EUROPEAN UNIVERSITY INSTITUTE
DEPARTMENT OF ECONOMICS

EUI Working Paper ECO No. 98/40

Risk Neutral Forecasting

Spyros Skouras

BADIA FIESOLANA, SAN DOMENICO (FI)

Printed in Italy in December 1998
European University Institute
Badia Fiesolana
I-50016 San Domenico (FI)
Italy

Risk Neutral Forecasting*

Spyros Skouras†

Abstract

This paper develops statistical and computational tools for modelling returns forecasts to be used by a risk neutral investor. Any forecast with the same sign as the conditional mean optimises the loss function derived from this agents' decision problem, so the class of optimal predictors is rather broad. We exploit the fact that optimal forecasting in this context can be seen as an extension of binary quantile regression in order to provide consistent estimators for optimal predictors. Further properties of these estimators are explored using simulations and favourable comparisons with least squares procedures are made. Unfortunately, our estimators are difficult to compute but an optimisation algorithm tailor-made for this purpose is provided. Our results provide a statistically valid method for selecting certain types of 'investment rules' according to popular optimality metrics.

Keywords: Asymmetric loss function, Investment rules, Forecasting, Estimation, Quantile regression of binary response, Computational optimisation, Genetic algorithm.

JEL: C51, C53, C63, G00, G19.

*I would like to thank Graham Mizon, seminar participants at the EUI workshops and particularly Soren Johansen and Ramon Marimon for some very useful comments on previous versions of this paper.

†European University Institute, Via dei Roccettini 9, I-50016, San Domenico di Fiesole, Italy, **Email:** skouras@datacomm.iue.it, **Fax**: ++301 3611588, **Tel**: 3623699.

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 the prediction problem facing a risk neutral investor and to propose techniques for the estimation of his optimal predictor. This agent solves a decision problem which (unlike that of most other agents) has a structure simple enough for a point forecast of returns to provide sufficient information for its solution. It therefore provides a natural starting point for studying prediction in an investment context as noted, for example, by Granger and Pesaran (1996).

The loss function corresponding to the risk neutral investor's decision problem has been widely used to assess the 'economic value' of various types of financial models. For example, linear (Pesaran and Timmerman, 1995) and non-linear (Satchell and Timmerman 1995) time-series models have been evaluated according to this metric, as have econometric models (Breen *et al.* 1989), technical trading rules (e.g. Sullivan *et al.* 1997), agnostic 'money-machines' such as neural nets (LeBaron 1998) or designs that take advantage of some 'market anomaly' (Sullivan *et al.* 1998).

The 'Risk Neutral Forecasting' techniques we propose make it feasible to estimate these models with the same criterion by which they are evaluated and as Granger (1993) notes, '*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*'. For some types of models, such as those based on technical trading rules, this is the *only* feasible estimation technique. Weiss (1996) discusses estimation of time

[1] See Tvede (1997).

[2] See Campbell *et al.* (1997), for an overview of the state of the art.

series models according to the relevant loss function, but many of the statements he makes do not apply without qualification when the loss function is that of the risk neutral investor.

The paper is organised as follows. In Section 2 we describe and motivate the risk neutral investment decision and build on this to define risk neutral best predictors. In Section 3 we relate the risk neutral best predictor to the conditional distribution of returns and the conditional distribution of their sign. Using these results we develop conditions which may be used for parametric modelling of risk neutral best predictors and which indicate that this is a generalisation of the problem of binary response prediction under asymmetric loss. Iin Section 4 we discuss why this approach is likely to be useful when there is risk of model misspecification. In Section 5 we derive conditions under which it is feasible to consistently estimate parametric models for risk neutral best predictors. We cannot analytically derive any further properties of our estimators but we investigate some of them using simple simulations. Section 6 discusses why computation of the estimators is difficult and proposes an algorithm that facilitates the estimation process. A summary of the main findings closes the paper.

2 Forecasting and Investments

2.1 The problem of forecasting returns in abstract

Let us begin with an abstract description of financial returns r_{t+1} from the current period t to $t+1$ as a random variable ($r_{t+1} : Z \to \mathbb{R}$, where Z is an unobserved sample space) satisfying:

$$\begin{aligned} r_{t+1} &= g(x_t) + u_t \\ E\left(u_t|\xi\right) &= 0,\ \xi \in X \end{aligned} \tag{1}$$

where x_t is a vector random variable $x_t : Z \to X \subseteq \mathbb{R}^K$, ($X$ is the sample space on which realisations ξ of x_t are observed), $g : X \to \mathbb{R}$ is a possibly

non-linear function, u_t is the disturbance and $E(*|\xi)$ is the expectation conditional on the event $x_t = \xi$.

The objective of any forecaster of r_{t+1} observing events in X is the determination of the functional form of a 'best predictor' of r_{t+1} conditional on $x_t = \xi$. The best predictor will minimise the expected value of a forecasters' loss when used as a forecast in a specific decision problem.

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

$$p(\xi) \in \arg\min_{\theta \in \mathbb{R}^1} \int L(r_{t+1}, \theta) dP|\xi, \ \xi \in X \tag{2}$$

where $P|\xi$ is the probability measure conditional on $x_t = \xi$ and $L : \mathbb{R}^2 \to \mathbb{R}$ is a loss function which gives the loss at $t+1$ when (at time t) it is predicted that r_{t+1} will be θ.

Obviously the form and the size of the set of solutions to (2) will depend crucially on the choice of the loss function (for examples illustrating this dependence, see Christoffersen and 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[3]. On the other hand, when the decision problem **is** known, it is always preferable to use loss functions derived from that decision problem.

In the following subsections we derive a loss function from a stylised model of a risk neutral agent's investment decision. The purpose of this is to understand the properties of a predictor which is best in investment contexts and to study its relation to the 'standard' measures of location which are often used as forecasts.

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

2.2 Forecasting returns in the context of a simple investment decision

Consider the simple one period cash-single asset allocation decision:

$$\begin{aligned} & \max_{\mu \in [-s,l]} E\left\{U(W_{t+1})|\xi\right\} \qquad (3) \\ s.t.\ W_{t+1} &= \mu W_t(1+r_{t+1}) + (1-\mu)W_t \end{aligned}$$

where μ is the fraction of wealth W_t invested in the asset and is constrained to be finite valued to capture borrowing and short-selling constraints.

The best predictor for an investor solving this decision problem can be derived only if there is a (known) function describing how the investment decision μ depends on a prediction θ of returns for r_{t+1}. This function may be used to determine the loss function $L\left(r_{t+1}, \theta\right)$. This loss function can then be plugged into (2) to obtain the best predictor. It is well known however that in a utility maximisation context, unless restrictive assumptions are imposed, μ is not a function of a scalar quantity such as θ, but of the whole conditional probability measure $P|\xi$ as well as the current level of wealth W_t. This implies that point forecasts of returns do not (in general) provide sufficient information for utility maximising investment behaviour.

Appropriate assumptions on $\{U, P|\xi\}$ can, of course, ensure that a scalar θ summarises the information necessary for utility maximising investors to solve (3), in which case point forecasts *are* sufficient for utility maximisation. These assumptions typically require investors to know certain carefully chosen properties of $P|\xi$ but not its mean[4]. Assuming also that this information is available, a function $\mu\left(\theta, \xi\right) \rightarrow \mathbb{R}$ can be derived (in some cases even analytically[5]), which can then be used to

[4]For example, under standard assumptions that make utility a function of the conditional mean and variance, for $\mu\left(\theta\right)$ to be derivable it would be necessary to know the conditional variance. Analogously, West *et al.*(1993) assume sucha utility function and that the mean is known so as to derive a loss function for predictions of the variance.

[5]See Campbell and Viceira (1996).

derive a loss function and thus a best predictor. Unfortunately, the necessary information on $P|\xi$ for the derivation of μ is never available in practice and use of 'crude assumptions' may lead to very misleading results.

