cdot \mu_R(x), \ x \in \mathcal{X}$$

Since for all $x \in \mathcal{X}$, $h(x) > 0$ this is equivalent to:

$$p(x) \in \arg\min_{\widehat{r}\in\mathbb{R}^1} -I[\widehat{r}] \cdot \mu_R(x) \cdot h(x), \ x \in \mathcal{X}$$

By substituting (2.28) and (2.29) into this expression, we obtain (2.10). ■

Proof of Proposition 4.

Suppose $m(*, b)$ is a RNBP almost everywhere on $\mathcal{X}$. By definition:

$$m(x, b) \in \arg\min_{\widehat{r}\in\mathbb{R}^1} -\int R \cdot I[\widehat{r}]\, dF|x, \text{ a.e. } \mathcal{X}$$

so it must be that for all $c \in B$:

$$-\int R \cdot I[m(X, b)]\, dF|x \leq -\int R \cdot I[m(X, c)]\, dF|x, \text{ a.e. } \mathcal{X}$$

Integrating over all inequalities in $\mathcal{X}$ with the probability measure of x :

$$-\int R \cdot I[m(X, b)]\, dF \leq -\int R \cdot I[m(X, c)]\, dF, \ \forall c \in B,$$

So (2.13) is indeed a necessary condition.

$\Leftarrow$

Notice that:

$$\begin{aligned} -\int R \cdot I\left[m\left(X,c\right)\right] dF &= -\int \mu_R\left(X\right) \cdot I(X,c))dF \\ &\geq -\int \mu_R\left(X\right) \cdot I(X,c) \cdot I_{\mu_R}(X)dF \\ &\geq -\int \mu_R\left(X\right) \cdot I_{\mu_R}(X)dF. \end{aligned}$$

for all $c \in B$.

Since the model is correctly specified, there exists a $b\prime$ that satisfies $m\left(X,b'\right) = \mu_R(X)$ so

$$-\int R \cdot I\left[m\left(X,c\right)\right] dF \geq -\int \mu_R\left(X\right) \cdot I(m\left(X,b'\right))dF$$

If b minimises (2.13) then it must be that $I\left[m\left(X,b\right)\right] = I(m\left(X,b'\right))$ almost surely. Then by Proposition 1 $m\left(X,b\right)$ is a RNBP almost surely. ■

Appendix B: The distribution of profits when forecasts and returns have a joint normal distribution

Here we derive general expressions for the distribution of a risk neutral investor's profits (defined as -1 times the loss function in (2.6)) under the condition that predictions and returns are jointly normal.[21] These results are used in Section 4.3 to derive losses of SAME and ML estimators.

Let:

$$\begin{bmatrix} Y_1' \\ Y_2' \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1 & \sigma_{12} \\ \sigma_{12} & \sigma_2 \end{bmatrix}\right)$$

Suppose Y_2' is a forecast for Y_1' and let X be the profits obtained by a risk neutral investor from the use of this forecast. Then,

$$X = Y_1' \cdot \mathbf{1}\left(Y_2' > 0\right)$$

Let

$$Y_1 \equiv \frac{Y_1' - \mu_1}{\sigma_1}, Y_2 \equiv \frac{Y_2' - \mu_2}{\sigma_2}, \rho = \frac{\sigma_{12}}{\left(\sigma_1\sigma_2\right)^{\frac{1}{2}}}$$

Then

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right)$$

[21] Acar (1998) has derived the expression for the mean of a closely related distribution.

So

$$X|Y_{1,}Y_2 = \left\{ \begin{array}{c} \mu_1 + \sigma_1 Y_1 \text{ if } Y_2 > -\frac{\mu_2}{\sigma_2} \\ 0 \text{ otherwise} \end{array} \right\},$$

which implies:

$$\begin{aligned} E(X) &= \mu_1 \Pr\left(Y_2 > -\frac{\mu_2}{\sigma_2}\right) + \sigma_1 E\left(Y_1 \cdot \mathbf{1}\left(Y_2 > -\frac{\mu_2}{\sigma_2}\right)\right) \\ &= \mu_1\left(1 - \Phi\left(-\frac{\mu_2}{\sigma_2}\right)\right) + \sigma_1 \int_{-\infty}^{\infty}\int_{-\frac{\mu_2}{\sigma_2}}^{\infty} Y_1 \cdot f_{Y_{1,}Y_2} dy_2 dy_1 \end{aligned}$$

where Φ is the c.d.f. of the standard normal and $f_{Y_{1,}Y_2}$ is the p.d.f. of (Y_1, Y_2).

Johnson & Kotz (1972, p. 113) report results which imply:

$$\int_{-\infty}^{\infty}\int_{-\frac{\mu_2}{\sigma_2}}^{\infty} Y_1 \cdot f_{Y_{1,}Y_2} dy_2 dy_1 = \frac{\rho}{\sqrt{2\pi}} \exp\left(-0.5\left(\frac{\mu_2}{\sigma_2}\right)^2\right)$$

Hence:

$$E(X) = \mu_1\left(1 - \Phi\left(-\frac{\mu_2}{\sigma_2}\right)\right) + \sigma_1 \frac{\rho}{\sqrt{2\pi}} \exp\left(-\left(\frac{\mu_2}{\sigma_2}\right)^2\right) \tag{2.30}$$

The Variance of this strategy can also be calculated, by using the fact that $Var(X) = E\left(X^2\right) - E(X)^2$, (2.30) and an expression for $E\left(X^2\right)$ provided by Johnson & Kotz (1972, p. 113).

Example: Use of an AR(1) model to forecast an AR(1) normal process.

Suppose that:

$$\begin{aligned} Y_1' &= R_{t+1} = b_0 + b_1 R_t + U_{t+1}; \quad U_{t+1} \sim N(0, \sigma_U) \\ Y_2' &= c_0 + c_1 R_t \end{aligned}$$

Then it follows that:

$$\left[\begin{array}{c} Y_1' \\ Y_2' \end{array}\right] \sim N\left(\left[\begin{array}{c} \frac{b_0}{1-b_1} \\ c_0 + c_1\frac{b_0}{1-b_1} \end{array}\right], \left[\begin{array}{cc} \frac{\sigma_U^2}{1-b_1^2} & c_1 b_1 \frac{\sigma_U^2}{1-b_1^2} \\ c_1 b_1 \frac{\sigma_U^2}{1-b_1^2} & c_1^2 \frac{\sigma_U^2}{1-b_1^2} \end{array}\right]\right)$$

Substituting this back into (2.30), we can obtain exact values for the mean profits obtained from using an AR(1) forecast for an AR(1) series:

$$\begin{aligned} E(X) = \frac{b_0}{1-b_1}\left[1 - \Phi\left(-\left(\frac{c_0}{|c_1|} + \frac{b_0}{1-b_1}\right)\frac{\sqrt{1-b_1^2}}{\sigma_U}\right)\right] \\ + \frac{\sigma_U b_1}{\sqrt{2\pi\left(1-b_1^2\right)}} \exp\left(-\left(\frac{c_0}{|c_1|} + \frac{b_0}{1-b_1}\right)^2 \frac{b_1^2 - 1}{\sigma_U^2}\right) \end{aligned}$$

For the parameters in Simulation 4.1, ($b_0 = 0.0015$; $b_1 = 0.0330$; $\sigma_U = 0.0108$) the profits from an optimal forecasting model ($b_0 = c_1, b_0 = c_1$) are:

$$E(X) = 2.3428 \cdot 10^{-4}$$

Chapter 3

Learning to profit with discrete investment rules

SUMMARY

In this chapter we attempt to explain the widespread use of discrete investment rules such as technical trading and market timing rules based on the fact that good rules of this form are relatively easier to learn. We show that a simple behavioural learning model (based on an estimator discussed in detail in Chapter 4) provides a good algorithm (from both a normative and positive perspective) with which investors' might learn good investment decision rules. In the context of this model and in a mean-variance optimisation framework we show formally that investment rule discretisation both simplifies and robustifies learning. Results are somewhat specific to the particulars of the proposed model, but are intended primarily to organise ideas for further development in future research.

3.1 Introduction

It is a fact that many popular investment decision rules are discrete in nature. As discussed in Chapter 1, many technical trading rules are binary as are various market timing rules. This is puzzling since it is hard to believe that such rules are indeed optimal for a broad range of investors and it is not obvious why deviations from optimality in this direction might arise.

In this Chapter we argue that these deviations are quite natural because they substantially facilitate a particular mode of learning at little cost to the investor. Furthermore, we argue that this mode of learning is a good description of how investors actually learn, not only because it is evolutionary (and hence closer to the way humans actually learn than learning 'like

[0] An early version of this paper appears in the Proceedings of the Sixth International Conference on Forecasting Financial Markets. The current version is still preliminary.

an econometrician') but also because it is potentially *better* than standard econometric learning methods in investment contexts.

Put more abstractly, our goal is to use a learning model to explain a puzzling empirical fact pertaining to individuals' observed behaviour. This is in contrast to the standard use of learning models which is to provide better explanations of various aggregate economic phenomena (Sargent 1993) such as the behaviour of financial time-series. To the best of our knowledge, this is the first attempt to use learning models for this purpose and this methodological contribution provides further motivation for this Chapter. As a side-product, we also deliver normative results in the form of an effective learning model for discrete decision rules that may be of use to investors (such as technical analysts) who are committed to the use of such rules.

In order to obtain our results, we restrict our attention to a specific decision environment. We consider agents using binary decision rules to optimise mean-variance objective functions. We propose that agents learn to choose their rule without explicit beliefs about the environment in which they operate. Instead, agents learn by using the rule that would have led them to maximise their utility had they used it in the past. This is consistent with an estimation method analysed in Chapter 4 and used by the Artificial Technical Analyst in Chapter 1. It is closely related to the behavioural learning models analysed by (Easley & Rustichini 1999) which are important because they come close to being consistent with features of observed learning behaviour.

In the context of a mean-variance utility maximiser constrained to use binary decision rules[1] learned using our behavioural/evolutionary method, we show that: (a) This learning method is good (generally better) relative to the standard benchmark of least-squares learning; and (b) The binary decision rules economise in 'rationality resources' relative to the use of continuous decision rules.

The next section introduces our simple model for investment decisions. In Section 3 we report results of simulations with which the profitability of alternative learning methods are compared and in Section 4 we discuss the merits of the learning method we propose. The paper is concluded with a summary of our main results and some conjectures they suggest.

[1] The study of binary (long-short) decision rules is common; for example, they arise in studies of technical trading (Sullivan, Timmermann & White 1998) and market timing rules (Henriksson & Merton 1981) and are constructed from forecasts based on a variety of econometric models (see for example Breen et al. (1989) and Pesaran & Timmermann (1995)).

3.2 Mean-variance investments with binary parametric decision rules

Consider an agent solving the following single-period cash-asset allocation problem at time t:

$$\begin{aligned} & \max_{a \in \mathcal{A}} E\left[W_{t+1}|x\right] - \alpha V\left[W_{t+1}|x\right] \qquad (3.1) \\ W_{t+1} &= W_t \cdot a \cdot (1 + R_{t+1}) + W_t \cdot (1 - a) \end{aligned}$$

where $E\left[*|x\right]$ and $V\left[*|x\right]$ denote mathematical expectation and variance of a random variable conditional on realisations $x \in \mathcal{X}$ of the 'information' random variable X_t, W_t is wealth in period t, R_{t+1} are returns to the risky asset (or portfolio of assets) and $\alpha > 0$ is a parameter determining the degree to which this agent is averse to risk (variance). The action a is restricted by the agents' budget constraint to lie in a subset $\mathcal{A}$ of $\mathbb{R}$.

In what follows, we make the following assumptions:

Assumption 3.1 *The agent knows of a parametric 'action rule' model* $a : \mathcal{X} \times B \rightarrow \mathcal{A}$ *such that for some* $b \in B$, $a\left(*, b\right)$ *is the solution to (3.1) almost everywhere on* $\mathcal{X}$.

Assumption 3.2 *The agent discretises his choice set by restricting* $\mathcal{A}$ *to a binary set of symmetric positions,* $\mathcal{A} = \{-k, k\}$.[2]

As we have noted, many agents actually do discretise their choice sets and we will provide an explanation for this in Section 4. For now note that assuming a binary choice set is particularly convenient since it allows us to obtain an utility-based performance measure for action rules that does not depend on α - and hence is common to all mean-variance agents, regardless of their degree of risk aversion.

3.2.1 Interpretation of decision rules

On the basis of Assumptions 1 and 2, we may write a as a function:

$$a\left(x, c\right) \equiv sign\left(m\left(x, c\right)\right) \equiv \left(2 \cdot I\left[m\left(x, c\right)\right] - 1\right) \cdot k,\ x \in \mathcal{X} \qquad (3.2)$$

where $I\left[*\right]$ is a sign indicator function taking the value one if the value in the brackets is positive and zero otherwise.

Two important interpretations we may give to $m\left(*, *\right)$ are:

- A parametric forecasting model for returns.
- A technical trading rule or market timing rule.

[2] Note that if the agent is risk neutral ($\alpha = 0$ in (3.1)) we can weaken this assumption to one of symmetry of $\mathcal{A}$ around zero without affecting the solutions of (3.1).

3.2.2 A common utility-based performance measure for decision rules

It is straightforward to show that the law of iterated expectations and Assumption 1 imply that a necessary and sufficient condition for $a(*, b)$ to be a solution to (3.1) is that :

$$\begin{aligned} b &\in \arg\max_{c \in B} E[W_{t+1}] - \alpha V[W_{t+1}] \\ W_{t+1} &= W_t \cdot a(X, c) \cdot (1 + R_{t+1}) + W_t \cdot (1 - a(X, c)) \end{aligned} \tag{3.3}$$

where $E[*]$ and $V[*]$ respectively denote the unconditional expectation and variance of a random variable.

Proposition 3 of Chapter 1 shows that when $E(R_{t+1}) > 0$, Assumption 2 implies:

$$\begin{aligned} \arg\max_{c \in B} E[W_{t+1}] - \alpha V[W_{t+1}] &= \arg\max_{c \in B} E[W_{t+1}] \\ W_{t+1} = W_t \cdot a(X, c) \cdot (1 + R_{t+1}) + W_t \cdot (1 - a(X, c)) & \end{aligned} \tag{3.4}$$

Hence, using also the linearity of the expectations operator, it follows that $a(*, b)$ is a solution to (3.1) if and only if:

$$b \in \arg\max_{c \in B} E\{R \cdot I[m(X, c)]\}. \tag{3.5}$$

So under our assumptions, the performance of a parametric decision rule for *any* mean-variance agent can be expressed as a function of a *common* profitability measure Π for each parameter c such that:

$$\Pi(c) = E\{R \cdot I[m(X, c)]\}. \tag{3.6}$$

3.3 Learning to parametrise discrete decision rules

Let us now assume that agents must make decisions on the basis of a finite history of observations of the state variables.

Assumption 3.3 *A sequence of draws $\{r_n, x_n\}_{n=1}^N$ from the distribution of (R, X) are available to an agent who wishes to solve (3.1) but has no prior beliefs regarding this distribution.*

This section discusses competing methods for formulating decision rules $d : \mathcal{X} \times (\mathbb{R} \times \mathcal{X}^N) \to \mathcal{A}$ interpreted as methods for learning the optimal action rule $a(*, b)$. Such methods can be viewed as statistical decision functions as defined by Wald (1971) and their use can be associated with what Neyman (1938) referred to as *inductive behaviour*.

3.3.1 Performance of 'econometric', least squares learning.

It is possible to estimate $m(*, c)$ as a forecasting model for $E(R|*)$ and use the forecasts thus derived in the decision rule d. For example, the popular 'least squares learning' model used for example by Bray & Savin (1986), Fourgeaud, Gourieroux & Pradel (1986) and Marcet & Sargent (1989) is precisely in this spirit. If the model is correctly specified[3], i.e. $m(x, b) = \mu_R(x)$ for some $b \in B$ and for all $x \in \mathcal{X}$, then standard regularity conditions ensure that the OLS estimator b^{ls} will converge to b as N tends to infinity. The following is a simple simulation designed to measure the profitability of this learning method in finite samples.

Simulation 3.1

Consider the following DGP used also in Simulation 5.1 of Chapter 2:[4]

$$\begin{aligned} R_{t+1} &= 0.00015 + 0.0330 \cdot R_t + U_{t+1} \\ & U_{t+1} \overset{iid}{\sim} N(0, 0.0108). \end{aligned} \tag{3.7}$$

The parametric model to be learned by OLS is

$$m(R_t, c) = c_0 + c_1 R_t.$$

We will denote the OLS estimated parameters $\left(b_0^{ls}, b_1^{ls}\right)$. The model used in the discrete decision rule of an agent operating by least squares learning is:

$$d\left(R_t, b^{ls}\right) = I\left[b_0^{ls} + b_1^{ls} \cdot R_t\right] \tag{3.8}$$

Assuming $b_1^{ls} > 0$ and that $b^{ls} \equiv \frac{b_0^{ls}}{b_1^{ls}}$ is independent of R, the expected profits Π of this rule are given by (see (3.6) and (3.7)):

$$\Pi\left(b^{ls}\right) = \int_{-b^{ls}}^{\infty} (0.00015 + 0.0330 \cdot R_t) \cdot dF_{R_t} \tag{3.9}$$

where F_{R_t} is the marginal c.d.f. of R_t. An exact analytical expression for (3.9) in terms of the underlying parameters is given in Appendix B of Chapter 2.

Evidently, the profitability of an estimate $\left(b_0^{ls}, b_1^{ls}\right)$ depends only on b^{ls}. The closeness of b^{ls} to $b \equiv \frac{b_0}{b_1}$ (the true parameter) therefore serves as a good

[3] Note this is much stronger than Assumption 1.

[4] The parameters of this DGP were fixed using OLS to estimate an AR(1) model on a series drawn from the empirical distribution of 2049 observations on IBM daily closing prices from 1st January 1990 through to 6th November 1997. (Source: *DATASTREAM*).

indicator of the profitability of this estimation technique. Unfortunately, b^{ls} does not inherit the neat convergence properties of $\left(b_0^{ls}, b_1^{ls}\right)$ because it is a ratio of normally distributed random variables which will therefore have an infinite variance.[5] The expected profitability from using the least squares estimator is given by

$$\Pi^{ls} = \int \Pi\left(b^{ls}\right) dF_{b^{ls}}$$

where $F_{b^{ls}}$ is the c.d.f. of the least squares estimator.