Brandt (1998) recently argued that it may therefore be expedient to depart from the objective of predicting r_{t+1} to focus instead on direct prediction of the optimal proportion of invested wealth μ conditional on ξ. Under appropriate conditions, he shows that this can be achieved using a non-parametric model for the mapping $\mu : X \to [-s, l]$. This is an interesting way around the problem of the lack of necessary information on $P|\xi$ but it leads to a complete departure from the returns forecasting framework.

There is one important case in which we are not forced to choose between crude assumptions on $P|\xi$ or a departure from the objective of forecasting returns. This is the case with which we deal with in this paper and it arises when U is linear, i.e. investors are risk-neutral[6].

Risk neutral investment decisions merit special attention for the following reasons. Firstly, some 'very wealthy' investors seem to behave in this way (at least in making their decisions at the margin) as do certain institutional investors and individuals investing a very small proportion of their wealth. Secondly, the binary 'investment rules' researchers have hitherto examined (discussed in the introduction) are often selected to maximise expected returns (e.g. LeBaron 1998, Moody *et al.*, 1998)[7]. This practice is an implicit attempt to estimate solutions of the risk neutral investor's prediction problem and can therefore be better understood in a formal prediction framework. Lastly, risk neutrality is a widely used benchmark in financial economics that serves to develop an understanding of more general problems. The results of Merton (1981) and Skouras (1998) show that the optimal behaviour of risk neutral agents is useful information for more general agents.

[6]In this case, Brandt's approach which relies on estimation of Euler equations is not applicable, because there is no Euler equation to estimate!

[7]They are sometimes also selected to maximise Sharpe ratios, but Skouras (1998) shows that under fairly general conditions these two criteria are identical.

For these reasons, we feel safe that even under the following assumption, relevant and interesting conclusions can be drawn from our analysis.

Assumption 2.1 Investors are risk-neutral, i.e. $U(W_t) = aW_t + b$

Under A2.1 the investment decision (3) can be rewritten as:

$$\max_{\mu \in [-s,l]} \mu E(r_{t+1}|\xi), \ \xi \in X \tag{4}$$

In order to precisely characterise the set of solutions to (4), let us first introduce the following definition:

Definition 2.2 A sign-preserving transform is a mapping $\tau : \mathbb{R}^1 \to \mathbb{R}^1$ such that $\tau(y) > 0 \Leftrightarrow y > 0$.

The set of all sign preserving transforms[8] will be denoted $\mathbb{T}$. As $\mathbb{T}$ will play an important role in what is to come, we give some examples of its elements in Figure 1.

Insert Fig. 1 here

It is straightforward to show that a necessary and sufficient condition for $\mu(*)$ to be a solution to (4) is that:

$$\mu(\xi) \in (l+s) \cdot \mathbf{1}[\tau(g(\xi)) > 0] - s, \ \tau \in \mathbb{T}, \xi \in X \tag{5}$$

where $\mathbf{1}[*]$ takes the value 1 if the logical expression in the brackets is true and 0 otherwise.

Now suppose the risk neutral investor's forecast for r_{t+1} is θ. His loss is the difference in utility at time $t+1$ between the utility of a correct

[8] See Manski 1988b, p.737 for a characterisation of a small but important subset of $\mathbb{T}$

forecast and that obtained given that the forecast is θ[9]. From (4) it is obvious that this loss is:

$$\begin{aligned} L(r_{t+1},\theta) &= ((l+s)\cdot\mathbf{1}\left[r_{t+1}>0\right]-s)\,r_{t+1}- \\ &\quad ((l+s)\cdot\mathbf{1}\left[\theta>0\right]-s)\,r_{t+1} \end{aligned} \tag{6}$$

As is evident from the definition of a best predictor (2), a loss function can be replaced with any increasing linear transformation without affecting the set of best predictors. Hence the prediction problem of risk neutral investors is invariant with respect to feasible position size (as long as positions are finite) and we may multiply and subtract constants with respect to θ in (6) to obtain an equivalent loss function:

$$L(r_{t+1},\theta) = -r_{t+1}\cdot\mathbf{1}\left[\theta>0\right] \tag{7}$$

This expression may be treated as a simpler form of (6) but note that it may also be interpreted as the loss function derived from an agent solving (4) who is constrained from borrowing or short-selling ($s=0, l=1$). We define the risk neutral best predictor as the best predictor of any agent solving (4), henceforth ignoring all obvious time subscripts.

Definition 2.3 A **Risk Neutral Best Predictor** (RNBP) is a mapping $p: X\to\mathbb{R}$ that satisfies:

$$p(\xi)\in\arg\min_{\theta\in\mathbb{R}^1} -\int r\cdot\mathbf{1}\left[\theta>0\right]dP|\xi,\ \xi\in X \tag{8}$$

3 Statistical properties of the Risk Neutral Best Predictor

In this section we derive some statistical properites of the RNBP which clarify its relation to other predictors and are particularly useful for parametric modelling.

[9]We may also think of this loss as the price that would be paid *ex post* to the investment decision (i.e. at $t+1$) to have known the true value of r_{t+1} *a priori* (at t) rather than to rely on the available forecast of θ.

3.1 The relation to the conditional mean.

Simple manipulations of a risk neutral best predictor's definition formalise its relation to the conditional mean of returns.

Proposition 1 A function $p : X \to \mathbb{R}$ is a risk neutral best predictor if and only if:

(a) It satisfies:

$$p(\xi) \in \arg\min_{\theta \in \mathbb{R}^1} -\mathbf{1}[\theta > 0] \cdot g(\xi), \ \xi \in X \tag{9}$$

Equivalently,

(b) It is *a sign-preserving transform of the conditional mean,* i.e. for some $\tau \in \mathbb{T}$ it satisfies:

$$p(\xi) = \tau(g(\xi)), \ \xi \in X \tag{10}$$

Proof:

(a) The definition of a risk neutral best predictor (8) is equivalent to (9).

(b) This follows since (10) is equivalent to (9).■

Proposition 1 implies $g(*)$ is a RNBP which is not surprising since it is well known that risk neutral investors can make optimal decisions on the basis of conditional means. A little less obvious is the fact that any sign-preserving transform of $g(*)$ is also a RNBP. The importance of this derives form the fact that the space of functions included in $\mathbb{T}$ is large and that it may therefore be that there is a $\tau \in \mathbb{T}$ such that $\tau(g(*))$ is a simple function even though $g(*)$ is quite complicated (see figure 1). In this case, the information relevant to a risk neutral investor (only the sign of returns) will have a simple structure even though the conditional mean does not.

3.2 The relation to the sign of returns.

Rather than thinking of θ as a prediction for r, some researchers prefer to think of $\mathbf{1}[\theta > 0]$ as a prediction of (an indicator of) the sign of returns $\delta \equiv \mathbf{1}[r > 0]$. This interpretation indicates the need to understand the relationship between the sign of the best predictor and the distribution of the sign of returns.

In the first instance one might think that the sign of the best predictor is the forecast that maximises the probability of correctly forecasting the sign of returns. However this is not the case: A predictor that does not maximise the probability of a sign 'hit' but which is very good at getting the sign right when the stakes are high will be preferred by a risk neutral investor. The following numerical example illustrates this point.

Example 3.1: Suppose r_{t+1} is a discrete *i.i.d.* random variable such that $\Pr(r_{t+1} = 0.1) = 0.2$, $\Pr(r_{t+1} = -0.001) = 0.8$. The optimal predictor is a constant k because r_{t+1} is *i.i.d.*. Consider now the prediction $k < 0$. 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 $k > 0$ has only a 20% chance of getting the sign right but results in a long position and positive expected profits.

It is therefore clear that the 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 Define $A : X \to [0, 1]$ as:

$$A(\xi) \equiv \frac{E(|r| \,|\delta = 1, \xi)}{E(|r| \,|\delta = 1, \xi) + E(|r| \,|\delta = 0, \xi)}, \ \xi \in X \qquad (11)$$

A necessary and sufficient condition for a mapping $p : X \to \mathbb{R}$ to be a risk neutral best predictor is that it satisfies:

$$p(\xi) \in \arg\min_{\theta \in \mathbb{R}^1} -\mathbf{1}[\theta > 0] \cdot (A(\xi) - \Pr(\delta = 0|\xi)), \ \xi \in X \qquad (12)$$

Proof: See Appendix

Proposition 2 reveals the relationship between the best predictor and the conditional distribution of the sign of returns (the binary random variable δ). One should think of $A(*)$ as a measure of the magnitude of returns when they are positive ($|r|\,|\delta = 1$) in relation to their magnitude when they are negative ($|r|\,|\delta = 0$). The proposition implies that if the distribution of r is skewed towards the right 'enough', then the risk neutral investor should be long even if $\Pr(\delta = 0|\xi) > 0.5$. The measure $A(*)$ quantifies what is 'enough' in relation to $\Pr(\delta = 0|\xi)$.

The proposition also indicates that there is a relation between the conditional distribution of the sign of returns δ and the best predictor. The precise nature of this relation can easily be derived using our previous results as we now show.

Proposition 3 Let $Q_\alpha(\xi)$ be the α'th quantile of $\delta|\xi$ so that:

$$Q_\alpha(\xi) \equiv \min_{\zeta\in[0,1]} \zeta : \Pr(\delta \square \zeta|\xi) \geq \alpha,\ \xi \in X \tag{13}$$

Let also $Q_A(\xi)$ be the $A(\xi)$'th quantile of $\delta|\xi$ so that::

$$Q_A(\xi) \equiv \min_{\zeta\in[0,1]} \zeta : \Pr(\delta \square \zeta|\xi) \geq A(\xi),\ \xi \in X \tag{14}$$

The $A(\xi)$'th quantile of $\delta|\xi$ determines the sign of any risk neutral best predictor $p(\xi)$, since:

$$Q_A(\xi) = \mathbf{1}[p(\xi) > 0],\ \xi \in X$$

Proof:

¿From the c.d.f. of δ it is easy to verify that

$$Q_A(\xi) = \left\{ \begin{array}{c} 1 \text{ if } A(\xi) > \Pr(\delta = 0) \\ 0 \text{ if } A(\xi) \square \Pr(\delta = 0) \end{array} \right\},\ \xi \in X$$

So by (12) we obtain the desired result■

As simple as this proposition may be, its implications are quite surprising. In particular, we believe that the fact that the risk neutral best prediction problem turns out to be a problem in determining a moving (in the sense that it depends on the realisation of x) quantile of the sign of returns is quite unexpected. This fact can considerably simplify the problem of best prediction in certain simple cases. Here is an example of such a case:

Example 3.2: Suppose it can be established that $E(|r| \,|\delta = 1, \xi) = E(|r| \,|\delta = 0, \xi)$ for all ξ. Then $A(*) = \frac{1}{2}$ and hence $Q_A(*)$ is the median of $\delta|x$. Knowledge of the median of the conditional distribution of the sign of returns is sufficient for a risk neutral investor to make his optimal decisions.

More generally, when $A(*) = a$, risk neutral best predictors can be found by determining the a'th quantile of the binary response δ, a problem that has received considerable attention in the literature (e.g. Manski and Thompson, 1989). It is a striking feature of risk neutral forecasting that it can be seen as an extension of this problem.

3.3 Moment extremum properties for parametric modelling

It is often the case that forecasting can be effectively formulated as a parametric estimation problem. In these cases, a model $s : X \times B \to \mathbb{R}$ is specified ($B \subset \mathbb{R}^K$ is a parameter space and K is a positive integer) and it is assumed that the model contains the desired best predictor $p(*)$, i.e. for some $b \in B$ it is the case that $s(\xi, b) = p(\xi) \; \forall \; \xi$. The forecasting problem is then reduced to that of finding the most effective way of using the available data to estimate the parameter b. Sample analogs of moment extremum conditions known to be satisfied by $p(*)$ may in some cases be used to estimate b. The following proposition provides two such conditions satisfied by $p(*)$ which are shown in Section 5 to be usable for estimation of b.

Proposition 4 If there is a mapping $s : X \times B \to \mathbb{R}^1$ and an (unknown) parameter $c' \in B$ such that $s(*, c')$ is a Risk Neutral Best Predictor, then a necessary and sufficient condition for $s(*, b)$ to be a Risk Neutral Best Predictor *almost everywhere* on P_X (the measure on X) is:

(a)

$$b \in \arg\min_{c \in B} - \int r \cdot \mathbf{1}\left[s(x, c) > 0\right] dP \tag{15}$$

Equivalently,

(b)

$$b \in \arg\min_{c \in B} - \int \left(\delta - A\left(x\right)\right) \cdot \mathbf{1}\left[s\left(x, c\right) > 0\right] dP \tag{16}$$

Proof: See Appendix

Proposition 4 states something that should be intuitively obvious. Since by definition the best predictor maximises the expected value of profits conditional on all ξ it must also maximise the expected value of profits taken over the probability measure on X. Therefore it maximises unconditional expected profits, as expressed by conditions (15-16).

In the context of our discussion on the relation of the RNBP to the sign of returns, note that if $A(*)$ is a known constant, equation (16) corresponds exactly to the moment extremum condition typically used in parametric prediction of a binary response variable under an asymmetric loss function (see Manski 1988) or equivalently quantile regression.

The relationship (16) will only be useful for deriving b if $A(*)$ is known, which in most applications it is not. However, there are strong indications that it may be feasible to model $A(*)$ quite accurately. Two related observations suggest this: Firstly, that the conditional expected value of absolute returns $E\left(|r| \,|x\right)$ are highly predictable (Taylor 1986, Schwert 1989, Granger and Ding 1994a,1993, Mills 1996, Fornari and Mele 1994) at least in univariate contexts. This indicates that accurate

models for $E(|r||\delta, x)$ may be feasible from which $A(*)$ can be immediately derived. Secondly, it is often approximately the case that $|r|$ is independent of the sign of returns δ (Granger and Ding 1994b, Henriksson and Merton 1981) in which case we are in the scenario of Example 3.2 where we know that $A(*) = \frac{1}{2}$.

4 Parametric modelling of the risk neutral best predictor vs. modelling the conditional mean.

Since the conditional mean is itself a risk neutral best predictor, it may seem that modelling the latter is pointless since the risk neutral investor's optimal decision can be derived from a model of the conditional mean. If our models could be perfectly accurate, this conclusion would be valid. Unfortunately, in most applications they are known to be no more than working approximations. What matters then is to find approximations that are good ecnough for the purpose at hand. Metrics for judging the quality of approximations for the conditional mean are usually based on various statistical criteria such as least squares or nonparametric conditions and may not reflect the decision problem in which the model will be used. By contrast, in modelling risk neutral best predictors directly we will take into account the context for which the model is being developed.

The currently dominant approach to modelling conditional means of returns is OLS parametric estimation. The next section will compare this approach to one based on parametric modelling of the risk neutral best predictor using conditions (15-16).

4.1 Parametric models of the conditional mean

Suppose we postulate a parametric model $s(*, c)$, $c \in B$ for $g(*)$. We cannot know for sure whether for some $b \in B$, $s(\xi, b) = g(\xi), \xi \in X$, i.e. whether *the model is correctly specified*, but we hope that this is the case.

The least squares approach to modelling $g(*)$ involves the determination of a parameter b^{ols} such that:

$$b^{ols} \in \arg\min_{c \in B} \int (r - s(x,c))^2 \, dP \tag{17}$$

If - as hoped - the model is correctly specified, it can be shown that $s(\xi, b^{ols}) = g(\xi)$, $\xi \in X$. When this is not the case, little can be said about how 'good' a model $s(*, b^{ols})$ will be for $g(*)$ without a definition of 'good' and some information on the form of $\{g(*), s(*,*), P\}$. Similarly, Proposition 4 establishes that (15-16) may be used *only* when $s(*,c)$ is a correctly specified model for some sign-preserving transform of the conditional mean, i.e. $\exists\, \tau \in \mathbb{T} : s(\xi, c) = \tau(g(\xi)), \xi \in X$. It should be clear that whilst both modelling approaches require strong conditions on the accuracy of model specification, conditions for modelling the RNBP are much weaker than those required for parametric modelling of the conditional mean.