We took 10000 draws from the distribution of $F_{b^{ls}}$ for various sample sizes N and computed estimates for the mean and standard deviation of $F_{b^{ls}}$ as well as the associated profits Π^{ls} from using the least squares estimator. These are given in Table 3.1.

N	100	200	400	800	1600	3200	6400	∞
$10^3 \cdot \widehat{\mu}_{b^{ls}}$	-14.338	-4.1141	-10.918	1.4320	6.9320	8.3079	6.2292	4.5455
$10^2 \cdot \widehat{\sigma}_{b^{ls}}$	79.465	47.077	61.121	38.030	33.415	30.628	5.8787	-
$10^4 \cdot \widehat{\Pi}^{ls}$	0.99471	1.1420	1.3233	1.5401	1.8026	2.0400	2.1994	2.3428

Table 3.1 Each column corresponds to a sample size N for which we compute the mean $\widehat{\mu}_{b^{ls}}$ and standard deviation $\widehat{\sigma}_{b^{ls}}$ of the simulated distribution of b^{ls}. The column corresponding to ∞ gives the true value b and the profits obtained if this value is used.

3.3.2 Performance of a 'behavioural' learning method.

Suppose the agent uses the rule that has performed best in terms of in sample profitability $\Pi(c)$. By (3.6), the decision rule thus obtained will be parametrised by b_N solving:

$$\max_{c \in B} E_N \left\{R \cdot I\left[m\left(X, c\right)\right]\right\} \tag{3.10}$$

where E_N is the empirical mean given a sample of size N.

The properties of this estimator are analysed in Chapter 2 where we show that under certain regularity conditions (satisfied by the DGP considered in our simulations) the estimator is asymptotically consistent. Furthermore, simulations here suggested that b_N converges at cube-root rate to a non-normal distribution.[6]

[5] This problem is exacerbated by the fact that b_0 and b_1 are small relative to the variance of U and therefore their own variance. Hence the ratio $\frac{b_0}{b_1}$ is particularly ill-behaved since with large probability both b_0 and b_1 will be close to zero. Such parametrisations are natural in financial series because they have a very low signal-to-noise ratio. It may be that in other environments, least squares learning performs somewhat better.

[6] This is perhaps due to its structural similarity to the maximum score estimator of (Manski 1975).

While the convergence rate of b_N to B is slower than the rate of convergence of $\left(b_0^{ls}, b_1^{ls}\right)$ to (b_0, b_1), this does not mean that b^{ls} converges faster than b_N to B. It may therefore be the case that use of the estimator b_N is more profitable in finite samples. We confirm this is the case in reasonably sized samples with the following simulation.

Simulation 3.2

Using the same draws as in Simulation 3.1, we examine the distribution of b_N and its profitability - which is given by an expression analogous to (3.9). Results are reported in Table 3.2 which consolidates Table 3.1 of this Chapter and Table 2.1 of Chapter 2.

N	100	200	400	800	1600	3200	6400	∞
$10^3 \cdot \widehat{\mu}_{b^{ls}}$	-14.338	-4.1141	-10.918	1.4320	6.9320	8.3079	6.2292	4.5455
$10^3 \cdot \widehat{\mu}_{b_N}$	1.2797	1.3502	2.4048	2.9877	3.9049	4.099	4.4858	4.5455
$10^2 \cdot \widehat{\sigma}_{b^{ls}}$	79.465	47.077	61.121	38.030	33.415	30.628	5.8787	-
$10^2 \cdot \widehat{\sigma}_{b_N}$	1.3603	1.3378	1.2694	1.1699	1.0162	0.86258	0.69860	-
$10^4 \cdot \widehat{\Pi}^{ls}$	0.99471	1.1420	1.3233	1.5401	1.8026	2.0400	2.1994	2.3428
$10^4 \cdot \Pi_{b_N}$	1.7105	1.7328	1.8048	1.8772	1.9784	2.0592	2.1408	2.3428

Table 3.2 Each column corresponds to a sample size N for which we compute the mean and standard deviation of the simulated distribution of b_N and b^{ls}. The column corresponding to ∞ gives the true value b.

3.3.3 The normative advantages of 'behavioural' learning

Our analysis and simulations suggest that, for the restricted decision problem we have considered:

- The estimator based on parametrising decision rules is more profitable in reasonably sized samples. A heuristic account of why this occurs is that, while it may be less efficient at exploiting the information in a sample $\{r_n, x_n\}_{n=1}^N$, it complements the information it does extract with information about the form of the decision rule $d(*, c)$ in a way that makes it effective at determining the optimal decision rule $d(*, b)$. By contrast, 'statistical' forecasting methods can only exploit information in $\{r_n, x_n\}_{n=1}^N$ and therefore may be at a disadvantage even when they perform this task 'well'.

- Learning methods based on estimators of parametric models of the underlying environment are generally consistent when these models are correctly specified. Correct specification of a parametric decision rule for an optimal decision rule is a weaker condition as discussed also in Chapter 2 and this means that convergence to the optimal rule may occur more generally with behavioural-type learning algorithms.

- When the assumption of correct specification is not satisfied, the rule learned is nevertheless the 'best' permitted by the model class, i.e. it minimises unconditional Risk (see Section 4.3 of Chapter 2). By contrast, the rule obtained by least squares has no natural interpretation.
- The estimator can serve to parametrise decision rules based on technical trading rules and market timing rules which do not have interpretations as forecasts for returns and for which methods such as least squares learning are inapplicable.

Further study of these issues is of course warranted, particularly to compare the performance of these methods on real data.

3.4 An explanation of the prevalence of discrete decision rules

It is easy to see how agents may come to use parametrised decision rules: it is sufficient that through experience they obtain enough knowledge to reduce uncertainty about their optimal decision rule to uncertainty about the value of a finite dimensional parameter. It is much harder to explain why the decision rules they use are discrete, even though this often turns out to be the case in reality.

Here we provide an explanation in terms of 'bounded rationality' by arguing that the 'rationality resources' conserved using the discretisation may be larger than the utility losses incurred because of its use. While the explanation necessarily remains informal because of the lack of an applicable formal theory of 'rationality resources', we can make a qualitative comparison of the effects of the discretisation in the context of the decision model introduced in Section 2.

At first sight, one may observe that the utility loss due to discretisations may be 'small' but, on the other hand, it is unlikely that use of continuous decision rules is substantially more costly than the use of discrete decision rules. However, when decision rules need to be *learned* we show that the advantages to using a discrete set of actions may be substantial. This is shown in the context of the behavioural model for learning introduced in Section 3. We have already discussed some of the advantages of this type of model; the fact that it helps us explain the observation of discrete decision rules is an important additional advantage.

In the context of our model, discretisation facilitates learning because it:

1. **Reduces the dimensionality of information required for optimal action.** Information about the state variable's distribution necessary for the agent to act optimally may be *drastically reduced* and

this may greatly simplify the aspects of the environment that need to be learned. This is illustrated in our model in which we show that discretisation makes the solution to (3.1) independent of α and hence makes learning about second moments unnecessary.

2. **Robustifies the optimal rule with respect to the objective function**. Optimal discrete rules may be *invariant* within some large class of objective functions. This is also illustrated in our model because the same rule is optimal for agents with different variance-aversion α. This facilitates learning because:

   (a) It *facilitates 'learning from others'*. In particular, it makes it possible for an agent to act optimally by copying what more experienced agents with unknown preferences are doing or have done.

   (b) It means that the *same rules remain optimal* even if the agent's objective function changes so that the agent does not need to learn from scratch. This may occur either because of changes in the agent's preferences or because of changes a principal makes to his incentives.

3. **Robustifies performance with respect to the DGP.** The robustness of a rule in relation to, possibly non stationary, unknown or even unknowable aspects of the DGP (e.g. its tails) interact with decisions may be altered so that smaller utility losses are incurred by the agent. For example, in our model it is only knowledge of the ratio $\frac{b_0}{b_1}$ which is necessary to derive an optimal decision rule. As long as this can be determined, the precise values of (b_0, b_1) are unnecessary; methods which attempt to determine (b_0, b_1) such as OLS can lead to poor estimates for $\frac{b_0}{b_1}$ - which is what matters.

It should be noted that these advantages were particularly important in the pre-computer era when learning was particularly costly and the properties of financial series were hardly understood. Once the use of discrete decision rules had been established, there was likely to be lock-in to their use. The reason for this is that accumulated knowledge which is useful within the discretised decision framework (as determined by restriction of choice to a 'narrow' parametric class of decision rules) may be difficult to translate into valuable information for decision making in other settings. For example, knowledge that a technical trading rule is profitable may not be easy to translate into knowledge about the distribution of returns which would be useful for a Bayesian agent using continuous decision rules. Hence we should not necessarily expect improved knowledge about properties of decision rules to necessarily weaken the incentives for using the discretisation and we do not expect discretisation to be a temporary phenomenon.

3.4.1 On further positive applications of behavioural learning models

As we have previously mentioned, behavioural learning methods such as the one proposed in Section 2 is consistent with what are believed to be some of the 'stylised facts' of how people actually learn. By serving as a model of how agents *do* learn, rather than as an *as if* model for agent behaviour, it can provide sharper characterisations of both mico and macroeconomic aspects of the markets in which they operate. At the macro level, for example, it can help us understand how information is incorporated into prices and quantify the degree of a market's efficiency (Section 5, Chapter 1). At the micro level, we have argued that it can help answer a variety of 'puzzles' concerning the behaviour of investors such as why they use discrete rules, such as technical trading and market timing rules. We look further applications in this direction.

3.5 Conclusion

Our goal has been to explain the widespread use of discrete decision rules by showing that it facilitates learning. To do this, we have had to utilise an appropriate model of learning in a particular decision setting. Our learning algorithm is based on the Risk Neutral Forecasting estimator of Chapter 2 which we emphasise is an appropriate way of learning the parameter of a discrete decision rule. Using simple simulations, we show that our estimator performs well as a normative model for learning. We also discuss why it is useful in positive modelling of certain aspects of investment behaviour.

While our results are provided within the context of a fairly narrow decision model, we conjecture that some of them generalise broadly. In particular, suppose a learning problem can be reduced to a 'parametric learning' problem, i.e. a problem of learning a parameter b of the state variables' DGP. Then instead of constructing a model for the DGP, it may be preferable to use the fact that the parameter b solves the agent's decision problem:

$$\max_{c \in B} E\left[U\left(X, d\left(X, c\right)\right) | X_t\right]. \tag{3.11}$$

to develop a natural learning method for b in the form of an estimator:

$$b_N \in \arg\max_{c \in B} E_N\left[U\left(X, d\left(X, c\right)\right)\right] \tag{3.12}$$

Correspondingly, a natural decision rule for this agent is $d\left(*, b_N\right)$.
The reason this is a 'good' learning method is because:

1. It leads to decisions which have a behavioural foundation, i.e. learning is 'realistic', and hence are useful as positive models of investors' learning.

2. b_N is an asymptotically consistent estimate for b under general conditions.

3. We conjecture that for small enough sample sizes (or when there is appropriate non-stationarity in the DGP):

$$E\left(U\left(X, d\left(X, b_N\right)\right)\right) > E\left(U\left(X, d\left(X, \widehat{b}\right)\right)\right)$$

where $\widehat{b}$ is a classical estimate for b obtained with a learning method lacking a 'behavioural foundation' such as OLS. This is because while other estimators may be more efficient at exploiting a sample's information about the distribution of state variables, they do not exploit information about how the estimated parameter will be used, which is most relevant.

4. If (unlike in this paper) $d(*, b)$ is a continuous function of b for all x, this learning method reduces to the application of the method of moments to the first order conditions of (3.12). The well-understood regular properties of this estimator will therefore also apply to the learning method and its analysis will be greatly simplified. Efficiency gains might be obtained by using the Generalised method of moments.

These conjectures will be studied in future work. Furthermore, a model of a financial market in which agents learn using a method of this form will be developed in an attempt to replicate some of the outstanding features of financial series.

Chapter 4

Estimators for the sign of a regression mapping

SUMMARY

In this Chapter we show that there exist a number of important applications in which the interest of a statistician or an economic agent focuses on the sign of a (mean) regression mapping rather than the mapping *per se*. We propose three related estimators that are robust with respect to model mis-specification in that only a weak specification condition is required for consistency in the relevant regression attribute. Furthermore, they are also more parsimonious than standard estimators which may be poor estimators of a regression's sign. The relative merits of each estimator is discussed and simulations are used to study some intractable properties of one of the estimators.

4.1 Introduction

Consider the random variables Y and $X \in \mathcal{X}$ related by:

$$\begin{aligned} Y &= \mu_Y(X) + U, \\ E(U|x) &= 0, \end{aligned} \tag{4.1}$$

where $\mu_Y : \mathcal{X} \to \mathbb{R}$ is an unknown and for now unrestricted mean regression mapping, U is an unobserved scalar disturbance term and $E(U|X)$ is the expectation function of U conditional on realisations x of X. It will be unnecessary to make any further assumptions about the behaviour of $F_Y|x$, the conditional distribution of Y given x. The 'sign of the regression mapping

$\mu_Y(*)$' is an indicator function $I_{\mu_Y}(*) \equiv I(\mu_Y(*))$ where:[1]

$$I(z) \equiv \left\{ \begin{array}{c} 1 \text{ if } z > 0 \\ 0 \text{ otherwise} \end{array} \right\}.$$

In this paper we deal with the problem of estimating I_{μ_Y} using N draws $\{y_n, x_n\}_{n=1}^N$ from the unconditional distribution F of the vector random variable (Y, X). The estimators we will propose are best characterised as semiparametric since they require restriction of uncertainty about I_{μ_Y} to a finite parameter c, yet do not require parametric restrictions on U. Evidently, they will be applicable quite broadly.

To the best of our knowledge, we are the first to focus on this problem, so we devote some effort to discussing its importance. We show that there are a number of applications in which the sign of a regression mapping is the object of interest. In particular, many decision problems with discrete choice sets or discrete solutions can be solved if the sign of a regression mapping is known. We provide a simple example of a job search decision (an optimal stopping problem) in which the choice set is discrete (accept or reject a job at a given wage) and the optimal decision is a function of the sign of a regression mapping. We also discuss an important generalisation of the risk neutral investor decision problem of Chapter 2 to the case where there is an interest bearing riskless asset and transaction costs. The solutions of this problem are functions of the sign of certain regression mappings. Finally, we show that *calibration* (Eisenhart 1939) and *econometric equation inversion* (Hendry & Ericsson 1991) can be seen as problems requiring the estimation of a regression mapping's zeros and discuss the advantages of doing this using the methods introduced in this paper.

Estimation of the sign of a regression mapping may also be useful to an econometrician when the regression mapping μ_Y is of interest but models based on different assumptions lead to different estimates for μ_Y. In this case, it can serve as an informal specification check or model selection criterion since it can be used to check which estimates for μ_Y best agree with an estimated model for its sign.[2]

The standard approach to estimating the sign of a mean regression mapping would be to view this as a particular function of the regression model parameters. Estimation of the sign can thus be transformed to the problem of estimating a function of model parameters rather than the parameters themselves and a number of (primarily Bayesian) techniques for achieving this general objective have been developed (Zellner 1978, Park & Zellner 1979). The contribution of this Chapter is the introduction of novel estimators

[1] This definition of the sign of g distinguishes only between positive and non-positive values of g. In Section 4 we will also consider situations where the further distinction between negative and zero values of g must be made.

[2] Analogously, Stoker (1986) proposes to use estimates for the *scale* of parameters as a specification check for models used to estimate the *values* of the parameters.

with desirable properties obtained by exploiting special properties of the *particular* (sign) function of the parameters in which we are interested.

Estimation of the sign is closely related to estimation of the zeros of a regression function which is a problem that has attracted some interest. In particular, Haerdle & Nixdorf (1987) and Tsybakov (1988) have derived computationally convenient recursive nonparametric estimators for the zeros of a regression function. They note that nonparametric estimation of a regression function requires a great deal of learning which is redundant when only its zeros are of interest. Exploiting this fact, they provide estimators which learn only about the regression functions' zeros and are therefore *computationally cheaper*. Similarly, we exploit the fact that our interest is concentrated on the sign of a regression mapping to develop parametric estimators that require little prior information on model specification, are relatively parsimonious and deliver accurate estimates of the desired attribute (the sign).

This chapter is organised as follows. In Section 2, we discuss reasons for which less prior information (in a sense to be formalised) is required to correctly specify a model for I_{μ_Y} than for μ_Y and propose three consistent estimators for I_{μ_Y}. In Section 3, we discuss further properties of our estimators and reasons for which they may be better than estimators derived from a model for μ_Y even when it is correctly specified. In Section 4 we motivate our focus on the sign of regression mappings by discussing a number of significant applications in which knowledge of a regression mapping's sign is the object of interest. We conclude with a survey of our results and a discussion of plans for future research. An Appendix follows containing all lemmata used in the proofs of our results.

4.2 Estimators for the sign of regression mappings

4.2.1 Model specification assumptions

Our analysis will be restricted to the situation where a potentially misspecified parametric model $\mathcal{G}$ for $F_Y|x$ is postulated, taking the form:

$$\mathcal{G} = \left\{ G_Y|\left[x, c\right], c \in B \subseteq \mathbb{R}^k \right\}$$

The conditional mean of the model with parameter c is:

$$m\left(x, c\right) = \int Y \cdot dG_R|\left[x, c\right]$$

Our analysis will be restricted to pairs $(\mathcal{G}, F_Y|x)$ such that $m : \mathcal{X} \times B \to \mathbb{R}^1$ is correctly specified for the sign of the conditional mean on B - that is,

for some $b \in B$, $I(m(x,b)) = I_{\mu_Y}(x)$ for almost all x.[3] We formalise our specification requirement in the following assumption.

Assumption 4.1 *Let $m : \mathcal{X} \times B \to \mathbb{R}^1$ be a model such that for some $b \in B$, $m(x,b)$ has the same sign as $\mu_Y(x)$ almost everywhere, i.e.:*

$$m(x,b)\, w(x) = \mu_Y(x) \quad a.e., \tag{4.2}$$

for some strictly positive mapping $w : \mathcal{X} \to \mathbb{R}^1_{++}$.

An example illustrating the weak restriction this imposes on the relation of m and μ_Y is given in Figure 4.1 below.