Furthermore, if a sign-preserving transform of the true conditional mean $g(*)$, is **not** an element of the parametric model ('the model is false'), then although a predictor that satisfies (15-16) is not a best predictor, it is a 'best *ex ante* predictor' (in the sense of Manski, 1988). Such a model is the best from the permissible class $s : X \times B \to \mathbb{R}^1$ at predicting r if the prediction as a function of x must be made before the realization of x is observed. Whilst such a predictor may be suboptimal *ex post* to the observation of $x = \xi$ for some $\xi \in X$, it is optimal averaging over all ξ according to their probability measure. This is a very desirable property for a model to possess even if it is only a second-best property and implies that the chosen model will maximise the risk neutral investor's utility (from within the permissible class) *ex ante* to the observation of ξ. By contrast, use of (17) in this case leads to a predictor which is best *ex ante* according to the irrelevant metric of least squares and hence the predictor itself has no useful interpretation.

4.1.1 The interpretation of models under each specification possibility.

The implications of the points we have made for our choice of modelling strategy are best illustrated by considering how their relative merits vary depending on the relation of the parametric model to the true conditional mean.

Case 1:

$$s(\xi, c) = g(\xi), \xi \in X \text{ for some } c \in B$$

In this (implausible) case, (15-16) and (17) are both valid conditions for the derivation of risk neutral best predictors. The only difference is that the size of the solution set to (15-16) may be larger.

Case 2:

$$\exists\, c \in B, \tau \in \mathbb{T} : s(\xi, c) = \tau(g(\xi)), \xi \in X$$

$$\begin{aligned} s(\xi, c) &\neq g(\xi) \ \forall c \in B, \xi \in \overline{X} \\ \text{where } \overline{X} &\subset X \text{ is a set with non-zero measure} \end{aligned}$$

In this case, using (17) will typically lead to an incorrect b. While $s(*, c)$, is mis-specified for the conditional mean, it is correctly specified as a model of some sign-preserving transform of it and hence use of (15-16) will lead to the selection of a parameter b such that $s(*, b)$ is a risk neutral best predictor. We illustrate this case with a simple example.

Example 4.1

Suppose the DGP of returns is given by the following non-linear process:

$$\begin{aligned} r &= x_2 (x_1 - 0.5)^3 + u \qquad (18) \\ (x_1, x_2) &\in X = (\mathbb{R}, \mathbb{R}_{++}) \\ E(u|\xi) &= 0, \xi \in X \end{aligned}$$

but is incorrectly believed to be linear in x and even worse the coefficient of x_1 and x_2 are wrongly fixed on the basis of *a priori* considerations so that

$$s(x, c) = c_0 + 0.1 x_1, c_0 \in R^1$$

(i.e. it is believed that there is no dependence on x_2.

The OLS predictor solves:

$$b^{ols} \in \arg\min_{c_0 \in \mathbb{R}^1} \int \left(x_2 \left(x_1 - 0.5\right)^3 - c_0 - 0.1x_1\right)^2 dP$$

and there is no reason to expect $s\left(*, b^{ols}\right)$ to satisfy (8), so the model is not a best predictor.

However, if we apply (15) we obtain:

$$b = \arg\min_{c_0 \in \mathbb{R}^1} - \int x_2 \left(x_1 - 0.5\right)^3 \cdot 1\left[c_0 + 0.1x_1 > 0\right] dP$$

It can easily be verified that there exists a $\tau \in T$ such that $-0.05 + 0.1\xi_1 = \tau\left(\xi_2 \left(\xi_1 - 0.5\right)^3\right)$ for all $\xi_1, \xi_2 \in X$ (as illustrated in figure 2) and hence $b = -0.05$ is a solution. Thus, $s\left(*, -0.05\right)$ is a sign-preserving transform of $g\left(*\right)$ (the conditional mean) and hence by Proposition 2(b) it is a best predictor for the risk-neutral investor despite the fact that the model used is completely mis-specified for the conditional mean.

Clearly, use of (15) should be preferred since it leads to the risk neutral best predictor whereas the OLS condition (17) does not.□

Insert Figure 2

Case 3:

$$\exists\ \xi \in X : s\left(\xi, c\right) \neq \tau\left(g\left(\xi\right)\right)\ \ \forall\ c \in B, \tau \in \mathbb{T}$$

This is the most probable scenario (at least in multivariate applications) and in this case the best predictor is not an element of the parametric model.

We are therefore called to reassess the interpretation of our metric for choosing c. Given that the best predictor is not attainable, what is a 'good' predictor? As we have mentioned, the best *ex ante* predictor given by (15) is good in a well defined sense: if a risk neutral investor had to choose amongst the use of a predictor in $s\left(*, c\right), c \in B$ before observing

the realisation of x, then the best *ex ante* predictor is the one he would choose. Instead of a solution to (3) the *ex ante* predictor solves[10]:

$$\begin{aligned} & \max_{c \in B} E\left\{W_{t+1}\right\} \\ s.t.\ W_{t+1} &= \mu W_t(1+r_{t+1}) + (1-\mu)W_t \\ \mu &= \left((l+s)\cdot \mathbf{1}\left[s\left(x,c\right)>0\right]-s\right) \end{aligned}$$

The following example shows that even when the DGP is simple and the model is only slightly mis-specified, the OLS predictor and the one derived from models of the RNBP will diverge.

Example 4.2

Suppose the returns' DGP is given by a simple AR(1) process:

$$\begin{aligned} r_{t+1} &= 0.001 + 0.1r_t + u \\ u &\sim N(0,\sigma) iid \\ \sigma &= 0.15 \end{aligned}$$

and that the parametric model $s\left(r_t, c\right) = c + 0.1r_t$, $c \in B$ 'approximately' contains the conditional mean but not quite. In particular, let:

$$B = (-\infty, -0.299] \vee [0.302, \infty)$$

Clearly this is a contrived restriction on the parameter set, but it allows us to illustrate (in the context of a simple process) how the criterion by which a parameter is chosen becomes crucial when mis-specification is even 'slight'.

Because of the symmetricity of the least-squares criterion, according to this criterion $c' = -0.3$ and $c'' = 0.302$ are equally good solutions and therefore $b^{ols} = -0.299$ is the best feasible parameter choice. However the same is not true if b^{ols} is evaluated as a solution to (15) because the objective function involved is asymmetric so c'' is better than b^{ols}. These facts are evident in Figure 3 below:

[10]The predictor given by (16) no longer coincides with that of (15) since the correct specification assumption of Proposition 4 is violated.

Insert figure 3

Our analysis suggests that use of $c'' = 0.302$ should be preferred by the risk-neutral investor because this choice leads to larger expected profits.□

4.2 Non-parametric models of the conditional mean.

We have argued that use of conditions (15-16) to model the RNBP is preferable over use of parametric conditions on the conditional mean (such as the least squares condition 17) particularly when it is likely that our model is mis-specified. However, non-parametric conditions on the conditional mean may provide complementary information about the RNBP e.g. to evaluate the correctness of the model specification. The striking feature of non-parametric estimation of conditional means of returns for the purpose of risk neutral best prediction is that we need only determine the behaviour of this function around zero (because then we can determine its sign everywhere). Non-parametric methods may be more effective in providing such local information about the conditional mean[11] than they have been as estimators of the entire mean's functional form.

5 Estimation of risk neutral best predictors

In this section we derive assumptions under which sample analogues of conditions (15-16) can be applied to consistently estimate parametric models of risk neutral best predictors. Our proofs draw on results developed by Manski (1988) for estimation of best predictors of a binary response under asymmetric loss functions. We report the results from some

[11]Tsybakov (1987) provides methods which may be used to recursively estimate the zeros of conditional means in an iid environment. Time series extensions of these results should be feasible and useful for our purposes.

simple simulations with which we explore the asymptotic behaviour of our estimators and compare them to least squares-maximum likelihood estimators.