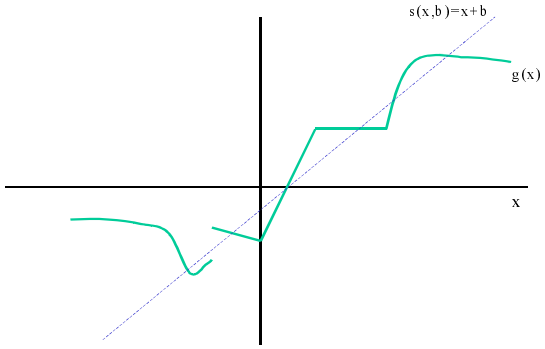

Figure 4.1. Two functions $m(*,b)$ and $\mu_Y(*)$ satisfying Assumption 4.1.

Correct specification of $m(*)$ as a model for μ_Y is much stronger than Assumption 4.1 in that it requires (4.2) to hold for $w(x) = 1$ rather than for some $w : \mathcal{X} \to \mathbb{R}^1_{++}$. It is easy to imagine a situation where we cannot be confident that a parsimonious model is correctly specified for the conditional mean yet we can postulate such a model for the sign of the conditional mean. For example, if x is a scalar and μ_Y is believed to be a highly nonlinear and discontinuous function satisfying a single-crossing condition at zero (e.g. Manski & Thompson 1989) with $\lim_{x \to -\infty} \mu_Y(x) < 0$, then Assumption 4.1 is satisfied for $m(x,c) = c + x$. If the sign of the regression mapping is of interest, direct estimation of the model $I(x+c)$ would be preferable to an estimate for I_{μ_Y} derived from a model for μ_Y that would be mis-specified and might therefore lead to inconsistent estimates.

It should be noted that given our weak specification requirement, models in $\mathcal{G}$ need not be specified beyond models for $m(*,*)$. We may therefore take a semi-parametric perspective and think of our analysis as being the modelling of conditional means (in which case we would think in terms of

[3] This is a requirement of correct specification for a very particular conditional attribute of a model. See White (1994) for definitions of correct specification in other more standard conditional attributes, such as the mean or median.

the pair $(m(*,*), \mu_Y(*))$ or even the modelling of the signs of conditional means (in which case we would think in terms of the pair $\left(I(m(*,*)), I_{\mu_Y}\right)$). For notational simplicity, we will denote $I(m(x,c))$ as $I(x,c)$;

4.2.2 Sufficient conditions for estimator consistency

We now introduce some further assumptions to be used in proving the asymptotic consistency of our estimators. The estimators we propose are all moment extremum estimators, based on conditions known to be satisfied by I_{μ_Y}. Having introduced the estimators and having shown they are consistent, we will discuss their relationships in the next Section.

Assumption 4.2 (Compactness) *The parameter space $B \subseteq \mathbb{R}^k$ is compact (or discrete, in which case only the next assumption is necessary).*

Assumption 4.3 (SLLN) *The draws $\{y_n, x_n\}_{n=1}^N$ from F satisfy a strong law of large numbers (a variety of which can be found in White (1984)).*

Assumption 4.4 (Identifiability) *There exists a unique $b \in B$ s.t.:*

$$-\int Y \cdot I(X,b)\, dF = \min_{c \in B} -\int Y \cdot I(X,c) dF$$

Whether this identifiability assumption holds will depend on the interaction of $\{m, F, B\}$ and must be ensured on a case-by-case basis by appropriate specification of $m(*)$ and B given our priors regarding the behaviour of F. We derive some sufficient conditions for this in Appendix B.

Assumption 4.5 (Equicontinuity) *There exists a strictly positive mapping $w : \mathcal{X} \to \mathbb{R}^1_{++}$ such that $w(x)\, m(x,c)$ is equicontinuous on B, i.e. $\forall$ $\alpha > 0$,*

$$\exists\ \gamma_\alpha : |a - c| < \gamma_\alpha \Rightarrow \sup_{x \in \mathcal{X}} |w(x)\, m(x,a) - w(x)\, m(x,c)| < \alpha, \quad (a,c) \in B \times B$$

*Equicontinuity of a given function such as $m(*,*)$ is a directly verifiable condition; however, since $w(*)$ will usually be unknown, equicontinuity involves an implicit assumption about the form $w(*)$ can take for a given specification of $m(*,*)$. Sufficient conditions for this condition are provided by Manski (1988*a*, Lemma 7, pp. 109-110) reproduced for reference purposes in Appendix B. The role of this assumption is to introduce appropriate* smoothness *in $\int Y \cdot I(X,c) dF$ that does not depend on the behaviour of F.*

Assumption 4.6 (Boundary) *The following boundary condition is satisfied:*

$$\lim_{\alpha \to 0} \sup_{c \in B} \int |Y| \cdot I(\alpha - |m(X,c)|)\, dF = 0,$$

This is an assumption that ensures that the probability of drawing x's such that $m(x,c)$ is close to zero for all c is small in an appropriate sense. It serves to ensure *continuity* of $\int Y \cdot I(X,c)dF$.

It is relevant to note that if $m(*,c)$ is linear (for example because we restrict our attention to best linear prediction) A4.4-4.6 become immediately satisfied under regularity conditions given in Appendix B, but *identification can only be to scale* so B must not include c and c' such that $c = ac'$ (where a is a positive scalar). When on the other hand $m(*,c)$ is non-linear it is difficult to derive conditions on F ensuring A4-6 will hold. We note that when the parameter set B is discrete (as in Chapter 1 where the Artificial Technical Analyst estimates technical trading rules using the first of the following estimators) *consistency only requires A3.*

4.2.3 A step function M-estimator

The following proposition provides a simple estimator for the sign of a regression mapping which resembles Manski's (1975) maximum score estimator. A theorem of Manski (1988*a*) is exploited to ensure consistency of the estimator under the assumptions we have introduced. This theorem essentially extends standard conditions under which M-estimators (Huber 1996) are consistent to the case where they are step functions of parameters.

Proposition 4.1 *Let*

$$b_N \in \arg\min_{c \in B} - \int Y \cdot I(X,c)\, dF_N, \tag{4.3}$$

where F_N is the empirical c.d.f. of $\{y_n, x_n\}_{n=1}^N$.

Under Assumptions 4.1-4.6,

$$\Pr\left(\lim_{N\to\infty} |b_N - b| = 0\right) = 1 \tag{4.4}$$

where $I(x,b) = I_{\mu_Y}(x)$ almost everywhere.

Proof. Theorem 3', Chapter 7 of Manski (1988*a*) applies to b_N and ensures

$$\Pr\left(\lim_{N\to\infty} |b_N - b| = 0\right) = 1$$

with

$$b = \arg\max_{c \in B} \int Y \cdot I(X,c)\, dF.$$

Lemma 1 in the Appendix ensures that the above is a necessary and sufficient condition for

$$I(x,b) = I_{\mu_Y}(x) \text{ a.e.}$$

■

Notice that the estimator b_N is defined as an element of a set which we show converges almost surely to a parameter b defining a model identical to the sign of the regression mapping almost everywhere. We have avoided measurability issues, but it should be noted that here we make the implicit assumption that b_N is chosen from this set in a manner ensuring b_N is a measurable function of (Y,X). As noted by Amemiya (1985) this is made possible by a theorem of Jennrich (1969, p. 637).

This is the estimator that has been used in Chapters 1,2 and 3 to estimate models to be used in investment decisions. It is a particular attractive estimator for that purpose since even when Assumption 4.1 is not satisfied, this estimator will converge to models that are 'good' in a precisely defined sense.

4.2.4 A quasi generalised step function M-estimator

A modification of the previous estimator yields another consistent estimator which is very closely related to what Gourieroux & Monfort (1995) refer to as a 'quasi generalised M estimator' (Definition 8.2, p. 214). The distinguishing feature of such estimators is that they optimise objective functions which themselves contain consistent estimates of certain nuisance parameters (in this case functions). The analysis of Gourieroux & Monfort (1995) does not apply to our estimator however, because it is a step function of the parameter. With the following proposition we establish the estimator's consistency.

Proposition 4.2 *Let $A : \mathcal{X} \to [0,1]$ be a mapping such that:*

$$A(x) \equiv \frac{E(|Y| \, | I(Y) = 1, x)}{E(|Y| \, | I(Y) = 1, x) + E(|Y| \, | I(Y) = 0, x)}, \; x \in \mathcal{X}. \tag{4.5}$$

Let $A_N(x)$ be an estimated model for $A(x)$ such that:

$$\Pr\left(\lim_{N \to \infty} \sup_{x \in \mathcal{X}} |A_N(x) - A(x)| = 0 \right) = 1.$$

Define b_N^a by:

$$b_N^a \in \arg\min_{c \in B} - \int (A_N(x) - I_Y) \cdot I(x,c) dF_N \tag{4.6}$$

Under Assumptions 4.1-4.6,

$$\Pr\left(\lim_{N\to\infty}|b_N^a - b| = 0\right) = 1 \tag{4.7}$$

where $I(x,b) = I_{\mu_Y}(x)$ *almost everywhere.*

Proof. Lemmata 3, 4 and 5 of Appendix A imply that:

$$\Pr\left(\lim_{N\to\infty}|b_N^a - b| = 0\right) = 1$$

where

$$b = \arg\min_{c\in B} - \int (A(x) - I_Y)\cdot I(x,c)dF.$$

Lemma 2 ensures that the above is a necessary and sufficient condition for

$$I(x,b) = I_{\mu_Y}(x) \text{ a.e.}$$

■

Interpretation as generalised binary quantile regression

If the 'nuisance function' $A(x)$ is a known constant ($A_N(x) = A(x) = a$) this estimator is identical to a standard estimator used in binary quantile regression, i.e. to fit a model for the a'th quantile of I_Y conditional on x (see e.g. Manski & Thompson 1989). This is because $I(x,b)$ will be the a'th quantile of I_Y. More generally, Proposition 2.3 of Chapter 2 implies that $I(x,b)$ will be the $A(x)$'th quantile of I_Y conditional on x. Clearly then estimating the sign of a regression mapping when $A(x)$ is a known constant can be achieved by binary quantile regression. When $A(x)$ is not a known constant, our problem can be approached using the estimator introduced here - which may be viewed as a generalisation of binary quantile regression.

One interesting application in which $A(x)$ may be a known constant is when Y are financial returns R_{t+1} in period $t+1$ and X are lagged returns. This is because it seems acceptable to assume that $|R_{t+1}|$ is independent of the sign of returns $I_{R_{t+1}}$ (Granger & Ding 1994*a*, Henriksson & Merton 1981) in which case $A(x) = \frac{1}{2}$ and for $A_N(x) = A(x) = \frac{1}{2}$ our estimator coincides with the maximum score estimator.

Remark 4.1 *If it is indeed believed that* $A(x) = a$ *for all* x*, there exist statistical techniques (Zheng 1998) which allow us to judge whether* $m(*,*)$ *is a correctly specified model for the* a*'th quantile of* I_Y *and hence evaluate whether Assumption 1 holds. It may be possible to extend these results to*

the case where $A(x)$ varies with x to provide a general specification test for models of the sign of a regression mapping. Given that many optimal decision rules of interest can be expressed as functions of the sign of a regression mapping, we would thus also obtain specification tests for models of optimal decisions.

Financial returns also provide series for which it may be possible to model $A(x)$ very effectively because the conditional expected value of absolute returns $E(|R_{t+1}||R_t, R_{t-1}, ...R_{t-T})$ are highly predictable (Taylor 1986, Schwert 1989, Granger & Ding 1994*b*, Mills 1996, Fornari & Mele 1994). This indicates that it may be feasible to accurately model

$$E\left(|R_{t+1}||I_{R_{t+1}}, R_t, R_{t-1}, ...R_{t-T}\right)$$

and hence also $A(R_t, R_{t-1}, ...R_{t-T})$. To be precise, what we would need is a model converging to $A(R_t, R_{t-1}, ...R_{t-T})$ uniformly on $(R_t, R_{t-1}, ...R_{t-T})$. While this may be an extreme assumption, it seems reasonable that with a good model for $A(x)$ this estimator might be an interesting competitor to the step function M-estimator of the previous section. Of course this remains to be verified in the context of a specific application.

4.2.5 A smoothed estimator

Horowitz (1992) provides a smoothed maximum score estimator which he shows has various desirable tractable properties that the maximum score estimator does not. Analogously, we smooth our simple step function M-estimator to obtain another consistent estimator.

Proposition 4.3 *Let $K : \mathbb{R} \to \mathbb{R}$ be a continuous function satisfying $|K(v)| < \infty$, $\lim_{v\to-\infty} K(v) = 0$, $\lim_{v\to\infty} K(v) = 1$.*

Let b_N^s be an estimator for b defined by:

$$b_N^s \equiv \arg\min_{c\in B} -\int Y \cdot K\left[\frac{m(X,c)}{\zeta_N}\right] dF_N$$

where $\{\zeta_N\}$ is a sequence s.t. $\zeta_N > 0$ and $\lim_{N\to\infty} \zeta_N = 0$.

Under Assumptions 4.1-4.6,

$$\Pr\left(\lim_{N\to\infty} |b_N^s - b| = 0\right) = 1 \tag{4.8}$$

where $I(x, b) = I_{\mu_Y}(x)$ almost everywhere.

Proof. Let

$$\begin{aligned} f(c) &\equiv \int Y \cdot I(X,c) dF, \\ f_N(c) &\equiv \int Y \cdot I(X,c) dF_N, \\ S_N(c) &\equiv \int Y \cdot K\left[\frac{m(X,c)}{\zeta_N}\right] dF_N. \end{aligned}$$

Notice that for any $\alpha > 0$:

$$\begin{aligned} |S_N(c) - f_N(c)| \leq & \int \left| Y \cdot \left[K\left(\frac{m(X,c)}{\zeta_N}\right) - I(X,c) \right] \right| \cdot I\left(|m(X,c)| - \alpha\right) dF_N \\ & + \int \left| Y \cdot \left[K\left(\frac{m(X,c)}{\zeta_N}\right) - I(X,c) \right] \right| \cdot I\left[\alpha - |m(X,c)|\right] dF_N \end{aligned}$$

Since $\Pr(m(X,c) = 0)$ and by our assumptions, K is bounded with

$$\lim_{N\to\infty} K\left(\frac{m(x,c)}{\zeta_N}\right) = \left\{ \begin{array}{l} 1 \text{ if } m(x,c) > 0 \\ 0 \text{ if } m(x,c) < 0 \end{array} \right\}$$

it follows that the first term on the RHS of the previous expression converges to zero uniformly on c as $N \to \infty$.

The second term is smaller than:

$$k \int |Y| \cdot I\left(\alpha - |m(X,c)|\right) dF_N$$

uniformly over c (where k is some positive constant).

By equicontinuity, there exists a finite δ_α such that for all a, c:

$$|a - c| < \delta_\alpha \Rightarrow |m(x,c) - m(x,a)| < \alpha$$

Fix $a > 0$. Then for $c \in B$ such that $|a - c| < \delta_\alpha$

$$\begin{aligned} & |m(x,c))| < \alpha \\ \Rightarrow & |m(x,c)| + |m(x,a) - m(x,c)| < 2\alpha \\ \Rightarrow & |m(x,a)| < 2\alpha \end{aligned}$$

Hence:

$$\begin{aligned} & |a - c| < \delta_\alpha \\ \Rightarrow & \int |Y| \cdot I\left(\alpha - |m(X,c)|\right) dF_N \leq \int |Y| \cdot I\left(2\alpha - |m(X,a)|\right) dF_N \end{aligned}$$

Since B is a compact set, we may define the finite set B_α such that for all c in B there is an a in B_α such that $|a - c| < \delta_\alpha$.

Then there exists an a in B_α such that for all c in B :

$$\int |Y| \cdot I\left(\alpha - |m(X,c)|\right) dF_N \leq \int |Y| \cdot I\left(2\alpha - |m(X,a)|\right) dF_N$$

So for all c in B :

$$\int |Y| \cdot I\left(\alpha - |m(X,c)|\right) dF_N \leq \max_{a \in B_\alpha} \int |Y| \cdot I\left(2\alpha - |m(X,a)|\right) dF_N$$

By the Strong Law of Large Numbers, as $N \to \infty$,

$$\max_{a \in B_\alpha} \int |Y| \cdot I\left(2\alpha - |m(X,a)|\right) dF_N \to \max_{a \in B_\alpha} \int |Y| \cdot I\left(2\alpha - |m(X,a)|\right) dF \quad \text{a.s.}$$

Hence, for all c in B, for $N > N_\alpha$

$$|S_N(c) - f_N(c)| \leq k \max_{a \in B_\alpha} \int |Y| \cdot I\left(2\alpha - |m(X,a)|\right) dF$$

So for $N > N_\alpha$

$$\begin{aligned} \sup_{c \in B} |S_N(c) - f_N(c)| &\leq k \max_{a \in B_\alpha} \int |Y| \cdot I\left(2\alpha - |m(X,a)|\right) dF \\ &\leq k \sup_{a \in B} \int |Y| \cdot I\left(2\alpha - |m(X,a)|\right) dF \end{aligned}$$

Since the above holds for all α, for N large enough:

$$\sup_{c \in B} |S_N(c) - f_N(c)| \leq k \lim_{\alpha \to 0} \sup_{a \in B} \int |Y| \cdot I\left(2\alpha - |m(X,a)|\right) dF$$

which by the Boundary Condition implies

$$\Pr\left(\lim_{N \to \infty} \sup_{c \in B} |S_N(c) - f_N(c)| \leq 0\right) = 1$$

Using also the fact that by Manski's (1988a) Lemma 6, p.106:

$$\Pr\left(\lim_{N \to \infty} \sup_{c \in B} |f(c) - f_N(c)| \leq 0\right) = 1$$

it follows that

$$\Pr\left(\lim_{N \to \infty} \sup_{c \in B} |S_N(c) - f(c)| \leq 0\right) = 1$$

By Lemma 5 we obtain that

$$\Pr\left(\lim_{N \to \infty} |b_N^s - b| = 0\right) = 1$$

Lemma 1 now ensures that the above is a necessary and sufficient condition for

$$I(x,b) = I_{\mu_Y}(x) \text{ a.e.}$$

■

4.3 Further estimator properties

It would be very convenient to have analytical results characterising the rates of convergence and asymptotic distribution of the estimators we have introduced as this would, for example, allow us to make comparisons between them. As we have been unable to obtain analytical results in this direction, we now report the results of a simulation study of the properties of the step function M-estimator and discuss likely properties of the other estimators. We use the step function M-estimator as a benchmark not only because it is simpler, but also because it is interesting even when Assumption 4.1 is violated (and is therefore our chosen estimator in previous Chapters).