5.1 Consistent estimators for the best predictors

In what follows, we make the following assumptions (which are extensions and adaptations of Manski's (1988) Conditions 6, 7 (p. 96-97), 8' (p. 108), 1a. (p.92) and 9, (p.103)):

Assumption 5.1 The parameter space $B \subseteq \mathbb{R}^K$ specifying potential solutions to the best predictor problem is compact. In the special case where B is discrete, only the next assumption is necessary.

Assumption 5.2 The empirical probability measure P^N, consisting of the observations $\{r_i, x_i\}_{i=1}^N$ satisfies uniform laws of large numbers. A simple case arises when the (r_i, x_i)'s are independent draws from P but see e.g. White (1984) for feasible extensions when time dependence is present.

Assumption 5.3 $\exists$ unique $b \in B$ s.t.:

$$-\int r \cdot \mathbf{1}\left[s\left(x, b\right) > 0\right] dP = \min_{c \in B} -\int r \cdot \mathbf{1}(s(x, c) > 0) dP$$

Whether this identifiability assumption holds will depend on the interaction of $\{s(x, c), P, B\}$ and must be ensured on a case-by-case basis by appropriate specification of B given our priors regarding the behaviour of P. A result we show that is of central importance for many applications is that if $s(x, c)$ is linear, identifiability (to scale) is ensured under weak regularity conditions (see Appendix B).

Assumption 5.4 There is a sign-preserving transform $\tau \in \mathbb{T}$ of $s(x, a)$ such that $\tau\left(s(x, *)\right)$ is equicontinuous on B, i.e. $\forall \alpha > 0$, $(\xi, a, c) \in (X \times B \times B)$,

$$\exists\ \delta_\alpha : |a - c| < \delta_\alpha \Rightarrow \left|\tau\left(s(\xi, a)\right) - \tau\left(s(\xi, c)\right)\right| < \alpha,\ \xi \in X$$

Equicontinuity of a parametric model $s(x, *)$ can be directly verified; some examples of such models are provided by Manski (Lemma 7, pp. 109-110) and are reproduced in Appendix B. The role of this assumption is to introduce appropriate *smoothness* in $\int r \cdot \mathbf{1}(s(x,c) > 0)dP$ without imposing assumptions on P.

Assumption 5.5 Boundary condition[12]:

$$\lim_{\alpha \to 0} \sup_{c \in B} \int_{X c \alpha} |r| \, dP = 0, \;\; X_{c\alpha} \equiv \{\xi \in X : -\alpha < s(\xi, c) < \alpha\}$$

This is an assumption that ensures that the probability ξ occurs s.t. $s(\xi, c)$ is close to zero is small. It serves to ensure *continuity* of $\int r \cdot \mathbf{1}(s(x,c) > 0)dP$. We derive some sufficient conditions for this and provide them in Appendix A.

It is relevant to note that if $s\,(*, c)$ is linear in x (for example because we restrict our attention to best linear prediction) A5.3-A5.5 become immediately satisfied under regularity conditions given in the Appendix, but identification can only be to scale, i.e. B must not include c and c' such that $c = ac'$, where a is a positive scalar. However, when $s\,(*, c)$ is non-linear it becomes difficult to accept that A5.3-A5.5 necessarily hold. This should serve as a warning to numerous researchers who routinely optimise neural-nets, technical trading rules and other non-linear functions over continuous parameter sets (e.g. Moody *et al.*, 1998, LeBaron 1998, Pictet *et al.* 1992) that their procedures may be inconsistent. We note that if parameter sets B are discrete, e.g. in studies that optimise technical trading rules according to a metric equivalent to (15) (e.g. Skouras 1997), consistency requires only A4.2.

The role of the assumptions we have imposed is to guarantee sufficient continuity of the expressions we are considering for laws of large numbers to guarantee uniform convergence of $\int r \cdot \mathbf{1}(s(x,c) > 0)dP^N$. Our contribution is to show that the structure of the risk neutral best predictor is such that a result of Manski (1988) is applicable.

[12] When we use (16), we assume a similar condition holds, replacing r with δ.

Proposition 5 Let B^N be the sample analogue of (15):

$$B^N \equiv \arg\min_{c \in B} - \int r \cdot \mathbf{1}\left[s\left(x,c\right) > 0\right] dP^N \tag{19}$$

Then under A5.1-5.5, as the sample size $N \to \infty$, the sample analogue converges almost surely to the parameter (15) defining either an *ex ante* or an *ex post* risk neutral best predictor[13], i.e.

$$B^N \longrightarrow b \equiv \arg\min_{c \in B} - \int r \cdot \mathbf{1}\left[s\left(x,c\right) > 0\right] dP \text{ almost surely}$$

Proof: Theorem 3', Chapter 7 of Manski (1988) applies to B^N and provides the desired result■

We confirm and illustrate the proposition with a simple simulated example. We will refer to any estimator for b as an estimator for Risk Neutral Forecasting.

Example 5.1

As in Example 4.2, suppose the DGP of returns is an AR(1) process:

$$\begin{aligned} r_{t+1} &= 0.001 + 0.1 \cdot r_t + u \\ u &\sim N(0, 0.15) \text{ iid} \end{aligned} \tag{20}$$

Let $c_0 \in B = [-0.01, 0.01]$ and $s(x,c) = c_0 + 0.1x$

Note that sufficient conditions (Appendix B) for A5.1-5 apply so consistent estimation is guaranteed by Proposition 5.

We simulate this series (setting $r_0 = E\left(r\right)$) and obtain 501 observations on r_t. We then estimate $\int r_{t+1} \cdot 1\left[c_0 + 0.1 r_t > 0\right] dP^N$ for $c_0 \in [-0.01, 0.01]$, $N = 500$ and plot this function in figure 4.

Insert Figure 4

[13]The chosen interpretation depends on whether we believe $s\left(*, c\right)$ is correctly specified as discussed in Section 4.

By Proposition 5, the maximum of this plot should converge to the maximum of $\int r_{t+1} 1\left[c_0 + 0.1 r_t > 0\right] dP$ a plot of which has been given in figure 3 . Indeed, the minima are very close. However, after 500 observations the objective function as a whole remains rather erratic, which we shall see makes computation of the minimum rather complicated.□

IWhen the conditions of Proposition 4 are satisfied, (16) is equivalent to (15) and we can also use a sample analogue of (16) to determine the risk neutral best predictor as the following proposition indicates:

Proposition 6 Let B_A^N be the sample analogue of (16), i.e.:

$$B_A^N \equiv \arg\min_{c \in B} - \int \left(\delta - A\left(x\right)\right) \mathbf{1}\left[s\left(x, c\right) > 0\right] dP^N \qquad (21)$$

If A5.1-5.5 are satisfied and there is a function $s : X \times B \to \mathbb{R}^1$ (where $B \subset \mathbb{R}^K$ is a parameter space and K is a positive integer) such that for a parameter $c \in B$, $s\left(*, c\right)$ is a Risk Neutral Best Predictor, as the sample size N increases, B_A^N converges almost surely to a parameter defining the (*ex post*) Risk Neutral Best Predictor a.e. P_X, i.e.:

$$\begin{array}{rcl} B_A^N & \longrightarrow & b \text{ almost surely} \\ \text{where } s\left(\xi, b\right) & \in & \arg\min_{\theta \in \mathbb{R}^1} - \int r \cdot \mathbf{1}\left[\theta > 0\right] dP|\xi, \text{ a.e. } P_X \end{array}$$

Proof: Theorem 3', Chapter 7 of Manski (1988) implies:

$$B_A^N \longrightarrow b = \arg\min_{c \in B} - \int \left(\delta - A\left(x\right)\right) \cdot \mathbf{1}\left[s\left(x, c\right) > 0\right] dP \text{ almost surely}$$

and Proposition 4b implies $s\left(x, b\right)$ is a risk neutral best predictor a.e..■

When the assumptions of this theorem are satisfied, the following loose argument indicates that there may be an efficiency advantage in using (21) over (19). Since δ takes the same values as $\mathbf{1}\left(g\left(x\right) > 0\right)$ with

some (hopefully large) probability (depending on the behaviour of u), the error term u does not affect (21) and hence with some probability there is no noise in the estimation process. Intuitively speaking, since $A(*)$ captures some of the relevant structure of $g(*)$, using this information should improve our estimators.