4.3.1 Step function M-estimator

Kim & Pollard (1990) show that a broad class of estimators optimising step functions converge at cube-root rate to an analytically intractable distribution. We have not been able to show that our step function M-estimator belongs in this class but it seems reasonable to conjecture that it behaves in a similar way to the estimators of Kim & Pollard (1990)[4] given that its only unusual feature is that it optimises a step function - which according to Kim & Pollard (1990, p.194) "is the main distinguishing feature of estimation problems that exhibit cube-root asymptotics". The following simulation provides corroborating evidence for this conjecture.

SIMULATION 4.3.1

Consider the following DGP, used also in the simulations of Chapters 2 and 3:

$$\begin{aligned} R_{t+1} &= b_0 + b_1 R_t + U \\ U &\sim N\left(0, \sigma^2\right) \text{ iid} \\ R_0 &\sim N\left\{\frac{b_0}{1-b_1}, \sigma^2\left[1-(b_1)^2\right]^{-1}\right\} \\ b_0 &= 0.00015;\ b_1 = 0.0330;\ \sigma = 0.0108 \end{aligned}$$

The parameters of this DGP were determined using OLS to estimate an AR(1) model on a series of returns drawn from the empirical distribution of IBM daily closing prices[5] from 1st January 1990 through to 6th November 1997 (2049 observations).

[4] The structurally closest estimator to this estimator for which there exists some theory is the maximum score estimator of Manski (1975). Kim & Pollard (1990) show that the maximum score estimator converges at cube-root rate to the maximum of a Gaussian process.

[5] Obtained from DATASTREAM on the last day in the dataset.

Suppose the parametric form of the DGP is known, but that the values of the parameters are not. As discussed, our estimator can only provide a consistent estimate of the scale of the parameters of a linear model. Hence, the model to be estimated takes the form:[6]

$$m(X,c) = c + X.$$

It is worth noting that we now have a more parsimonious description of our model. An estimate for $\frac{b_0}{b_1}$ is b_N satisfying:[7]

$$b_N \in \arg\min_c - \int R_{t+1} \cdot I\left[c + R_t\right] dF_N \tag{4.9}$$

We took $T = 10^4$ draws from the c.d.f. F_{b_N} of b_N to obtain the simulated distribution $\widehat{F}_{b_N}$ and repeated for various sample sizes. In particular, we set: $N = \left\{100 \cdot 2^{l-1}\right\}_{l=1}^{7}$. The size of T was chosen so that the standard deviation $\widehat{\sigma}_{b_N}$ of $\widehat{F}_{b_N}$ did not change 'much' between $\frac{T}{2}$ and T for any N. The distributions $\widehat{F}_{b_N}$ were then used to investigate various properties of F_{b_N} that are of interest.

Asymptotic consistency

According to Proposition 4.5, b_N is asymptotically consistent so it must converge to $\frac{0.00015}{0.0330} = 4.5455 \times 10^{-3}$ almost surely as N becomes large. This is confirmed in Table 4.1, which illustrates the convergence of $\widehat{F}_{b_N}$ to b.

N	100	200	400	800	1600	3200	6400	∞
$10^3 \cdot \widehat{\mu}_{b_N}$	1.2797	1.3502	2.4048	2.9877	3.9049	4.099	4.4858	4.5455
$10^2 \cdot \widehat{\sigma}_{b_N}$	1.3603	1.3378	1.2694	1.1699	1.0162	0.86258	0.69860	-

Table 4.1 Each column corresponds to a sample size N for which we compute the mean $\widehat{\mu}_{b_N}$ and standard deviation $\widehat{\sigma}_{b_N}$ of the simulated distribution of b_N. The column corresponding to ∞ gives the true value b.

Finite sample bias

Table 4.1 also indicates that b_N is biased downwards. This means that forecasts tend to lead to long positions more often than they should. This occurs because the cost of being erroneously long is smaller (on average) than the cost of being erroneously short for this DGP (since $E(R) > 0$) and the asymmetric nature of the risk neutral forecasters' loss function implied by this is reflected in the direction of the bias.

[6] Strictly speaking, we should also estimate the model $m(X,c) = c - R$ unless we know the sign of b_1, but (for simplicity) we will asume here (and whenever we estimate linear models) that this is indeed known. Often either theory or related empirical evidence suggest this sign.

[7] To fix b_N at a particular point, we took (both here and later) the value closest to zero satisfying (4.9).

Rate of convergence

In our preceding discussion we conjectured that b_N will converge to a complicated asymptotic distribution D at cube-root rate, i.e. that:

$$N^{\frac{1}{3}}\left(b_N - \frac{b_0}{b_1}\right) \overset{a}{\sim} D \tag{4.10}$$

If our conjecture is true and σ_D, the standard deviation of D is finite, an implication is that:

$$\lim_{N\to\infty} N^{\frac{1}{3}}\sigma_{b_N} = \sigma_D \tag{4.11}$$

where σ_{b_N} is the standard deviation of F_{b_N}. While far from being a proof, verification of (4.11) would provide strong supportive evidence for our conjecture.

Figure 4.2. is a plot of $(N_k)^{\frac{1}{3}}\widehat{\sigma}_{b_N}$ for each N_k which supports (4.11) and provides an estimate for σ_D, ($\widehat{\sigma}_D \simeq 0.13$). In this figure σ_D will be the asymptote to which $N^{\frac{1}{3}}\widehat{\sigma}_{b_N}$ tends. If convergence was not at cube-root rate, this asymptote would not exist.

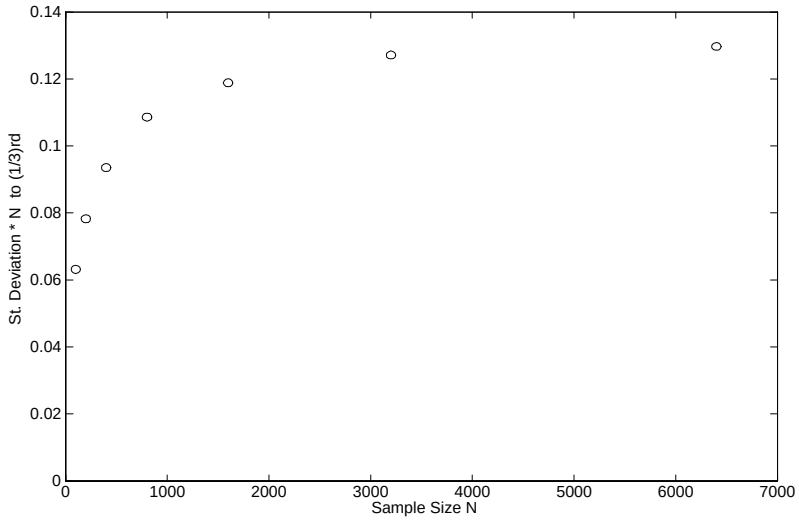

Figure 4.2. This figure provides supportive evidence for the conjecture that convergence occurs at cube-root rate.

This simple simulation provides supportive evidence for the conjecture that the estimator converges at cube-root rate. However, our simple simulation does not even begin providing a characterisation of the class of DGPs for which this will be the case - an issue we must leave for future research.

Asymptotic distribution

That D is a non-normal distribution is clearly illustrated by the following QQplot of a normal distribution and the distribution of $N^{\frac{1}{3}}\widehat{F}_{b_N}$ for $N = 6400$.

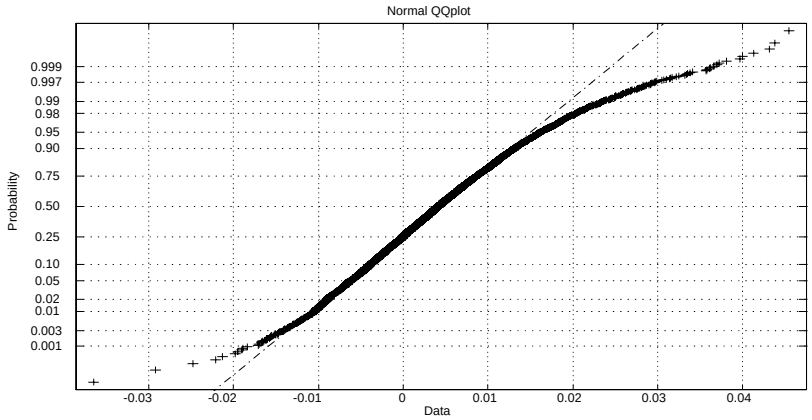

Figure 4.3. This figure plots the quantiles of D_N. The axes are set so that quantiles of any normal distribution lie on a staight line.

It may be feasible to obtain a more precise characterisation of the asymptotic distribution of b_N using the results of Kim & Pollard (1990). However, they do not provide results which would allow estimation of this asymptotic distribution and hence this exercise is of limited value since it cannot provide the basis for asymptotic inference.

4.3.2 Quasi generalised step function M-estimator

In addition to the Assumptions required for consistency of the step function M-estimator, consistency of this estimator requires knowledge of a model for $A(x)$ satisfying the requisite almost sure uniform convergence condition of Proposition 2. We conjecture that when this condition is indeed satisfied, this estimator will be *more efficient* than the simple step function M-estimator. This is because we would expect use of any additional available information regarding the structure of F to improve our estimators and A captures some of this structure. Note that the structure captured in A is still weaker than that of μ_Y. To illustrate, suppose for simplicity that $A_N(x) = A(x)$; then for reasonable distributions of U, I_Y takes the same values as I_{μ_Y} with positive (possibly even large) probability. For all observations of (y_n, x_n) such that this is the case, the realised model error u_n does not affect the values of the estimator which will therefore be less volatile. Hence we would expect that as A_N converges, the generalised binary quantile regression estimator is less volatile than the simple step function M-estimator.

No further properties of this estimator can be derived analytically, as the intractabilities of the simple moment extremum estimator are compounded.

4.3.3 Smoothed estimator

This estimator has been designed to modify the simple moment extremum estimator analogously to the modification Horowitz's (1992) smoothed maximum score estimator provides to the maximum score estimator of Manski (1975). Based on this fact, we conjecture that if we make certain additional assumptions analogous to the conditions of Horowitz (1992), this estimator will be asymptotically normal and convergence will take place at a rate which is faster than cube-root. Proof of this conjecture is a particularly desirable goal for future research, since it would allow us to conduct asymptotic inference about the behaviour of I_{μ_Y}.

4.3.4 Assessment of the estimators

The proposed estimators are attractive when the objective is to estimate a model the conditional mean of which has the same sign as the 'true' regression mapping . We have shown that the proposed estimators are robust at achieving this in the sense that they lead to models with the desired property even under relatively severe mis-specification. In particular, standard estimators such as least squares, (pseudo-) maximum likelihood and Bayesian estimators are only likely to perform well when $m(*,*)$ is correctly specified for the conditional mean, i.e. when Assumption 1 holds for $w(x) = 1$. Confidence in this stronger version of Assumption 1 requires a much greater degree of prior information about the behaviour of μ_Y that may not always be available. Hence the estimators previously introduced will in general be much more robust than 'standard' estimators.

In any case, even in the restrictive cases where we believe that we have an accurate model for the conditional mean, standard estimators are not necessarily the best way forward. This is because the sign of the estimated model's conditional mean may be badly estimated, even when the conditional mean is estimated well. Indeed, it is quite a general observation that functions of estimates are badly behaved (see for example the simulations in Chapters 2 and 3); a variety of estimators have been developed to get around this problem, mainly by taking parameter uncertainty into account when providing an estimate of the desired function (see e.g. Zellner (1978), Park & Zellner (1979)). Because our estimator directly estimates the sign of the conditional mean, it does not necessitate accounting for parameter uncertainty, making it more convenient.

Comparing the three estimators we propose, we suggest the step function M-estimator as the benchmark because (i) it is much simpler than both other estimators; (ii) it is useful even when a good model for $A(x)$ cannot be postulated; and (iii) it has some desirable properties in the risk neutral forecasting context even when Assumption 4.1 is not satisfied. Of course if a good model for $A(x)$ is available there is a good case for using the quasi-

generalised modification. However, as we show in the next Chapter, these estimators are difficult to compute and while a specially designed algorithm is proposed, it is a definite advatage of the smoothed modification that it is easier to compue; this advantage is reinforced by the fact that it may be possible to derive conditions under which it converges at a faster rate to a tractable asymptotic distribution.

4.4 Applications

We now turn to examples of situations where the sign of a regression mapping is the object of interest. The current practice in most of these situations is to simply derive estimates for this sign from estimates of the regression mapping itself. However, for reasons we have just discussed, these estimates may be quite poor.

In the first two applications we illustrate situations where it is desirable to estimate the sign of a mapping μ_{Yj} where $\mu_{Yj}(x) \equiv \mu_Y(x) - a_j$ for some regression mapping μ_Y and a known constant a_j. The third and fourth illustrate situations where it is desirable to estimate the *zeros* of μ_{Yj} - which are simply obtained by estimating the sign of μ_{Yj} and $-\mu_{Yj}$ and finding the x's such that neither are positive. Note that for these applications we require that the postulated model $m(*, c)$ is correctly specified for the sign of $\mu_Y + a_j$. But even if we impose the unnecessarily strong requirement that this holds *for all* a_j, this remains a substantially weaker assumption than that of correct specification of $m(*, c)$ for μ_Y.

4.4.1 Estimating solutions to discrete decision problems

Brandt (1999) proposes a nonparametric method for estimating solutions to decision problems which can be described in terms of Euler equations. However, discrete decision problems have solutions which are not solutions of Euler equations and hence Brandt's method is inapplicable. We show that some decision problems of this form are solved by known functions of the sign of a regression mapping and hence that the methods introduced in this paper can be used to estimate solutions to such discrete decision problems.

A broad class of binary choice dynamic programming problems of agents at time τ can be written in the general form:

$$\max_{d^t \in \{0,1\}} E\left[\sum_{t=\tau}^{T} \beta^{t-\tau}\left[\left(R_1^t - R_0^t\right) d^t + R_0^t\right] | x_\tau\right] \tag{4.12}$$

where T is possibly infinite, $0 < \beta < 1$ is a discount rate, $E(*|X_\tau)$ is the conditional expectation of a random variable with respect to x_τ (information at time τ) and R_j^t is the (random at $t-1$) reward to action j obtained by an individual choosing action j at time t.

Many such problems have the property that $\left\{d^t = 1\right\}$ is an optimal action at time t if and only if:

$$E\left(Y_{t+1}|X_t\right) \geq a_t \tag{4.13}$$

where Y_{t+1} is the realisation of the agent's value function if $d^t = 1$ is chosen at time t. In particular, this is the form of solutions to optimal stopping problems - that is problems in which if the optimal decision is $d_t = 1$ (stop) at t, then $d_\tau = 1$ is also optimal for $\tau > t$ (see e.g. Eckstein & Wolpin (1989) for a review of the substantial literature on problems of this form).[8]

Suppose now that $E\left(Y_{t+1}|X_t\right)$ is unknown to an agent solving this problem who must estimate his optimal decision using a sample of past data $\{y_n, x_n\}_{n=1}^N$. Then as is evident from (4.13) it suffices that he estimates the sign of the regression mappings $\mu'_{Y_{t+1}}(x_t) = E\left(Y_{t+1}|x_t\right) - a_t$ for each t.

Example 4.1 ***A canonical two period job-search problem***

Let $R_0^t = b$ be the reward from remaining unemployed at time $t = 1, 2$ and $R_1^1 = Y_1$ be the reward from being employed in the first period at a known wage Y_1. If a job is accepted in the first period, no new job offers are received in the second (employment leaves no time to search for new jobs) but the worker can keep the job he has. However, if a job is not accepted, a new wage offer Y_2 is received. The reward from being employed in the second period is then:

$$R_1^2 = \left\{ \begin{array}{l} Y_1 \text{ if } d^1 = 1 \\ Y_2 \text{ if } d^1 = 0 \end{array} \right\}$$

where $d^t = 1$ indicates a job offer is accepted in period t and $d^t = 0$ indicates it is not. Note that at $t = 1$, Y_2 is a random variable.

The agent's optimisation problem at $t = 1$ is a special case of (4.12) which may be more simply written as:

$$\max_{d^1, d^2 \in \{0,1\}} d^1\left(Y_1 - b\right) + b + \beta\left[d^2\left(d^1 \cdot \left(Y_1 - E\left(Y_2|x\right)\right) + E\left(Y_2|x\right) - b\right) + b\right]$$

where β is his discount factor and x is a vector containing public information at $t = 1$ predicting the distribution of wages Y_2 (such as unemployment rates).

The optimal decision rule (trivially obtained by backward induction) is:

$$\begin{aligned} d^2 &= 1 \text{ if } Y_1 > b \text{ and } d^1 = 1 \\ d^2 &= 1 \text{ if } Y_2 > b \text{ and } d^1 = 0 \\ d^1 &= 1 \text{ iff } E\left(Y_2|x\right) < (1+\beta)\, Y_1 - b. \end{aligned}$$

[8] A very different dynamic programming problem with a solution of the same form is the 'cost-loss ratio' problem extensively analysed by meteorologists - see for example Katz & Murphy (1990).

which can be solved without knowledge of $E(Y_2|X_1)$ *if the* ***sign*** *of* $\mu_Y(x) = E(Y_2|x) - (1+\beta)Y_1 + b$ *is known. Our estimators can therefore be applied to the estimation of this agent's optimal decision if a sample* $\{y_n, x_n\}_{n=1}^N$ *of draws from the distribution of* (Y, X) *are available. For example, the agent might need to estimate the sign of* $\mu_Y(x)$ *when* x *is a set of variables describing the state of the business cycle.* [9]

4.4.2 Estimating solutions to continuous decision problems with discrete solutions

When decision spaces are continuous, optimal actions will not be characterised by Euler equations when they are corner solutions. The following example a special case of which plays a central role in Chapter 2, illustrates the fact that estimation of the sign of a regression mapping will sometimes suffice for optimal decision-making in such contexts.

Example 4.2 ***Risk Neutral bond-portfolio investment decision with transaction costs***

Consider a Risk Neutral Investor's single-period problem of allocating wealth between a 'riskless' asset such as a bond and a single risky asset (or portfolio). This can be described as:

$$\max_{a_t \in [-s,l]} E\left\{W_t\left[(1+R_{t+1})\,a_t + \left(1+R_{t+1}^f\right)(1-a_t)\right]|x_t\right\} - \kappa W_t\left|a_t - a_t'\right| \tag{4.14}$$

where R_{t+1} *are returns to the portfolio from* t *to* $t+1$ *that are predictable with* x_t, R_{t+1}^f *are the returns to the bond,* W_t *is wealth at time* t, κ *are proportional transaction costs,* a_t' *is the proportion of* W_t *held in the asset before a decision is made at* t *and* a_t *is the proportion of wealth chosen for investment in the risky asset. The investor can borrow at the riskless interest rate up to* $(l-1)*100\%$ *of his wealth (*$l \geq 1$*) and take a short position in the asset of a size up to* $s*100\%$ *of his wealth (*$s \geq 0$*).*[10]

A solution to (4.14) also solves:

$$\max_{a_t \in [-s,l]} E\left(R_{t+1}^e|X_t\right)a_t - \kappa\left|a_t - a_t'\right| \tag{4.15}$$

where $R_{t+1}^e \equiv R_{t+1} - R_{t+1}^f$ *are 'excess returns'.*

[9] Granger & Pesaran (2000) give an example of a single-period decision problem that could be solved by estimating the sign of a regression mapping.

[10] The 'Risk Neutral Forecasting' decision problem of Chapter 2 is a special case of this with $\gamma = 0$ and $R^f = 0$.

Let $\mu_{Y,1}(x) \equiv E\left(R^e_{t+1}|X_t = x\right) - \kappa$ *and* $\mu_{Y,2}(x) \equiv E\left(R^e_{t+1}|X_t = x\right) + \kappa$. *A solution to (4.15) can be expressed as:*

$$a_t(x) = \left\{ \begin{array}{c} -s \text{ if } 0 \geq \mu_{Y,2}(x) \\ a'_t \text{ if } \mu_{Y,2}(x) > 0 \geq \mu_{Y,1}(x) \\ l \text{ if } \mu_{Y,1}(x) > 0 \end{array} \right\}. \tag{4.16}$$

So we can estimate optimal decisions a_t *by estimating the* ***sign*** *of the regression mappings* $\mu_{Y,1}(*)$ *and* $\mu_{Y,2}(*)$.

4.4.3 Estimation of the zeros of a regression function

The set of $\mathcal{Z}$eros of μ_Y is:

$$\mathcal{Z} \equiv \{x : \mu_Y(x) = 0, x \in \mathcal{X}\} \tag{4.17}$$

which is identical to

$$\left\{x : I_{\mu_Y}(x) = I\left(-\mu_Y(x)\right), x \in \mathcal{X}\right\}. \tag{4.18}$$

Evidently, our estimators can be used to estimate the set $\mathcal{Z}$ by estimating the sign of the regression mappings μ_Y and $-\mu_Y$ (note that if Assumption 4.1 applies to (m, μ_Y) then it will also apply to $(m, -\mu_Y)$) and obtaining

$$\widehat{\mathcal{Z}} \equiv \left\{x : I(x, b_N) = I'\left(x, b'_N\right), x \in \mathcal{X}\right\} \tag{4.19}$$

where $I'(x, c)$ is a model for $I\left(-\mu_Y(x)\right)$ with associated estimated parameter b'_N and $I(x, c)$ is a model for $I\left(\mu_Y(x)\right)$ with associated estimated parameter b_N. We now discuss various interesting applications in which $\widehat{\mathcal{Z}}$ may be useful. As we have noted in the introduction, an alternative computationally convenient non-parametric approach for estimating $\mathcal{Z}$ can be found in Haerdle & Nixdorf (1987) and Tsybakov (1988).

Calibration

Suppose X in (4.1) is deterministic, μ_Y is some unknown function and $\{y_j\}_{j=1}^J$ are $J \geq 1$ observations of Y drawn from $F_Y|X = x^*$ (x^* unknown). The calibration problem is the task of estimating x^* given $\{y_j\}_{j=1}^J$ and a sample $\{y_n, x_n\}_{n=1}^N$ of draws from F. The 'classical' approach to calibration (Eisenhart 1939) draws on the fact that the set:

$$\mathcal{Z}^1 = \left\{x : \mu_Y(x) - \frac{\Sigma_{j=1}^J y_j}{J} = 0, x \in \mathcal{X}\right\} \tag{4.20}$$

should provide a good estimate for x^* under weak conditions on $F_Y|X = x^*$. Since μ_Y is not known, however, the 'classical' approach relies on assuming

that $\mu_Y(x)$ can be substituted with a 'good' estimated model $m(x, \widehat{c})$ in this expression, so that

$$\widehat{\mathcal{Z}}^1 = \left\{ x : m(x, \widehat{c}) - \frac{\Sigma_{j=1}^J y_j}{J} = 0 \right\}$$

is a 'good' estimate for x^*.

However, for reasons closely related to the point made in the Simulation of Chapter 3, this does not seem to be the case (Kruthckoff 1967) and this approach is now considered questionable. Consequently, a number of new estimators have been proposed (Brown 1979, Hunter & Lamboy 1981, Kalotay 1971, Lwin & Maritz 1980), each with its own advantages. The estimators developed in this paper can be used to directly estimate $\mathcal{Z}^1$ which itself should be a good estimate for x^*. Hence, our estimators provide an interesting addition to the pool of available estimators for this problem which may be particularly attractive when no model is certain to be a correct specification for μ_Y.

Econometric equation inversion

Let X_i denote the i'th element of X and X_{-i} denote $(X_1, ..., X_{i-1}, X_{i+1}, ..., X_I)$. It is sometimes desirable to estimate the mapping $x_i = z(y^*, x_{-i})$ such that $\mu_Y(X_{-i} = x_{-i}, X_i = x_i) = y^*$ under the assumption that the relationship between Y and X is characterised by (4.1). This is sometimes referred to as a problem of econometric equation inversion, since the inverse of $\mu_Y(x_i, x_{-i})$ with respect to x_i needs to be estimated. This can be treated as a problem of estimating the zeros of a mean regression model, as the following two examples illustrate.

Decisions with Targets Consider the following classic problem in the control of passively observed systems which arises routinely in Economic models. A decision maker has some control over the distribution of X_i which he uses to ensure that the expected value of Y is equal to a 'target level' y^*, given any realisation of X_{-i}. It is typically assumed that $\mu_Y(x)$ is not affected by any changes in the distribution of X_i which the decision maker effects. In this case, if μ_Y is unknown, the decision maker will be interested in estimating:

$$\mathcal{Z}^2 \equiv \{ x_i : \mu_Y(x_i, x_{-i}) - y^* = 0, x \in \mathcal{X} \}.$$

which is a set estimable using the estimators we have proposed.

Macroeconometric models Substantial research effort has been devoted to analysing the implications of econometric equation inversion in the context of macroeconomic models - see for example Hendry & Ericsson (1991),

Ericsson et al. (1998) and the references therein. The standard setting in which this arises is when (4.1) is assumed to hold for money demand as a function of prices, interest rates etc. Based on these assumptions, it is often required to obtain an expression for prices (or interest rates) as a function of money demand and interest rates (or prices).

Two common ways of estimating such a relationship which are now recognised as problematic are:

1. Invert $m(x, c)$ with respect to x_i and estimate the model $x_i = m'(y, x_{-i}, c)$ as if it were a correctly specified model for $\mu_{X_i}(y, x_i)$. Of course, this will only be the case under very special assumptions such as joint normality of (Y, X), which are probably a very bad approximation to any interesting reality. Estimation following this course may therefore lead to inconsistent estimates.

2. Invert an estimated model $m(x, b_N)$ for $\mu_Y(x)$ with respect to x_i to obtain an expression of the form $x_i = m'(y, x_{-i}, b_N)$. This will generally deliver bad estimates for $\mu_{X_i}(y, x_{-i})$ except under special conditions, even though it is sometimes used for this purpose. It is more natural to expect it would serve as a good estimate for

$$\mathcal{Z}^3 \equiv \{x_i : \mu_Y(x_i, x_{-i}) - y^* = 0, x \in \mathcal{X}\}$$

although for reasons related to those discussed in relation to Calibration, this should also be questioned, particularly if adjustments are not made to account for parameter uncertainty.

Evidently, the only relationship that can be estimated in a satisfactory way is $\mathcal{Z}(x_{-i}, y^*)$, which can also be achieved using our estimator. An example of $\mathcal{Z}(x_{-i}, y^*)$ could give the prices x_i conditional on which, for a given level of income and interest rates (x_{-i}) money demand Y is expected to equal y^*. The advantages are, as always, parsimony and consistency under weaker conditions on correct specification.

4.5 Conclusions & future directions

In this chapter we develop estimators for a semiparametric model of a regression mapping when our interest focuses on the sign of this mapping. We have given a number of examples of important modelling situations in which this is the case and have argued that in these situations the proposed estimators are particularly advantageous.

The reasons for this are threefold: Firstly, it will be easier to correctly specify a model for the sign of the regression mapping and therefore obtain a consistent estimator for it. Secondly, as illustrated in the context of our simulation, models for the sign are likely to be more parsimonious than models

for the regression mapping; and thirdly, estimators for the sign derived from estimators of the regression mapping are likely to require some modification to account for parameter uncertainty. While a convincing assessment of the value of these methods is pending their use in an appropriate empirical application, these arguments indicate that they might be useful in a broad range of such applications.

The three semiparametric estimators we propose are consistent under weak conditions, including for example arbitrary heteroskedasticity of errors and a non-functional relationship between Y and X. While the properties of the benchmark step function M-estimator and the quasi-generalisation thereof seem to be hopelessly intractable, the properties of the smoothed estimator may be tractable and we intend to investigate this in future research. This is particularly interesting, since it could provide a route to conducting asymptotic inference about the sign of a regression mapping and therefore also about the form of those optimal decisions that can be expressed as functions of this sign. Of course, it may be useful to also consider how other stages of the modelling process might be modified to improve modelling of the sign of regression mappings. Inevitably, this will require accumulation of practical experience with this particular problem.

4.6 Appendix

4.6.1 A. Lemmata used in proofs of propositions in text

Lemma 4.1 *Let $h : \mathcal{X} \to \mathbb{R}_{++}$ be any strictly positive function.*
Suppose Assumption 4.1 is satisfied.
$I(x,b) = I_{\mu_Y}(x)$ almost everywhere if and only if:

$$b \in \arg\min_{c \in B} - \int h(X) \cdot Y \cdot I(X,c) dF. \tag{4.21}$$

Proof. $\Rightarrow$

$$\begin{aligned} - \int h(X) \cdot Y \cdot I(X,c) dF &= - \int h(X) \cdot \mu_Y(X) \cdot I(X,c)) dF \\ &\geq - \int h(X) \cdot \mu_Y(X) \cdot I(X,c) \cdot I_{\mu_Y}(X) dF \\ &\geq - \int h(X) \cdot \mu_Y(X) \cdot I_{\mu_Y}(X) dF. \end{aligned}$$

By A4.1 there is a b such that $I(x,b) = I_{\mu_Y}(x)$. Clearly in this case, the above relations hold with equality. Therefore if c is a minimising value, the above relations must also hold with equality and hence:

$$\int h(X) \cdot \mu_Y(X) \cdot I(X,c) dF = \int h(X) \cdot \mu_Y(X) \cdot I_{\mu_Y}(X) dF.$$

This implies that $I(X,c) = I_{\mu_Y}(X)$ almost surely.

$\Leftarrow$

We have already shown that:

$$-\int h(X)\cdot Y\cdot I(X,c)dF \geq -\int h(X)\cdot \mu_Y(X)\cdot I_{\mu_Y}(X)dF$$

Hence if $I(X,b) = I_{\mu_Y}(X)$, b is a minimising value of the LHS. ■

Lemma 4.2 *Let $A:\mathcal{X}\to[0,1]$ be a mapping such that:*

$$A(x) \equiv \frac{E(|Y||I(Y)=1,x)}{E(|Y||I(Y)=1,x)+E(|Y||I(Y)=0,x)},\ x\in\mathcal{X} \tag{4.22}$$

Suppose Assumption 4.1 is satisfied.

$I(x,b) = I_{\mu_Y}(x)$ almost everywhere if and only if:

$$b\in \arg\min_{c\in B} -\int [A(X)-\Pr(I(Y)=0|X)]\cdot I(X,c)dF \tag{4.23}$$

Proof. Notice that for any function $h:\mathcal{X}\to\mathbb{R}$:

$$\int h(X)\cdot Y\cdot I(X,c)dF = \int h(X)\cdot \mu_Y(X)\cdot I(X,c)dF \tag{4.24}$$

where F_X is the c.d.f. of X.

It is easy to show (see Chapter 2, Proposition 2.2) that:

$$\mu_Y(x) = E(|Y||I(Y)=1,x) - \Pr(I(Y)=0|x)\left[E(|Y||I(Y)=0,x)+E(|Y||I(Y)=1,x)\right] \tag{4.25}$$

Define the strictly positive function $h:\mathcal{X}\to\mathbb{R}_{++}$ by:

$$h(x) \equiv \frac{1}{E(|Y||I(Y)=0,x)+E(|Y||I(Y)=1,x)} \tag{4.26}$$

Using (4.24),(4.25) and (4.26).

$$\int h(X)\cdot Y\cdot I(X,c)dF = \int [A(X)-\Pr(I(Y)=0|X)]\cdot I(X,c)dF$$

Lemma 1 can now be applied to obtain the desired result. ■

Lemma 4.3 *Let*

$$\begin{aligned} h(x,c) &\equiv (I_Y - A(x))\cdot I(x,c),\\ h_N(x,c) &\equiv (I_Y - A_N(x))\cdot I(x,c),\\ v(x) &\equiv I_Y - A(x),\\ v_N(x) &\equiv I_Y - A_N(x). \end{aligned}$$

If $A_N(x)$ converges uniformly to $A(x)$ almost surely, the SLLN holds, and B_α is a finite cardinality subset of B then

$$\Pr\left(\lim_{N\to\infty}\max_{c\in B_\alpha}\left|\int h_N(X,c)dF_N - \int h(X,c)dF\right| = 0\right) = 1 \tag{4.27}$$

and

$$\Pr\left(\lim_{N\to\infty}\max_{c\in B_\alpha}\left|\int v_N(X)dF_N - \int v(X)dF\right| = 0\right) = 1. \tag{4.28}$$

Proof. As the proofs of (4.27) and (4.28) are identical, we prove only (4.27).

Since $h_N(X,c)$ is a linear function of $A_N(X)$, the assumption directly implies that for all $\varepsilon > 0$ there exists a N_ε such that for $N > N_\varepsilon$, uniformly on $(\mathcal{X}, B)$:

$$|h_N(X,c) - h(X,c)| < \varepsilon \text{ a.s.}$$

So:

$$\int |h_N(X,c) - h(X,c)|\, dF_N < \int \varepsilon dF_N = \varepsilon$$

Since $\int |h_N(X,c) - h(X,c)|\, dF_N \geq \left|\int h_N(X,c) - h(X,c)dF_N\right|$ this implies:

$$\left|\int h_N(X,c) - h(X,c)dF_N\right| < \varepsilon$$

$$\Rightarrow \Pr\left(\lim_{N\to\infty}\left|\int h_N(X,c) - h(X,c)dF_N\right| = 0\right) = 1 \tag{4.29}$$

The Strong Law of Large Numbers ensures that for any c:

$$\lim_{N\to\infty}\left|\int h(X,c)dF_N - \int h(X,c)dF\right| = 0 \text{ a.s.} \tag{4.30}$$

So combining (4.29) and (4.30), for any c:

$$\lim_{N\to\infty}\left|\int h_N(X,c)dF_N - \int h(X,c)dF_N\right| + \left|\int h(X,c)dF_N - \int h(X,c)dF\right| = 0 \text{ a.s.} \tag{4.31}$$

Now notice that since $|a-b| + |b-c| \geq |a-c|$,

$$\left|\int h_N(X,c)dF_N - \int h(X,c)dF\right| \leq \left|\int h_N(X,c)dF_N - \int h(X,c)dF_N\right| + \left|\int h(X,c)dF_N - \int h(X,c)dF\right|$$

Using this fact and (4.31), we obtain that for any $c \in B_\alpha$:

$$\left| \int h_N(X,c) dF_N - \int h(X,c) dF \right| \to 0 \text{ a.s.}$$

■

Lemma 4.4 *Let*

$$\begin{aligned} h(x,c) &\equiv v(x) \cdot I(x,c), \\ h_N(x,c) &\equiv v_N(x) \cdot I(x,c), \\ v(x) &\equiv I_Y - A(x), \\ v_N(x) &\equiv I_Y - A_N(x). \end{aligned}$$

If $A_N(x)$ converges uniformly to $A(x)$ almost surely and A2-6 are satisfied, then

$$\Pr\left(\lim_{N\to\infty} \max_{c\in B} \left| \int h_N(X,c) dF_N - \int h(X,c) dF \right| = 0 \right) = 1.$$

Proof. We follow the logic of Manski (1988*a*), Lemmata 5 and 6, pp 104-108:

$$\begin{aligned} &\left| \int h_N(X,a) - h_N(X,c) dF_N \right| \\ &\qquad = \left| \int v_N(X)[I(X,a) - I(X,c)] dF_N \right| \\ &\qquad \leq \int |v_N(X)[I(X,a) - I(X,c)]| \, dF_N \end{aligned}$$

So

$$\left| \int h_N(X,a) - h_N(X,c) dF_N \right| \leq \int_{\mathcal{X}(a,c)} |v_N(X)| \, dF_N \tag{4.32}$$

where

$$\mathcal{X}(a,c) \equiv \{x \in \mathcal{X} : m(x,a) \leq 0 \leq m(x,c) \cup m(x,a) \geq 0 \geq m(x,c)\}.$$

For $\alpha > 0$ and $c \in B$, by the equicontinuity assumption (A5) which for notational simplicity (but without loss of generality) we assume holds for $w(x) = 1$, it follows that there exists a δ_α such that for all $(a,x) \in (B, \mathcal{X})$

$$|a - c| < \delta_\alpha \Rightarrow \left\{ \begin{array}{c} m(x,c) > \alpha \Rightarrow m(x,a) > 0 \\ m(x,c) < -\alpha \Rightarrow m(x,a) < 0 \end{array} \right\}$$

Hence there exists a δ_α such that for all $(a,x) \in (B, \mathcal{X})$

$$|a - c| < \delta_\alpha \Rightarrow \mathcal{X}(a,c) \subset \mathcal{X}_{c\alpha} \equiv \{x \in \mathcal{X} : -\alpha < m(x,c) < \alpha\}. \tag{4.33}$$

Using (4.32) and (4.33):

$$|a-c|<\delta_\alpha \Rightarrow \left|\int h_N(X,a)-h_N(X,c)dF_N\right| \leq \int_{Xc\alpha} |v_N(X)|\, dF_N$$

By identical reasoning, this condition holds if we replace $h_N(x,c)$ with $h(x,c)$, $v_N(x)$ with $v(x)$ and F_N with F.