Of course, even if we believe in these assumptions, we may not wish to be as bold as Henriksson and Merton (1981) who assume that $A(*)$ is known. In this case, it may be possible to estimate $A(*)$ with a model $A^N(*)$ satisfying certain desirable convergence properties as N becomes large. Such a model may be trivial to formulate if, for example, $A(*)$ is known to be constant (We have discussed empirical studies which suggest this may be the case). If $A^N(*)$ converges uniformly to $A(*)$, then a variant of the estimator B_A^N may be used to identify the risk neutral best predictor.

Proposition 7 Let $\overline{B}_A^N$ be:

$$\overline{B}_A^N \equiv \arg\min_{c \in B} - \int (\delta - A^N(x)) \mathbf{1}[s(x,c) > 0] \, dP^N \qquad (22)$$

Then if there is a function $s : X \times B \to \mathbb{R}^1$ (where $B \subset \mathbb{R}^K$ is a parameter space and K is a positive integer) and a parameter $c \in B$ such that $s(*,c)$ is a RNBP (satisfies (8)) and if A5.1-5.5 are satisfied, as the sample size $N \to \infty$, $\overline{B}_A^N$ converges to the parameter defining the risk neutral best predictor:

$$\begin{array}{rcl} \overline{B}_A^N & \longrightarrow & b \text{ almost surely} \\ \text{where } s(x,b) & \in & \arg\min_{\theta \in \mathbb{R}^1} - \int r \cdot \mathbf{1}[\theta > 0] \, dP|\xi, \ \xi \in X \end{array}$$

Proof: Since Lemma A (in Appendix A) states that:

$$\lim_{N \to \infty} \sup_{c \in B} \left| \begin{array}{c} \int \left(\delta - A(x)^N\right) \cdot \mathbf{1}[s(x,c) > 0] \, dP^N - \\ \int \left(\delta - A(x)^N\right) \cdot \mathbf{1}[s(x,c) > 0] \, dP \end{array} \right| = 0$$

it follows using Lemmata 4 and 5 of Manski (1988) that Theorem 1' of Manski (1988) applies and therefore it must be that:

$$\begin{aligned} \lim_{N\to\infty} \sup_{c\in\overline{B}_A^N} |c - b'| &= 0 \\ \text{where } b' &\in \arg\min_{c\in B} -\int (\delta - A(x))\mathbf{1}\left[s(x,c) > 0\right] dP \end{aligned}$$

The assumption that the model is correctly specified and Proposition 4 complete the proof as they imply that $s(x,b) \in \arg\min_{\theta\in\mathbb{R}^1} -\int r \cdot \mathbf{1}\left[\theta > 0\right] dP|\xi$, $\xi \in X$■

Remark 1 *If $A(*)$ is known to be constant, there exist techniques (Zheng 1998) which allow us to judge whether $s(*,c)$ is a correctly specified model for the $A(\xi)$'th quantile of $\delta|\xi$. If this is the case, then it follows that $s(*,c)$ must also be a correctly specified model for the risk neutral best predictor. These results are applicable since we have shown that in this circumstance the problem of risk neutral best prediction is equivalent to a problem of quantile regression. It may be possible to extend these results to the case where $A(*)$ is not constant and provide a general test for correct specification of risk neutral best predictors.*

5.2 Estimator asymptotic distribution

It would be *very* convenient to have some analytical results concerning the rates of convergence and asymptotic distribution of our estimators as this is a necessary condition for judging their asymptotic efficiency and conducting hypothesis tests. Unfortunately there are no available results for estimators of the form we have developed or for the loss incurred from their use. Some results for estimators that are closely related are:

1) Pakes and Pollard (1989) show that if (r,x) were $i.i.d.$ our estimators would be asymptotically normal. The strict independence assumptions they use to derive this result makes it inapplicable in time-series contexts such as the ones we are interested in.

2) West (1996) provides an asymptotic normality result for out-of-sample loss from general loss functions. However, his results are not applicable here because the loss functions he considers are continuous.

3) The results of Chamberlain (1986) imply that if $A(*)$ is a known constant, the estimator B_A^N does *not* converge at a rate $\frac{1}{\sqrt{n}}$ even when it is consistent.

Rather than embark on the difficult but worthwhile task of deriving results which are applicable to the discontinuous loss function, time dependent case which the risk neutral investor faces, we will provide a simple simulation to investigate the convergence properties of the estimator and the loss incurred from its use.

5.2.1 Simulation 5.1

Consider the following DGP:

$$\begin{aligned} r_{t+1} &= 0.00015 + 0.0330 \cdot r_t + u \\ u &\sim N(0, 0.0108) \text{ iid} \end{aligned}$$

The parameters of this DGP were determined by using OLS to estimate an AR(1) model on a series of IBM stock prices which we will describe in Section 6.

We draw $N+1$ simulated observations of r_t from this DGP (setting $r_0 = E(r)$) and repeat till we obtain $T = 10,000$ such draws of size $N+1$ from P.

Suppose the parametric model to be estimated is $s(x, c) = c_0 + c_1 x$. Our discussion of the assumptions introduced in this section indicate that a linear model for a risk neutral best predictor can only be estimated to scale. We therefore estimate $s(x, c) = c + x$ instead:

$$B^N = \min_c - \int r_{t+1} \cdot \mathbf{1}[c + r_t > 0] \, dP^N$$

According to Proposition 5, B^N should converge to $\frac{0.001}{0.1} = 0.01$ as N becomes large. We do not know whether B^N has an asymptotic

distribution, but we hope to find out by sampling the distribution of B^N for various N.

We are also interested in the asymptotic distribution of:

$$E\left(L\left(r_{t+1}, B^N\right)\right) = -\int r_{t+1} \cdot \mathbf{1}\left[B^N + r_t > 0\right] dP^N$$

as knowledge of this distribution allows us to assess the asymptotic profitability arising from the use of B^N. Given the distribution of B^N we can determine this asymptotic distribution on the basis of analytical results for the form of the function $E\left(L\left(r_{t+1}, *\right)\right)$ provided in the Appendix.

The figures below plot the histograms of B^N and $E\left(L\left(r_{t+1}, B^N\right)\right)$ for various values of N.

Insert figure 5a-h

The following table also provides some insight into the behaviour of the estimator

N	$\Sigma_T \frac{B_i^N}{T}$	$V\left(B_i^N\right)$	$\Sigma_T \frac{-E\left(L\left(r_{t+1}, B_i^N\right)\right)}{T}$ *10^{-3}	$V\left(E\left(L\left(r_{t+1}, B_i^N\right)\right)\right)$ *10^{-8}
200	0.0001	0.0004	0.1662	0.5081
500	0.0022	0.0002	0.1736	0.4520
1000	0.0034	0.0002	0.1833	0.3723
2000	0.0041	0.0001	0.1954	0.2499
True values	0.0045		0.2340	

Table 1. The first column indicates the size of the sample in which B^N is optimised. The second and third column provide the mean and variances of the estimator over the 10000 simulations, and the last two colums provide the mean and variance of the expected profits these estimators would have yielded for the risk neutral investor. The last row gives the true B and the expected profits from use of this true B.

We draw the following conclusions:

i. Convergence to the true parameter is confirmed.

ii. The asymptotic distribution of B^N does not appear to be normal. The asymptotic distribution of $E\left(L\left(r_{t+1}, B_i^N\right)\right)$, the expected loss of this estimator is certainly non-normal.

iii. The estimate $\Sigma_T \frac{B_i^N}{T}$ is smaller than its true value for all N and this finite sample bias does not take into account the nature of the loss function which (as seen in Example 4.2 and figure 3) makes errors to the left of b more costly than errors to the right.