Hence,

$$|a-c|<\delta_\alpha \Rightarrow \left|\int h(X,a)-h(X,c)dF\right| + \left|\int h_N(X,a)-h_N(X,c)dF_N\right| \leq \int_{\mathcal{X}_{c\alpha}} |v(X)|\, dF + \int_{\mathcal{X}_{c\alpha}} |v_N(X)|\, dF_N \quad (4.34)$$

Now notice that:

$$\left|\int h_N(X,a)dF_N - \int h(X,a)dF\right| = \left| \begin{array}{c} \int (h_N(X,a)-h_N(X,c))\, dF_N - \int (h(X,a)-h(X,c))\, dF + \\ + \int h_N(X,c)dF_N - \int h(X,c)dF \end{array} \right|$$

$$\Rightarrow \left|\int h_N(X,a)dF_N - \int h(X,a)dF\right| \leq \left|\int h_N(X,a)-h_N(X,c)dF^N\right| + \left|\int h(X,a)-h(X,c)dF\right| + \left|\int h_N(X,c)dF^N - \int h(X,c)dF\right|$$

Hence combining the above relationship with (4.34):

$$\begin{aligned} |a-c| &< \delta_\alpha \Rightarrow \\ &\left|\int h_N(X,a)dF_N - \int h(X,a)dF\right| \\ &\leq \int_{Xc\alpha} |v(X)|\, dF + \int_{Xc\alpha} |v_N(X)|\, dF_N + \left|\int h_N(X,c)dF_N - \int h(X,c)dF\right| \end{aligned}$$

Now Compactness of B implies that there exists a $B_\alpha \subset B$ such that $card(B_\alpha)<\infty$ and for any $a \in B$ there is a $c \in B_\alpha$ such that $|a-c|<\delta_\alpha$. Therefore for all $a \in B$,

$$\left|\int h_N(X,a)\, dF - \int h(X,a)\, dF\right| \leq \max_{c\in B_\alpha} \int_{\mathcal{X}_{c\alpha}} |v(X)|\, dF + \max_{c\in B_\alpha} \int_{\mathcal{X}_{c\alpha}} |v_N(X)|\, dF_N + \max_{c\in B_\alpha} \left|\int h_N(X,c)dF_N - \int h(X,c)dF\right|$$

By Lemma 1, for all α, $\eta > 0$, there exists a $N_{\alpha\eta} < \infty$ such that for all $N > N_{\alpha\eta}$

$$\begin{aligned}\sup_{a\in B}\left|\int h_N(X,a)dF_N - \int h(X,a)dF\right| &\leq 2[\max_{c\in B_\alpha}\int_{Xc\alpha}|v(X)|\,dF + \eta \\ &\leq 2[\sup_{c\in B}\int_{Xc\alpha}|v(X)|\,dF + \eta\end{aligned}$$

Now the Boundary condition can be written as:

$$\lim_{\alpha\to 0}\sup_{c\in B}\int_{c\alpha}|Y|\,dF = 0$$

implying that as $(\alpha,\eta)\to 0$ the required result is obtained. ■

Lemma 4.5 *Assume*

$$\int (A_N(x) - I_Y)\cdot I(x,c)dF_N$$

converges uniformly almost surely to

$$\int (A(x) - I_Y)\cdot I(x,c)dF$$

as N tends to infinity and A2. Then:

$$\Pr\left(\lim_{N\to\infty}\left|b_N^\alpha - b^j\right| = 0\right) = 1$$

Proof. See for example Amemiya (1985) Theorem 4.1.1 as modified by footnote 1. ■

4.6.2 B: Sufficient conditions for assumptions required for asymptotic consistency

Sufficient conditions for Identifiability

We give such conditions on the basis of the following Proposition which is a simple extension of a result in Manski (1985).

Proposition 4.4 *Sufficient conditions for Assumption 4 are:*

1. *For some $\tau\in\mathbb{T}$, there exists $w:\mathcal{X}\to\mathbb{R}$ such that:*

$$\tau(m(x,c)) = w(x)'c, \quad (x,c)\in(\mathcal{X},B),$$

and

2. *The support of* F_X *is not contained in any proper linear subspace of* $\mathbb{R}^k$, *and*

3. $b_l \neq 0$ *for some* l *and* $\forall X_{-l} \equiv (X_1, X_2, ..., X_{l-1}, X_{l+1}, ...X_k)$ *the distribution of* $X_l|X_{-l}$ *has everywhere positive Lebesgue density.*

Proof. The conditions we have assumed imply by Manski (1985, Lemma 2, p. 317) that for all $c \neq b$,

$$\int_{\mathcal{X}_c} dF_X > 0$$
$$\mathcal{X}_c \equiv \left\{ x \in \mathbb{R}^k : \mathbf{1}\left[w(x)'c \geq 0\right] \neq \mathbf{1}\left[w(x)'b \geq 0\right] \right\} \tag{4.35}$$

Let:

$$\mathcal{X}_c' \equiv \left\{ x \in \mathbb{R}^k : \mathbf{1}\left[w(x)'c > 0\right] \neq \mathbf{1}\left[w(x)'b > 0\right] \right\}$$

For $c \neq 0$, by the Proposition's conditions 2 and 3.

$$\int_{\mathcal{X}_c'} dF_X = \int_{\mathcal{X}_c} dF_X > 0.$$

For $c = 0$,

$$\int_{\mathcal{X}_c'} dF_X = 1 - \int_{\mathcal{X}_c} dF_X$$

Hence for all $c \neq b$ it is the case that $\int_{\mathcal{X}_c'} dF_X > 0$ which implies:

$$\int R \cdot \mathbf{1}\left[w\left(X\right)' c > 0\right] dF \neq \int R \cdot \mathbf{1}\left[w\left(X\right)' b > 0\right] dF$$

so by condition 1

$$\int R \cdot \mathbf{1}\left[m\left(X, c\right) > 0\right] dF \neq \int R \cdot \mathbf{1}\left[m\left(X, b\right) > 0\right] dF$$

and hence the minimum of the r.h.s. must be unique, ensuring identifiability. ■

Sufficient conditions for Equicontinuity

By Manski (1988*a*), Lemma 7, pp. 109-110:
For some $\tau \in \mathbb{T}$, *at least one* of (1), (2) or (3) hold:

1. $\mathcal{X} \times B$ is a compact metric space and $\tau\left(m\left(*, *\right)\right)$ is continuous on it.

2. $\tau(m(*,*))$ is bounded on $\mathcal{X} \times C$ and $m(x,*)$ is convex on C for all $x \in \mathcal{X}$, where $B \subset C \subset \mathbb{R}^k$ and C is an open convex set.

3. $\tau(m(x,c)) = w(x)'c$, $(x,c) \in (\mathcal{X}, B)$, $w : \mathcal{X} \to \mathbb{R}^k$.

Sufficient conditions for the Boundary condition

According to Manski (1988*a*, Lemma 8, pp. 110-111) the following three conditions must hold:

1. For some $\tau \in \mathbb{T}$, there exists a $w : \mathcal{X} \to Z \subset \mathbb{R}^k$ such that $\tau(m(x,c)) = w(x)'c$, $(x,c) \in (\mathcal{X}, B)$,

2. $\forall\, (c,\omega) \in B \times V$, where V is the range space of $|R|$, the probability measure $F_{w(x)'c}|R$ is absolutely continuous w.r.t. the Lebesgue measure μ and also $\exists\ \lambda < \infty$ s.t. $\frac{dF_{w(x)'c}|R}{d\mu} < \lambda$,

3. $\int |R|\, dF_R$ exists

Chapter 5

Global optimisation of the average of random step functions

SUMMARY

The average of a large number of random step functions is a discontinuous surface with an extremely large number of local optima even though, as the number of step functions tends to infinity, convergence may be to a smooth surface with a unique minimum. We propose an algorithm for computing the minimum of such a surface since standard gradient-based optimisation methods are inapplicable. The algorithm is based on the idea of substituting the discontinuous surface with a sequence of continuous surfaces that converge to it, yet are easy to minimise sequentially. Conditions under which the algorithm converges to a global minimum are provided and its performance is evaluated in a simple but relevant application.

5.1 Introduction

In this Chapter we study the significant difficulties involved in computationally minimising functions of the form

$$f_N(c) \equiv -\int Y \cdot I\left[m\left(X,c\right)\right] dF_N \tag{5.1}$$

where F_N is the empirical distribution function of (Y,X) compiled using N observations from F, the c.d.f. of (Y,X) and $m\left(X,c\right)$ is some real valued random function. I is an indicator function taking the value 1 if the expression in brackets is positive and zero otherwise. We are interested in obtaining the minimum b_N of $f_N(c)$ for $c \in B$.

Our interest in this problem clearly arises from the fact that the simple step function M-estimator of Chapter 4 used throughout this thesis minimises a function of this form. We will explain *why* $f_N(c)$ is difficult to minimise and - based on this explanation - propose an optimisation algorithm designed to perform well given the characteristics of f_N that we will establish. A theoretical convergence result is established and the algorithm is evaluated in the context of estimating a simple Risk Neutral Forecasting Model.

The problem we address may arise in computing a variety of econometric estimators, including the maximum score estimator of Manski (1975). Indeed, it has also been encountered by LeBaron (1998*a*) and Pictet et al. (1996).

5.2 Source of computational difficulties

The discontinuity of the integrand in (5.1) makes the computational derivation of b_N non-trivial. A direct implication of this discontinuity is that the range of the objective function must *necessarily* have *finite cardinality*. This is reassuring in as much as it implies that b_N always exists, but is also troublesome in that it implies its computation will involve the optimisation of a discontinuous function, which precludes the use of gradient descent methods. Considering that the randomness inherent in the sampling process is also carried over to the objective function, it becomes evident that $f_N(c)$ will be a highly rugged function with many local extrema even when $\int Y \cdot I\left[m\left(X,c\right)\right] dF$ is continuous.[1]

As N tends to infinity, our assumptions in Chapter 4 ensure that $f_N(c)$ converges uniformly almost surely to a continuous function and hence computational optimisation should become easier. For small N, minimisation of $f_N(c)$ is trivial since it takes on a small number of values in B. To illustrate this, think of the simple case where m is a linear function, i.e. $m\left(x,c\right) = x'c$, $x \in \mathcal{X}$, $c \in B$. Then the N hyperplanes defined by the collection of vectors $\{c : x_n'c = 0\}_{n=1}^{N}$ decompose B into at most $N^{\dim(\mathcal{X})} + 1$ regions in each of which $f_N(c)$ must be constant as a function of c.[2] Unfortunately, it turns out that the sample sizes we wish (or are able) to use in practice lie in the intermediate zone where optimisation of $f_N(c)$ is difficult and it is this case that we will consider. To get a feel for the properties of the objective functions we typically deal with, consider Figure 5.1 which is a plot of $\int R_{t+1} \cdot I\left[c_0 + 0.1R_t + c_2R_{t-1}\right] dF_N$ where F_N is the empirical distribution of the IBM series described in the simulations of previous Chapters.

[1] Assumptions A4.2-4.6 of Chapter 4 are sufficient to ensure this by Manski (1988*a*, Lemma 5, p.104).

[2] A similar point, but in a different context, is made by Manski (1985, p.320).

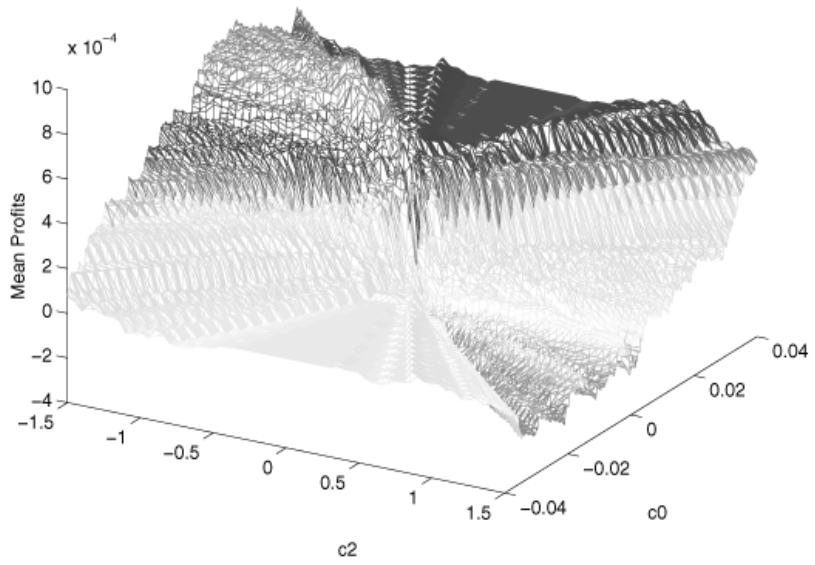

Figure 5.1. This objective function has a large number of local extrema and its discontinuities preclude the use of most sophisticated optimization procedures to find its global maximum.

We have tried a variety of optimization techniques (such as simplex search, gradient descent and a genetic algorithm) on this type of problem but they have all drastically failed to converge to a value for b_N which does not heavily depend on initial conditions. This problem has also been noted by LeBaron (1998*a*) and Pictet et al. (1996) for similar objective functions and, along with the lack of a theoretical characterisation of the statistical properties of b_N, has probably been the most important obstacle to the widespread use of the intuitive estimator proposed in Chapter 4.

5.3 Proposed algorithm

The computational procedure we propose is based on minimising f_N by combining a procedure for finding a point close to the global minimum with a *local* search. To achieve this, suppose there exists a sequence of functions $\left\{f_N^j\right\}_{j=0}^J$ s.t. *(i)* it is 'easy' to minimise f_N^1 globally, *(ii)* the minima of f_N^j and f_N^{j+1} are 'close' for all j and *(iii)* $f_N^J(c) = f_N(c)$, $c \in B$. If we can find such a sequence, we can minimize f_N^j globally using only a local search and the global minimum of f_N^{j-1} as a starting point. Hence, a global search is required only for f_N^0 and by construction this is 'easy'. The difficult problem of optimising f_N is thereby replaced with a sequence of $J+1$ easy

optimisation problems.[3]

A sequence of functions *conjectured* to have these properties is defined by

$$f_N^j(c) = -\int Y \cdot \left(1 + \exp\left(-\frac{m(X,c)}{\zeta^j}\right)\right)^{-1} dF_N, \ j = 0, 1, ..., J,$$

where $\{\zeta^j\}_{j=0}^J$ is an appropriate sequence of strictly decreasing positive constants. The function $f_N^j(c)$ is a smoothed version of $f_N(c)$ obtained by replacing the step function in (5.1) with a sigmoid function the steepness of which is controlled by ζ^j.[4] Using this sequence of functions, the algorithm for minimising $f_N(c)$ involves the following steps:

Step 1. Set ζ^0 as described in Appendix B to ensure that a large proportion of the values of $\frac{m(X,c)}{\zeta^0}$ lie in a region of the domain of $(1 + \exp(-y))^{-1}$ in which this function has some curvature. We illustrate the impact of this smoothing in Figure 5.2 which is a plot of $-f_N^0(c_0, c_2)$ with m and F_N as defined in Figure 5.1

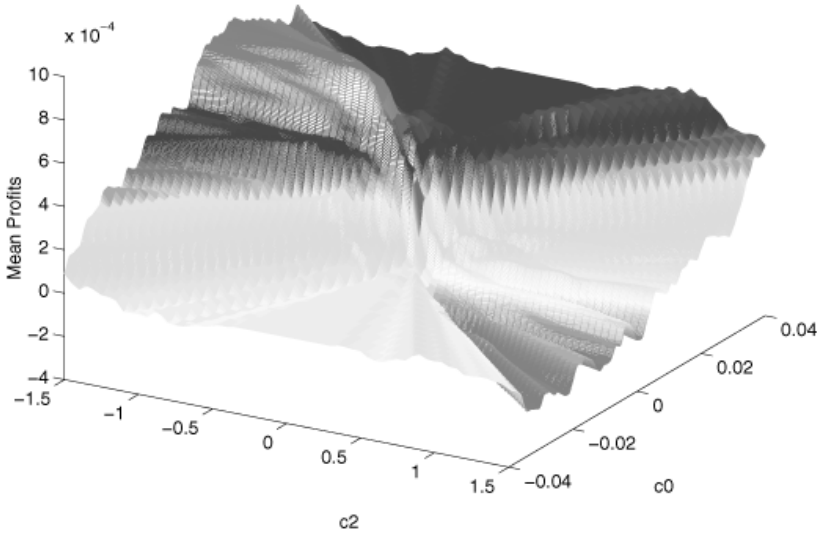

Figure 5.2. Although this objective function has more than one maximum it is sufficiently smooth for standard optimization procedures to be efficient in locating its global maximum.

To compute the minimum of f_N^0, optimisation methods that work well globally should be used, since f_N^0 has a plethora of local minima. In particular, we propose the use of a genetic algorithm to find an initial minimum

[3] Thanks are due to Domingo Tavella for a discussion which led to the development of this algorithm.

[4] This idea also underlies the smoothed estimator of Chapter 4.

(see Dorsey & Mayer (1995) for evidence on the suitability of such an algorithm) which will then be used as a starting point for a simplex search. The minimum obtained from this global search is our computed estimate for the minimum of f_N^0.

Let $j = 1$ and proceed to Step 2. Denote the computed estimate of the minimum of f_N^j as b_c^j.

Step 2. Using b_c^j as a starting point, compute an estimate for:

$$b_N^{j+1} \in \arg\min_{c \in B} - \int Y \cdot \left(1 + \exp\left(-\frac{m(X,c)}{\zeta^{j+1}}\right)\right)^{-1} dF_N$$

where $\zeta^{j+1} < \zeta^j$ and $\zeta^J = 0$.

Step 3. If $j = J$, end; otherwise, let $j = j + 1$ and return to Step 2.