5.3 Comparison to least squares estimators of the conditional mean

As we have discussed, in the (improbable) case that a parametric model includes a correct specification for the conditional mean, least squares conditions can be used to find the best predictor. Indeed, in this case it may be preferable to estimate c using least squares methods as they are consistent under more general conditions than those required for asymptotic consistency of B^N. Furthermore, these estimators may have desirable efficiency properties (e.g. when normality guarantees they are also maximum likelihood estimators) which the estimators developed here do not share. Finally, their computational derivation is far easier and is supported by standard software.

These considerations motivate the following simple simulation which is intended to compare our estimator to the OLS estimator in a simple situation highly favourable to the latter. In particular, we consider the case where the model $s(*, c)$ contains a correct specification for the conditional mean and furthermore where the OLS estimator is also a Maximum Likelihood estimator.

5.3.1 Simulation 5.2

Using the same series as in Simulation 5.1, we derive:

$$\left(b_0^{ols}, b_1^{ols}\right) = \min_{(c_0,c_1)} \int \left(r_{t+1} - c_0 + c_1 r_t\right)^2 dP^N$$

We plot the histograms of $B^{ols} \equiv \frac{b_0^{ols}}{b_1^{ols}}$ and $E\left(L\left(r_{t+1}, B^{ols}\right)\right)$ for various N.

Insert figure 6a-h

We may compare the following table to Table 1..

N	$\Sigma_T \frac{B_i^{ols}}{T}$	$V\left(B_i^{ols}\right)$	$\Sigma_T \frac{-E\left(L\left(r_{t+1}, B_i^{ols}\right)\right)}{T}$ *10^{-3}	$V\left(E\left(L\left(r_{t+1}, B_i^{ols}\right)\right)\right)$ *10^{-8}
200	-0.0816	77.7744	0.1636	0.5427
500	-0.0010	0.1252	0.1734	0.4504
1000	0.0102	0.2371	0.1849	0.3452
2000	0.0162	0.7903	0.1984	0.2298
True values	0.0045		0.2340	

Table 2. The first column indicates the size of the sample in which B^N is optimised. The second and third column provide the mean and variances of the estimator over the 10000 simulations, and the last two colums provide the mean and variance of the expected profits these estimators would have yielded for the risk neutral investor. The last row gives the true B and the expected profits from use of this true B.

A comparison of these results to those of simulation 5.1 indicate the following

i. Convergence of B^{ols} to the true parameter is slower than that of B^N. Furthermore, it converges with a huge variance due to b_1^{ols} often

taking values very close to zero. This would occur infrequently if occur when b_1 were far from zero, but in financial applications coefficients are typically small.

ii. Whilst this may not be evident in the histogram, the estimator B^{ols} does converge to an asymptotic distribution. This is a non-central Cauchy distribution since it is derived from the ratio of two variables with a known joint normal distribution (see Papoulis 1984, p.136).

iii. In addition to these undesirable properties, B^{ols} performs worse than B^N in terms of loss in small samples. In other words, a risk neutral investor with a sample $N < 1000$ should prefer use of B^N.

iv. Nevertheless, eventually B^{ols} becomes a preferable estimator in terms of loss. This occurs because loss functions are bounded and hence when $b_1^{ols} \simeq 0$ loss does not explode even though $\left|B^{ols} - b\right|$ does.

We believe Simulations 5.1-5.2 provide surprisingly supportive evidence in favour of our estimator over an OLS-ML estimation approach when the objective is risk neutral best prediction. They indicate that even in the unrealistic scenario that is most favourable to OLS estimation, their is no clear evidence that this estimation technique is preferable.

6 An algorithm for computing the proposed estimator

With the theory of the previous sections in place, it seems that we are ready to estimate best predictors from financial data. Unfortunately, derivation of our estimators is in practice hindered by some serious computational obstacles.

In this section we will explain why the computational difficulties arise and propose an optimisation algorithm specially designed for the

particular problem at hand. Our discussion henceforth will be couched in terms of the first of the proposed estimators B^N, but is equally applicable to our other estimators B_A^N and $\overline{B}_A^N$. Since B_A^N can be interpreted as the estimator of a best predictor of a binary response under asymmetric absolute loss (possibly varying with x), the techniques developed here may also be useful for binary quantile regression.

6.1 Origins of computational difficulties

Our estimator has been defined (19) as:

$$B^N \equiv \arg\min_{c \in B} - \int r \cdot \mathbf{1}\left[s(x,c) > 0\right] dP^N$$

The reason it is difficult to computationally derive B^N is the same reason for which we have had to make special assumptions in order to ensure asymptotic consistency, namely the discontinuity of $\mathbf{1}\left[s(x,c) > 0\right]$. A direct implication of this discontinuity is that the objective function must necessarily have finite cardinality. This in turn means that the minimum we are searching for exists but that the objective function will in general be set-valued and hence it too will be discontinuous.

To understand this, think of the simple case where $s\left(*, c\right)$ is a linear function, i.e. $s\left(\xi, c\right) = \xi' c$, $\xi \in X$. Then the N hyperplanes defined by $c = \left[\gamma : \xi_i' \gamma = 0\right]$, $i = 1, 2, ..., N$ decompose B into at most $N^{\dim(X)} + 1$ regions in each of which $\int r \cdot \mathbf{1}\left[s(*, c) > 0\right] dP^N$ must be constant as a function of c[14].

Considering that the randomness inherent in the sampling process is carried over to a set-valued discontinuous objective function, it becomes evident that $\int r \cdot \mathbf{1}\left[s(*, c) > 0\right] dP^N$ will be a highly rugged object even when $\int r \cdot \mathbf{1}\left[s(*, c) > 0\right] dP$ is itself continuous[15]. Of course, as N tends to infinity these problems disappear but unfortunately this asymptotic result is not reflected in realistic sample sizes. To get a feel for the

[14] A similar point is made in Manski 1985, p.320.

[15] Our assumptions A4.3 and A4.5 are sufficient to ensure this by Manski 1988, Lemma 5., p.104.

problem, we may revisit Figure 4 which is a plot of the highly irregular $\int r_{t+1} \cdot \mathbf{1}\left[c_0 + 0.1 r_t > 0\right] dP^N$ as a function of c_0 with $N = 500$. Obviously, as the dimensionality of the objective function increases, so does the difficulty of the problem. We illustrate this fact by plotting the graph of $\int r_{t+1} \cdot \mathbf{1}\left[c_0 + 0.1 r_t + c_2 r_{t-1} > 0\right] dP^N$ with N =2049 observations on IBM daily closing prices[16] from 1st January 1990 through to 6th November 1997.

Insert figure 7 here

We have tried traditional optimisation techniques (such as simplex search, gradient descent and the less traditional genetic algorithm) on this type of problem but have observed a drastical failure to converge to a specific value for B^N. LeBaron (1998) and Pictet *et al.* (1996) have also encountered this problem when trying to optimise similar objective functions. It is likely that these computational difficulties have been an important obstacle for researchers who have previously made informal attempts to estimate models according to the types of loss function considered here.

6.2 The proposed algorithm

The computational procedure we propose is based on the following idea: begin by approximating $\int r_{t+1} \cdot \mathbf{1}\left[s(\xi, c) > 0\right] dP^N$ with a smooth function of c. This imposes continuity and eliminates much of the ruggedness of the landscape making its minimum relatively easy to find with a powerful global search procedure[17]. Next, make this approximation closer to $\int r_{t+1} \cdot \mathbf{1}\left[s(x, c) > 0\right] dP^N$ and use the minimum of the previous approximation as a starting point for a local search, repeating till the problem actually solved is the desired one - but is solved using a starting point that is (if all goes well) very close to its global minimum. As long as the problems solved along the way are not 'too' different, this is a reasonable

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

[17] This first element of our procedure has independently been used by LeBaron (1997).

property to expect from our procedure. We now describe the proposed algorithm in detail[18].