5.4 Algorithm performance

An immediate objection to the use of the algorithm presented is that it will not be possible to verify whether elements of the sequence $\left\{b_N^j\right\}_{j=0}^J$ are sufficiently close to each other for a good approximation to b_N^j to be in the vicinity of b_N^{j+1} and therefore ensure the necessity of only a local search. Whether or not this happens will of course be an empirical issue and will depend on the interaction of the size of N, the form of F and m as well as the efficiency of the computational procedures used. We will provide sufficient conditions to rule out this problem and since these conditions are difficult to verify, we complement them with empirical evidence suggesting that the procedure performs well for various N in that it converges much faster than grid searches.

5.4.1 Analytical results

The following proposition states that if the minimum of the population analogue of the objective function is a continuous function of the smoothing parameter, the problem described can be ruled out when N is large.[5]

Proposition 5.1 *Define:*

$$\begin{aligned} f^j(c) &\equiv \int Y \cdot \left(1 + \exp\left(-\frac{m(X,c)}{\zeta^j}\right)\right)^{-1} dF, \\ f(c) &\equiv \int Y \cdot I\left[m(X,c)\right] dF, \end{aligned}$$

[5] Of course, as N becomes large the original objective function becomes increasingly smooth and therefore some of the computational difficulty disappears. However, it would be disconcerting if our algorithm did not work even when N was large.

where ζ^j is a constant.

Define also:

$$b^j \equiv \arg\min_{c \in B} -f_j(c),$$

And the sample analogues $f_N^j(c)$, $f_N(c)$ and b_N^j respectively.
Let b_c^j be a numerical estimate for b_N^j.
If:

1. *Assumptions A4.2-4.6 of Chapter 4 are satisfied.*
2. *$f(c,\zeta)$ is a quasi-concave function, such that there exists a $D_\zeta : \mathcal{X} \to [0,\infty)$ with $\int D_\zeta(X)\, dF_X < \infty$ and $\left| Y \cdot \left(1 + \exp\left(-\frac{m(x,c)}{\zeta}\right)\right)^{-1} \right| \leq D_\zeta(x)$, for all (x,c) in $\mathcal{X} \times B$.*
3. *The optimisation algorithms satisfy:*
   (a) *The optimum of f_N^0 can be computed exactly (i.e. $b_c^0 = b_N^0$),*
   (b) *For every $j > 0$, optimisation works in an ε-neighbourhood of the solution, i.e. for a starting point π and some positive constant ε^*,*

$$\left| \pi - b_N^j \right| \leq \varepsilon^* \Rightarrow b_c(j,N) = b_N^j.$$

   (c) *For every $j > 0$, the starting point π for optimisation of f_N^j is given by b_c^{j-1}.*

Then:

There exists a strictly decreasing finite sequence of positive constants $\{\zeta^j\}_{j=0}^J$, with $\zeta^J = 0$ such that

$$\Pr\left(\lim_{N \to \infty} \left| b_c^J - b_N \right| = 0 \right) = 1$$

Proof. See Appendix A. ■

While this proposition is reassuring, the quasiconcavity assumption is unsatisfactory because we cannot provide conditions on our primitives $\{F, m\}$ which would guarantee it holds. Furthermore, it is a property which is computationally costly to verify using Monte Carlo simulations even in simple cases.[6] We conclude that while the usefulness of the computational algorithm proposed can only be evaluated in the context of a specific application, it is intuitively sensible, will work under reasonable (if unverifiable) conditions and has been found to be very effective applied to a range of real data (illustrated in the next subsection).

[6] Nevertheless, we have confirmed that quasi-concavity is the case for the AR(1) model used in simulations in previous Chapters by plotting $f(c,m)$. Hopefully, quasi-concavity is a generic feature of this type of objective function.

5.4.2 Empirical evidence

Using the parametric model

$$m(X, c) = c_0 + c_1 R_t + c_2 R_{t-1}$$

we employ our algorithm to compute the simple step function M-estimator of Chapter 4, for the IBM data previously described. Since (as discussed in Chapter 4), linear models can only be identified to scale, we set $c_1 = 0.1$. The estimator to be computed is:

$$\begin{aligned} b_N &\in \arg \min_{(c_0,c_2)\in B} - \int R_{t+1} \cdot I\left[c_0 + 0.1 r_t + c_2 R_{t-1}\right] dF_N \qquad (5.2) \\ B &= [-10, 10] \times [-1, 1] \end{aligned}$$

We now describe the step-by-step results from an implementation of the proposed algorithm.

Step 1: Obtain computational estimates of the minimising value of f_0 on B where:

$$f^0 \equiv - \int R_{t+1} \cdot \left(1 + \exp\left(-\frac{c_0 + 0.1 R_t + c_2 R_{t-1}}{\zeta^0}\right)\right)^{-1} dF_N \qquad (5.3)$$

These are:

$$b_c^0 = (0.0043, -0.5658)$$

Steps 2 and 3: Obtain a computational estimate for the solution to (5.2) after all[7] recursions of Step 2:

$$b_c^J = (0.0036, -0.4765)$$

where $\zeta^j = 0.85\zeta^{j-1}$ for $J > j > 0$ and $J = 100$.

The computational estimate for the minimum of f_N is:

$$f_N\left(b_c^J\right) = -9.4025 * 10^{-4} \qquad (5.4)$$

Figure 5.3 plots the sequence $\{b_c^j\}_{j=0}^J$ and the computational approximation to b_N obtained by a grid search described below:

[7] 39% of the recursions of this step resulted in improvements in the objective function indicating that the chosen sequence $\{m^j\}$ did not cause too many redundant optimisations which could be a source of wasted computing time.

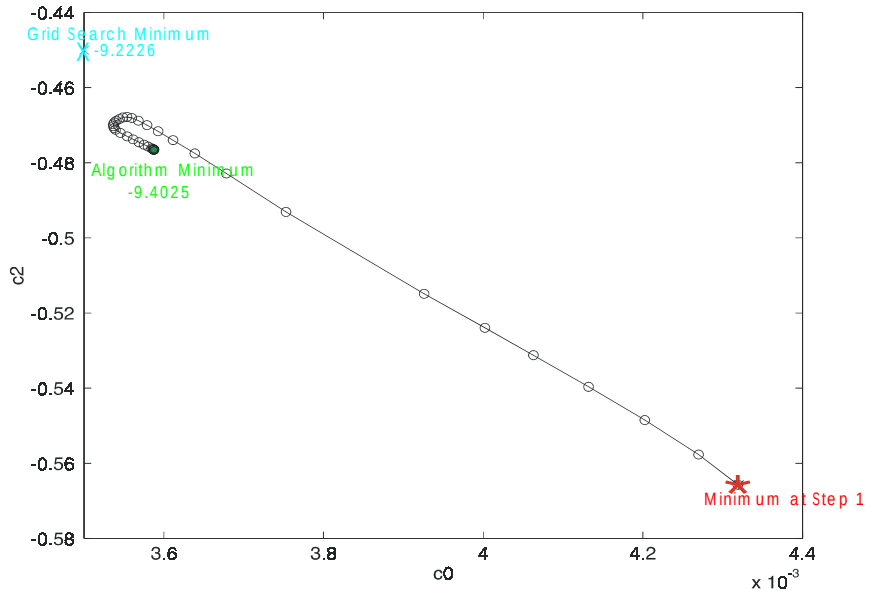

Figure 5.3. The circles display b_c^j computed at each recursion of Step 2 of the computational algorithm. The $*$ is the starting point b_c^0 computed by the genetic algorithm and the X is the grid-search parameter. The grid search is less accurate than the computational algorithm as indicated by an evaluation of f at the minima obtained by each method.

Grid Search: We performed a grid search (already used to plot Figures 7 and 8) over $81 \times 301 = 24381$ points spaced evenly in rectangles of dimension (0.0005×0.01) on (c_0, c_2)-space and obtained the following results:

The grid-search estimate for the solution to (5.2) is:

$$b_{grid} = (0.0035, -0.45)$$

The grid-search estimate for the minimum of f_N is[8] (compare to (5.4)):

$$f_N\left(b_{grid}\right) = -9.2226 * 10^{-4} \tag{5.5}$$

Our empirical results suggest the following:

- The proposed computational procedure is very accurate (slightly more accurate than a grid search which takes 10 times as long).
- The recursions of Step 2 lead to significant improvements over the point estimated in Step 1.

[8] Note that the grid-search minimum of f_0 was also obtained and performed worse than the minimum obtained by our genetic algorithm, confirming convergence to a global minimum in Step 1.

It therefore seems that the results of Proposition 5.1 seem to apply and that the algorithm is very effective in that it shifts the speed-accuracy frontier significantly outwards relative to standard techniques. Note that we should expect the algorithm's performance relative to the grid search to improve further as the number of parameters increases.

5.5 Conclusions

We have analysed the difficulties involved in computing the global minimum of the average of a certain type of random step functions. We proposed an intuitive optimisation algorithm and derived conditions under which we can show that it will perform well. Indeed, it does perform well in a simple but relevant empirical application.

Further research will compare the performance of this algorithm to competitors more sophisticated than grid-search, such as simulated annealing. It is hoped that the theoretical results may be extended, in particular by finding primitive conditions on the c.d.f. F of (Y, X) that imply the quasi-concavity condition used in Proposition 5.1.

5.6 Appendix

5.6.1 A. Proof of lemmata and proposition

Lemma 5.1 *Let A4.6 hold. Then $f^j(c)$ converges uniformly to $f(c)$ as $\zeta^j \to 0$, i.e.*

$$\lim_{\zeta_j \to 0} \sup_{c \in B} \left| f^j(c) - f(c) \right| = 0$$

Proof. Notice that for any α :

$$\begin{aligned} & \left| f^j(c) - f(c) \right| \\ \leq & \left| \int Y \cdot \left(\left(1 + \exp\left(-\frac{m(X,c)}{\zeta^j} \right) \right)^{-1} - \mathbf{1}\left[m(X,c) > 0 \right] \right) \cdot \mathbf{1}\left[|m(X,c)| \geq \alpha \right] dF \right| \\ & + \left| \int Y \cdot \left(\left(1 + \exp\left(-\frac{m(X,c)}{\zeta_j} \right) \right)^{-1} - \mathbf{1}\left[m(X,c) > 0 \right] \right) \cdot \mathbf{1}\left[|m(X,c)| < \alpha \right] dF \right| \end{aligned}$$

If $\alpha > 0$, the first term converges to zero uniformly over c as ζ^j tends to zero. The second term is smaller than:

$$\int |Y| \cdot \mathbf{1}\left[|m(X,c)| < \alpha \right] dF$$

So by the boundary condition (A6) for all ζ^j the second term also converges to zero uniformly over c. ■

Lemma 5.2 *Assume A4.2-4.4 and that there exists a* $D_\zeta : \mathcal{X} \to [0, \infty)$ *with* $\int D_\zeta(X)\, dF_X < \infty$ *and* $\left| Y \cdot \left(1 + \exp\left(-\frac{m(x,c)}{\zeta}\right)\right)^{-1} \right| \leq D_\zeta(x)$, *for all* (x, c) *in* $\mathcal{X} \times B$ *and* $\zeta > 0$. *Then for all* $\zeta^j \in \left[0, \zeta^0\right]$, $f_N^j(c)$ *converges uniformly almost surely to* $f^j(c)$ *as* $N \to \infty$

$$\Pr\left(\sup_{c \in B} \left| f_N^j(c) - f^j(c) \right| = 0\right) = 1$$

Proof. For all $\zeta^0 \geq \zeta^j > 0$, by Manski (1988*a*, Theorem 2', p.101), our assumptions imply:

$$\Pr\left(\lim_{N \to \infty} \sup_{c \in B} \left| f_N^j(c) - f^j(c) \right| = 0\right) = 1 \tag{5.6}$$

So $f_N^j(c)$ converges uniformly almost surely to $f^j(c)$ for all $\zeta^0 \geq \zeta^j > 0$. The same condition also holds for $\zeta^j = 0$ as is shown in the proof of Proposition 4. Hence convergence is uniform almost surely for all $\zeta^j \in \left[0, \zeta^0\right]$. ■

Lemma 5.3 *Assume* $f_N^j(c)$ *converges uniformly almost surely to* $f^j(c)$ *as* N *tends to infinity and that* c *is in a compact set* B. *Then* b_N^j *converges almost surely to* b^j *as* N *tends to infinity, i.e.*

$$\Pr\left(\lim_{N \to \infty} \left| b_N^j - b^j \right| = 0\right) = 1$$

Proof. See for example Amemiya (1985) Theorem 4.1.1 as modified by footnote 1. ■

Proof of Proposition 5.1. Suppose that for $j^* \in \{0, 1, ..., J\}$ it is the case that almost surely:

$$b_c^{j^*} = b_N^{j^*}.$$

By Lemma 3, for $N > N_\varepsilon$, there exists an N such that *for all* $j \in \{0, 1, ..., J\}$, when $N > N_\varepsilon$:

$$\left| b_N^j - b^j \right| < \varepsilon \text{ almost surely,} \tag{5.7}$$

so:

$$\left| b_c^{j^*} - b^{j^*} \right| < \varepsilon \text{ almost surely.} \tag{5.8}$$

Quasiconcavity of $f(c, \zeta)$ implies (by the theorem of the maximum) that b^j is a continuous function of ζ^j. Since continuity on a compact set implies equicontinuity, we have that b^j is an equicontinuous function of ζ_j on $[0, \zeta_0]$. Hence, there exists an α such that for all j:

$$\zeta^j - \zeta^{j+1} < \alpha \Rightarrow \left| b^{j+1} - b^j \right| < \varepsilon_\alpha \tag{5.9}$$

so using (5.8), for $N > N_\varepsilon$, almost surely:

$$\left|b_c^{j^*} - b^{j^*}\right| + \left|b^{j^*+1} - b^{j^*}\right| < \varepsilon + \varepsilon_\alpha.$$

Applying (5.7) for $j+1$, we obtain that for $N > N_\varepsilon$, almost surely:

$$\left|b_c^{j^*} - b^{j^*}\right| + \left|b^{j^*} - b^{j^*+1}\right| + \left|b^{j^*+1} - b_N^{j^*+1}\right| < 2\varepsilon + \varepsilon_\alpha,$$

so for $N > N_\varepsilon$, almost surely:

$$\left|b_c^{j^*} - b_N^{j^*+1}\right| < 2\varepsilon + \varepsilon_\alpha.$$

Let $2\varepsilon + \varepsilon_\alpha < \varepsilon^*$. Then, Assumptions 3b and 3c ensure that for $N > N_{\varepsilon_\alpha}$, almost surely:

$$b_c^{j^*+1} = b_N^{j^*+1}. \tag{5.10}$$

which means that (5.7) holds for all $j > j^*$. Now by the second assumption made for this proposition, (5.7) holds for $j^* = 0$, therefore it must hold for all j. Taking any *finite* sequence $\left\{\zeta_j\right\}_0^J$, such that $\zeta^j - \zeta^{j+1} < \alpha$ and for which $\zeta^J = 0$ it follows that:

$$\left|b_c^J - b_N^J\right| = 0, \text{ almost surely.}$$

Since from Proposition 1 as N tends to infinity,

$$\left|b_N^J - b_N\right| = 0, \text{ almost surely.}$$

It follows that as N tends to infinity,

$$\left|b_c^J - b_N\right| = 0, \text{ almost surely.}$$

■

5.6.2 B. Setting the smoothing parameter ζ_0

We assume we are working with zero mean (or demeaned series).

To set ζ^0 we propose the following procedure:

Step 1: Obtain an OLS estimate b^{ls} of the parameters of $m(X, c)$.

Step 2: Estimate the standard deviation σ of $m\left(X, b^{ls}\right)$.

Step 3: Set $\zeta^0 = \frac{2\sigma}{10}$.

The logic of this is the following:

We expect that $\Pr\left[-2\sigma < m\left(X, b^{ls}\right) < 2\sigma\right]$ is large given that for zero mean series, the mean of $m\left(X, b^{ls}\right)$ is likely to be small.

Now the range in which the function $(1 + \exp(*))^{-1}$ is curved is (say) $[-10, 10]$.

We can therefore set $\zeta^0 = \frac{2\sigma}{10}$ and expect that for most observations, $\frac{m(X,b_N)}{\zeta^0}$ is indeed smoothed.

We assume we are working with zero mean (or demeaned series).

To set ζ^0 we propose the following procedure:

Step 1: Obtain an OLS estimate b^{ls} of the parameters of $m(X,c)$.

Step 2: Estimate the standard deviation σ of $m\left(X,b^{ls}\right)$.

Step 3: Set $\zeta^0 = \frac{2\sigma}{10}$.

The logic of this is the following:

We expect that $\Pr\left[-2\sigma < m\left(X,b^{ls}\right) < 2\sigma\right]$ is large given that for zero mean series, the mean of $m\left(X,b^{ls}\right)$ is likely to be small.

Now the range in which the function $(1+\exp(*))^{-1}$ is curved is (say) $[-10,10]$.

We can therefore set $\zeta^0 = \frac{2\sigma}{10}$ and expect that for most observations, $\frac{m(X,b_N)}{\zeta^0}$ is indeed smoothed.

Bibliography

Acar, E. (1998), Expected returns of directional forecasters, *in* E. Acar & S. Satchell, eds, 'Advanced Trading Rules', Butterworth Heinemann.

Allen, F. & Karjalainen, R. (1999), 'Using genetic algorithms to find technical trading rules', *Journal of Financial Economics* **51**(2), 245–71.

Allen, H. & Taylor, M. (1990), 'Charts, noise and fundamentals in the London foreign exchange market', *Economic Journal* (100), 49–59.

Allen, H. & Taylor, M. (1992), 'The use of technical analysis in the foreign exchange market', *Journal of International Money and Finance* **11**, 304–314.

Allen, P. & Phang, H. (1994), Managing uncertainty in complex systems: Financial markets, *in* L. Leyesdorff & P. V. D. Besselaar, eds, 'Evolutionary Economics and Chaos Theory', Pinter, London.

Amemiya, T. (1985), *Advanced Econometrics*, Basil Blackwell.

Arthur, B. W., Holland, J. H., Palmer, B. L. R. & Tayler, P. (1996), Asset pricing under endogenous expectations in an artificial stock market, Working paper, Santa Fe Institute.

Barberis, N. (1999), 'Investing for the long run when returns are predictable', *Journal of Finance* .

Berger, J. O. (1985), *Statistical Decision Theory and Bayesian Analysis*, 2nd edn, Springer.

Bessembinder, H. & Chan, K. (1998), 'Market efficiency and the returns to technical analysis', *Financial Management* **27**(2), 5–17.

Bjornstad, J. F. (1990), 'Predictive likelihood: A review', *Statistical Science* **5**, 242–65.

Blume, L., Easley, D. & O'Hara, M. (1994), 'Market statistics and technical analysis: The role of volume', *Journal of Finance* **XIIX**(1).

Brandt, M. W. (1999), 'Estimating portfolio and consumption choice: A conditional Euler equations approach', *Journal of Finance* **54**(5).

Bray, M. M. & Savin, N. E. (1986), 'Rational expectations equilibria, learning and model specification', *Econometrica* **54**, 1129–1160.

Breen, W., Glosten, L. R. & Jagannathan, R. (1989), 'Economic significance of predictable variations in stock index returns', *Journal of Finance* **Vol XLIV**(No. 5), 1177–1189.

Brock, W., Lakonishok, J. & LeBaron, B. (1991), Simple technical trading rules and the stochastic properties of stock returns, Working paper, Santa Fe Institute.

Brock, W., Lakonishok, J. & LeBaron, B. (1992), 'Simple technical trading rules and the stochastic properties of stock returns', *Journal of Finance* **47**(5), 1731–64.

Brown, D. P. & Jennings, R. H. (1989), 'On technical analysis', *Review of Financial Studies* **2**(4), 527–55.

Brown, G. (1979), 'An optimization criterion for linear inverse regression', *Technometrics* **21**(4), 575–579.

Brown, S. J., Goetzmann, W. N. & Kumar, A. (1998), 'The Dow theory: William Peter Hamilton's track record reconsidered.', *Journal of Finance* **53**(4), 1311–33.

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

Campbell, J. Y. & Viceira, L. M. (1999), 'Consumption and portfolio decisions when expected returns are time varying', *Quarterly Journal of Economics* **114**(2), 433–95.

Chan, L. K. C., Jegadeesh, N. & Lakonishok, J. (1996), 'Momentum strategies', *Journal of Finance* **51**(5), 1681–1713.

Chiang, T. (1992), Technical Trading Rules Based on Classifier Systems: A New Approach to Learn from Experience, PhD thesis, University of California at San Diego.

Christofferson, P. F. & Diebold, F. X. (1996), 'Further results on forecasting and model selection under asymmetric loss', *Journal of Applied Econometrics* **11**, 561–571.

Cooper, I. & Kaplanis, E. (1994), 'Home bias in equity portfolios, inflation hedging and international capital market equilibrium', *Review of Financial Studies* **7**(1), 45–60.

Cornell, W. & Dietrich, J. (1978), 'The efficiency of the market for foreign exchange under floating echange rates', *Review of Economics and Statistics* **15**(4), 369–387.

Cumby, R. E. & Modest, D. M. (1987), 'Testing for market timing ability: A framework for forecast evaluation', *Journal of Financial Economics* **19**, 169–189.

Curcio, R. & Goodhart, C. (1991), Chartism: A controlled experiment, Discussion paper 124, Financial Markets Discussion Group Series, LSE.

Diebold, F. X., Gunther, T. & Tay, A. (1998), 'Evaluating density forecasts, with applications to financial risk management', *International Economic Review* **39**, 863–83.

Diebold, F. X., Hahn, J. & Tay, A. (1999), 'Multivariate density forecast evaluation and calibration in financial risk management: High-frequency returns on foreign exchange', *Review of Economics and Statistics* **81**, 661–73.

Diebold, F. X. & Mariano, R. S. (1995), 'Comparing predictive accuracy', *Journal of Business and Economic Statistics* **13**, 253–265.

Dorsey, R. E. & Mayer, W. J. (1995), 'Genetic algorithms for estimation problems with multiple optima, nondifferentiability and other irregular features', *Journal of Business and Economic Statistics* **13**(1).

Easley, D. & Rustichini, A. (1999), 'Choice without beliefs', *Econometrica* **67**, 1157–84.

Eckstein, Z. & Wolpin, K. I. (1989), 'The specification and estimation of dynamic stochastic discrete choice models', *Journal of Human Resources* **24**(4), 562–98.

Eisenhart, C. (1939), 'The interpretation of certain regression methods and their use in biological and industrial research', *Annals of Mathematical Statistics* **10**.

Ericsson, N. R., Hendry, D. F. & Mizon, G. E. (1998), 'Exogeneity, cointegration and economic policy analysis', *Journal of Business and Economic Statistics* **16**(4), 370–387.

Fama, E. (1970), 'Efficient capital markets: A review of theory and empirical work', *Journal of Finance* **25**, 383–417.

Fama, E. (1991), 'Efficient capital markets II', *Journal of Finance* **XLVL**(5), 1575–1617.

Fisher, G., Wilson, D. & Xu, K. (1998), 'An empirical analysis of term premiums using significance tests for stochastic dominance', *Economics Letters* **60**, 195–203.

Fornari, F. & Mele, A. (1994), 'A stochastic variance model for absolute returns', *Economics Letters* **46**, 211–214.

Fourgeaud, C., Gourieroux, C. & Pradel, J. (1986), 'Learning procedures and convergence to rationality', *Econometrica* **54**, 845–68.

Frankel, J. A. & Froot, K. (1990), 'Chartists, fundamentalists and trading in the foreign exchange market', *American Economic Review* **80**(2), 181–5.

Geisel, M. S. & Zellner, A. (1968), Sensitivity of control to uncertainty and the form of the criterion function, *in* 'The Future of Statistics', Academic Press, pp. 269–89.

Geisser, S. (1993), *Predictive Inference: An Introduction*, Chapman and Hall, New York.

Gencay, R. (1996), 'Non-linear prediction of security returns with moving average rules', *Journal of Forecasting* **15**, 165–174.

Gencay, R., Ballochi, G., Dacorogna, M., Olsen, R. & Pictet, O. (1999), Real-time trading models and the statistical properties of foreign exchange rates, mimeo., Olsen Associates.

Goldbaum, D. (forthcoming), 'A nonparametric examination of market information: Application to technical trading rules', *Journal of Empirical Finance* .

Gourieroux, C. & Monfort, A. (1995), *Statistics and Econometric Models*, Vol. 1, Cambridge University Press.

Granger, C. W. (1993), 'On the limitations of comparing mean squared forecast errors: Comment', *Journal of Forecasting* **12**, 651–652.

Granger, C. W. & Ding, Z. (1994*a*), 'Modeling volatility persistence of speculative returns: A new approach', *Journal of Econometrics* **73**, 185–215.

Granger, C. W. & Ding, Z. (1994*b*), Stylized facts on the temporal and distributional properties of daily data from speculative markets, working paper 93-38, University of California at San Diego.

Granger, C. W. & Pesaran, M. H. (2000), A decision theoretic approach to forecast evaluation, *in* W. S. Chan, W. K. Li & H. Tong, eds, 'Statistics and Finance: An Interface', Imperial College Press, London, pp. 261–78.

Granger, C. W. & Pesaran, M. H. (forthcoming), 'Economic and statistical measures of forecast accuracy', *Journal of Forecasting* .

Haerdle, W. & Nixdorf, R. (1987), 'Nonparametric sequential estimation of zeros and extrema of regression functions', *IEEE Transactions on information systems* **IT-33**(3), 367–372.

Hanoch, G. & Levy, H. (1969), 'The efficiency analysis of choices involving risk', *Review of Economic Studies* **36**, 335–346.

Hansen, B. E. (1994), 'Autoregressive conditional density estimation', *International Economic Review* **35**, 705–30.

Hansen, L. P. & Jagannathan, R. (1991), 'Implications of security market data for models of dynamic economies', *Journal of Political Economy* **99**(2), 225–262.

Hansen, L. P. & Richard, S. (1987), 'The role of conditioning information in deducing testable restrictions implied by dynamic asset pricing models', *Econometrica* **55**, 587–613.

Hendry, D. F. & Ericsson, N. R. (1991), 'Modeling the demand for narrow money in the United Kingdom and the United States', *European Economic Review* **35**, 833–886.

Henriksson, R. D. & Merton, R. C. (1981), 'On market timing and investment performance II: Statistical procedures for evaluating forecasting skills', *Journal of Business* **54**, 513–533.

Horowitz, J. L. (1992), 'A smoothed maximum score estimator for the binary response model', *Econometrica* **60**(3), 506–531.

Huber, P. J. (1996), *Robust Statistical Procedures*, 2nd edn, Society for Industrial and Applied Mathematics.

Hudson, R., Dempsey, M. & Keasey, K. (1996), 'A note on the weak form efficiency of capital markets: The application of simple technical trading rules to UK stock prices - 1935 to 1994', *Journal of Banking and Finance* **20**, 1121–1132.

Hunter, W. G. & Lamboy, W. F. (1981), 'A Bayesian analysis of the linear calibration problem', *Technometrics* **23**(4), 323–350.

Hussman, J. (1992), 'Market efficiency and inefficiency in rational expectations equilibria', *Journal of Economics Dynamics and Control* **16**, 655–680.

Jennrich, R. (1969), 'Asymptotic properties of non-linear least squares estimators', *Annals of Mathematical Statistics* **40**, 633–643.

Johnson, N. L. & Kotz, S. (1972), *Continuous Multivariate Distributions*, Wiley.

Kalotay, A. (1971), 'Structural solution to the linear calibration problem', *Technometrics* **13**(4), 761–769.

Kandel, S. & Stambaugh, R. (1996), 'On the predictability of stock returns: An asset-allocation perspective', *Journal of Finance* **LI**(2).

Katz, R. W. & Murphy, A. H. (1990), 'Quality/Value relationships for imperfect weather forecasts in a prototype multistage decision-making model', *Journal of Forecasting* **9**, 75–86.

Kaufman, P. (1978), *Commodity Trading Systems and Methods*, John Wiley and Sons.

Kho, B. (1996), 'Time-varying risk premia, volatility and technical trading rule profits: Evidence from foreign currency futures markets', *Journal of Financial Economics* **41**, 249–290.

Kim, J. & Pollard, D. (1990), 'Cube root asymptotics', *Annals of Statistics* **18**, 191–219.

Knez, P. & Ready, M. (1996), 'Estimating the profits from trading strategies', *Review of Financial Studies* **9**(4).

Kruthckoff, R. (1967), 'Classical and inverse regression methods of calibration', *Technometrics* pp. 425–439.

Latham, M. (1986), 'Informational efficiency and information subsets', *Journal of Finance* **XLI**(1), 39–52.

LeBaron, B. (1998*a*), An evolutionary bootstrap method for selecting dynamic trading strategies, Working paper, University of Wisconsin.

LeBaron, B. (1998*b*), Technical trading rules and regime shifts in foreign exchange, *in* E. Acar & S. Satchell, eds, 'Advanced Trading Rules', Butterworth Heinemann.

Lee, C. I. & Mathur, I. (1996), 'Trading rule profits in European currency spot rates', *Journal of Banking and Finance* **20**, 949–962.

Leitch, G. & Tanner, J. E. (1991), 'Economic forecast evaluation: Profits versus the conventional error measures', *American Economic Review* **81**(3), 580–590.

LeRoy, S. F. (1989), 'Efficient capital markets and martingales', *Journal of Economic Literature* **XXVII**, 1583–1621.

Levich, R. & Thomas, L. (1993), 'The significance of technical-trading rules' profits in the foreign exchange market: A bootstrap approach', *Journal of International Money and Finance* **56**, 269–290.

Lo, A. W. & MacKinlay, A. C. (1997), 'Maximising predictability in the stock and bond markets', *Macroeconomic Dynamics* **1**, 102–134.

Luttmer, E. G. J. (1999), 'What level of fixed costs can reconcile consumption and stock returns?', *Journal of Political Economy* **107**(5), 969–997.

Lwin, T. & Maritz, J. (1980), 'A note on the problem of statistical calibration', *Applied Statistics* **29**(2), 135–141.

Malkiel, B. G. (1992), Efficient market hypothesis, *in* J. Eatwell, M. Milgate & P. Newman, eds, 'The New Palgrave Dictionary of Banking and Finance', McMillan.

Malkiel, B. G. (1996), *A Random Walk Down Wall Street*, W.W. Norton and Company, New York.

Manski, C. (1975), 'Maximum score estimation of the stochastic utility model of choice', *Journal of Econometrics* (3), 205–228.

Manski, C. (1985), 'Semiparametric analysis of discrete response', *Journal of Econometrics* pp. 313–333.

Manski, C. (1988*a*), *Analog Methods in Econometrics*, Vol. 39 of *Monographs on Statistics and Applied Probability*, Chapman and Hall.

Manski, C. (1988*b*), 'Identification of binary response models', *Journal of the American Statistical Association* **83**(403).

Manski, C. F. (1991), 'Regression', *Journal of Economic Literature* **XXIX**(1), 34–50.

Manski, C. F. & Thompson, S. T. (1989), 'Estimation of best predictors of binary response', *Journal of Econometrics* **40**(1), 97–123.

Marcet, A. & Sargent, T. (1989), 'Convergence of least squares learning mechanisms in self referential linear stochastic models', *Journal of Economic Theory* **48**(2), 337–368.

Marimon, R. (1996), Learning from learning in economics, *in* D. M. Kreps & K. F. Wallis, eds, 'Advances in Economics and Econometrics: Theory and Applications', Vol. 1, Cambridge University Press.

Marimon, R., McGrattan, E. & Sargent, T. (1990), 'Money as a medium of exchange in an economy with artificially intelligent agents', *Journal of Economic Dynamics and Control* **14**, 329–74.

Merton, R. C. (1981), 'On market timing and investment performance, I: An equilibrium theory of market forecasts', *Journal of Business* **54**, 363–406.

Mills, T. C. (1996), 'Non-linear forecasting of financial time-series: An overview and some new models', *Journal of Forecasting* **15**, 123–135.

Murphy, J. (1986), *Technical Analysis of the Futures Markets*, New York Institute of Finance, New York.

Neely, C., Weller, P. & Dittmar, R. (1996), Is technical analysis in the foreign exchange market profitable? A genetic programming approach, Technical Report 1480, CEPR.

Neftci, S. N. (1991), 'Naive trading rules in financial markets and Wiener-Kolmogorov prediction theory: A study of 'technical analysis'', *Journal of Business* **64**(4).

Neyman, J. (1938), 'L'estimation statistique traitee come une probleme classique de probabilite', *Actualites sci. ind.* (739).

Park, S. & Zellner, A. (1979), 'Minimum expected loss (MELO) estimators for functions of parameters and structural coefficients of econometric models', *Journal of the American Statistical Association* **74**(365), 185–93.

Pau, L. F. (1991), 'Technical analysis for portfolio trading by syntactic pattern recognition', *Journal of Economic Dynamics and Control* **15**, 715–730.

Pesaran, M. H. & Timmerman, A. (1998), A recursive modelling approach to predicting UK stock returns, Working paper.

Pesaran, M. H. & Timmermann, A. (1995), 'Predictability of stock returns: Robustness and economic significance', *Journal of Finance* **L**(4), 1201–1228.

Pictet, O. V., Dacorogna, M. M., Dave, R., Chopard, B., Schirru, R. & Tomassini, M. (1996), 'Genetic algorithms with collective sharing for robust optimization in financial applications', *Neural Network World* **5**(4), 573–587.

Ready, M. J. (1997), Profits from technical trading rules, working paper, University of Wisconsin.

Sargent, T. J. (1993), *Bounded Rationality in Macroeconics*, Oxford University Press.

Satchell, S. & Timmermann, A. (1995), 'An assesment of the economic value of non-linear foreign exchange rate forecasts', *Journal of Forecasting* **14**, 477–497.

Scholes, M. & Williams, J. (1977), 'Estimating betas from nonsynchronous data', *Journal of Financial Economics* pp. 309–327.

Schwert, W. G. (1989), 'Why does stock market volatility change over time', *Journal of Finance* **XLIV**(5).

Skouras, S. (2000), Pseudo optimal models, mimeo., University of Cambridge.

Stoker, T. M. (1986), 'Consistent estimation of scaled coefficients', *Econometrica* **54**(6), 1461–1481.

Sullivan, R., Timmermann, A. & White, H. (1998), Dangers of data driven inference, Working paper 98-16, University of California at San Diego.

Sullivan, R., Timmermann, A. & White, H. (1999), 'Data-snooping, technical trading rule performance and the bootstrap', *Journal of Finance* **54**(5), 1647–91.

Sylla, R. (1993), Chartists' language, *in* J. Eatwell, M. Milgate & P. Newman, eds, 'The New Palgrave Dictionary of Money and Finance', Macmillan, London.

Taylor, S. J. (1986), *Modelling Financial Time Series*, John Wiley and Sons.

Taylor, S. J. (1994), 'Trading futures using a channel rule: A study of the predictive power of technical analysis with currency examples', *Journal of Futures Markets* **14**(2), 215–235.

Tolley, H. & Pope, R. (1988), 'Testing for stochastic dominance', *Journal of Agricultural Economics* .

Treynor, J. & Ferguson, R. (1985), 'In defense of technical analysis', *Journal of Finance* **XL**(3).

Tsybakov, A. (1988), Passive stochastic approximation, Working paper A-207, University of Bonn.

Tvede, L. (1997), *Business Cycles: From John Law to Chaos Theory*, Harwood Academic Publishers.

Wald, A. (1971), *Statistical Decision Functions*, second edn, Chelsea Publishing Company, Bronx, New York.

Weiss, A. A. (1996), 'Estimating time series models using the relevant cost function', *Journal of Applied Econometrics* **11**, 539–560.

West, K. D. (1994), 'Asymptotic inference about predictive ability', *Econometrica* **64**(3), 1067–1084.

West, K. D., Edison, H. & Cho, D. (1993), 'A utility-based comparison of some models of exchange rate volatility', *Journal of International Economics* **35**, 23–45.

White, H. (1980), 'Using least squares to approximate unknown regression functions', *International Economic Review* **21**(1), 149–70.

White, H. (1984), *Asymptotic Theory for Econometricians*, Academic Press.

White, H. (1994), *Estimation, Inference and Specification Analysis*, Econometric Society Monographs, 22, Cambridge University Press.

Zellner, A. (1973), The quality of economic policy-making when targets and costs of change are mis-specified, *in* 'Selected Readings in Econometrics and Economics: Essays in Honour of Jan Tinbergen', number Part 2, Macmillan, pp. 147–64.

Zellner, A. (1978), 'Estimation of functions of population means and regression coefficients including structural coefficients: A minimum expected loss (MELO) approach', *Journal of Econometrics* **8**, 127–58.

Zheng, X. (1998), 'A consistent nonparametric test of parametric regression models under conditional quantile restrictions', *Econometric Theory* **14**, 123–38.