Step 1. Derive an estimator which is a smooth approximation to the desired estimator[19]:

$$B_0^N = \arg\min_{c \in B} - \int r \cdot \left(1 + \exp\left(-\frac{s(x,c)}{m^0}\right)\right)^{-1} dP^N \qquad (23)$$

where m^0 is a normalising constant[20] set so that a large proportion of the values of $\frac{s(x,c)}{m^0}$ lie in a region of the domain of $(1+\exp(-y))^{-1}$ where this function has some curvature. We illustrate the impact of this smoothening in figure 8 which shows its effect on the objective function obtained from the IBM series (figure 7).

Insert figure 8 here

To find B_0^N, optimisation methods that work well globally should be used. In particular, we propose the use of a genetic algorithm to find an initial maximum (see Dorsey and Mayer (1995) for evidence on the suitability of such an algorithm) from which we then initiate a simplex search. The computed estimate of B_0^N will be denoted $\boxed{B_0^N}$.

Let $i = 1$ and proceed to Step 2.

Step 2. Using $\boxed{B_{i-1}^N}$ as a starting point, derive $\boxed{B_i^N}$ a numerical approximation to B_i^N

[18]I would like to thank Domingo Tavella for a discussion that lead me in the direction of this algorithm.

[19]This smoothness makes West's (1996) results applicable and hence under general circumstances the out of sample losses from B_0^N will be asymptotically normal.

[20]These constants are determined as follows. First we estimate $s(x,c)$ by OLS. We then derive an estimate for the mean μ_{ols} and standard deviation σ_{ols} of $s\left(x, b^{ols}\right)$. Given that x has been demeaned, $\simeq 95\%$ of $s\left(x, b^{ols}\right)$ is in the range $(\mu_{ols} - 2\sigma_{ols}, \mu_{ols} + 2\sigma_{ols})$. Given also that μ_{ols} is usually small, we may assume that a large proportion of $s\left(x, B^N\right)$ lie in the region $(-2\sigma_{ols}, 2\sigma_{ols})$. The range in which the function $(1+\exp(*))$ is curved is (say) $[-10, 10]$. We therefore set $m_0 = \frac{2\sigma_{ols}}{10}$ to ensure that a large proportion of observations of $\frac{s\left(x_i, B^N\right)}{m^0}$ are in the desired region.

$$
\begin{aligned}
B_i^N &= \arg\min_{c\in B} -\int r\cdot\left(1+\exp\left(-\frac{s(x,c)}{m^i}\right)\right)^{-1} dP^N = 0 \\
m^i &= f\left(m^{i-1}\right)
\end{aligned}
$$

In our applications, we have used[21] $f(z) = 0.85 \cdot z.$.

Step 3. If $\int \left|-1+2\cdot\left(1+\exp\left(-\frac{s(x,c)}{m^i}\right)\right)^{-1}\right| dP^N = 1$ (which ensures that convergence to $-\int r\cdot 1\left(s(x,B^N)>0\right)dP^N$ has been achieved) then end, let $i=I$ and use $\boxed{B_I^N}$ as the estimate for B^N.

Otherwise, let $i = i+1$ and return to Step 2.

6.3 Some properties of the algorithm

All computational optimisation techniques aim to improve the speed of optimisation over an exhaustive grid search. A good technique is one that improves significantly over this speed without incurring a large cost in terms of a significant deterioration in performance.

One objection to the use of an algorithm such as the one proposed is that it could be that B_i^N varies drastically as a function of m^i. If so, even if one finds the global minimum of the first approximation, it may be that in the process of minimising better and better approximations to the desired problem, the algorithm gets stuck in a local minimum. 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 P and $s(x,c)$ as well as the efficiency of the computational procedures used.

With respect to this we note that exhaustive grid-searches nothwithstanding, we know of no alternative which would work on such a rugged landscape[22]. In trials we have run and which we will demonstrate

[21] A more sophisticated approach could let m^i be a function of $\left(\boxed{B_{i-1}^N}, \boxed{B_{i-2}^N}, ..., \boxed{B_0^N}\right)$ designed to accelerate and improve convergence

[22] Except perhaps the method of Pictet *et al.* (1996) the properties of which are still insufficiently understood.

below, the procedure has worked well for various N, as it is converges and does so much faster than grid searches.

Furthermore, as the following proposition shows, if b is a continuous function of m then this problem can be ruled out when N is 'large'. 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.

Proposition 8 Let:

$$b(c,m) \equiv \arg\min_{c\in B} -\int r \cdot \left(1+\exp\left(-\frac{s(x,c)}{m}\right)\right)^{-1} dP$$

$$b_i \equiv \arg\min_{c\in B} -\int r \cdot \left(1+\exp\left(-\frac{s(x,c)}{m^i}\right)\right)^{-1} dP$$

$$\overline{m} \equiv \max_i m^{i-1} - m^i$$

If (1) $b\,(c,m)$ is a quasi-concave function, (2) $B_i^N \to b_i$ almost surely[23] as $N \to \infty$, (3) The optimisation algorithms satisfy: (i) The optimum in Step 1 can be found (i.e. $\boxed{B_0^N} = B_0^N$) and (ii) for every subsequent step, optimisation works in an ε-neighbourhood of the solution (i.e. for a starting point B_s and some positive constant ε, $|B_s - B_i| \square\ \varepsilon \Rightarrow \boxed{B_i^N} = B_i^N$), **then** $\exists\ \overline{m} > 0$:

$$\boxed{B_I^N} \to b \text{ as } N \to \infty$$

Proof: See Appendix.

The most crucial assumption we make here is that $b\,(c,m)$ is a quasi-concave function. This is an unsatisfactory assumption because it does not relate to properties of P or $s\,(x,c)$. Unfortunately, we do not

[23] Conditions for (2) are weaker than the ones we have imposed for consistent estimation and are given by Manski 1988, Theorem 2', p.101)

know of conditions on $(P, s(*,*))$ which would guarantee this and it is difficult to check whether even in simple cases this is the case[24].

We conclude that the usefulness of the procedure we propose can only be evaluated in the context of a specific application. However, we believe the method is intuitively sensible and we have found it to be very effective on real data. The following simulation is an example of its effectiveness.

6.4 Empirical results from a simple model

Using the parametric model

$$s(x, c) = c_0 + c_1 x_1 + c_2 x_2$$

we employ our algorithm to compute an estimate of the risk neutral best predictor.

Since the model can only be identified to scale, set $c_1 = 0.1$ and compute:

$$\begin{aligned} B^N &\equiv \arg \min_{(c_0,c_2)\in B} -\int r_{t+1} \cdot \mathbf{1}\left[c_0 + 0.1 r_t + c_2 r_{t-1} > 0\right] dP^N \\ B &= [-10, 10] \times [-1, 1] \end{aligned}$$

We now describe the step-by-step results of the estimation procedure.

Step 1: The parameters estimated by the Genetic Algorithm are

$$B_0^N = \arg \min_{(c_0,c_2)\in B} -\int r \cdot \left(1 + \exp\left(-\frac{c_0 + 0.1 r_t + c_2 r_{t-1}}{m^0}\right)\right)^{-1} dP^N \tag{24}$$

$$\boxed{B_0^N} = (0.0043, -0.5658)$$

[24] We have been able to confirm that this is the case for the AR(1) model used in previous examples by plotting $b(c, m)$. Hopefully, quasi-concavity generalises to other cases.

The value of the objective function at this point is:

$$\int r \cdot \left(1 + \exp\left(-\frac{0.0043 + 0.1r_t - 0.5658r_{t-1}}{m^0}\right)\right)^{-1} dP^N = 8.4352 * 10^{-4}$$

The daily profits that would have been obtained (in sample) by a risk neutral investor using this estimated model are:

$$\int r_{t+1} \cdot \mathbf{1}\left[0.0043 + 0.1r_t - 0.5658r_{t-1} > 0\right] dP^N = 8.6936 * 10^{-4}$$

Step 2: The estimated parameter after all[25] recursions was:

$$B_I^N = (0.0036, -0.4765)$$

The daily profits that would have been obtained (in sample) by a risk neutral investor using this estimated model would be:

$$\int r_{t+1} \cdot \mathbf{1}\left[0.0036 + 0.1r_t - 0.4765r_{t-1} > 0\right] dP^N = 9.4025 * 10^{-4} \quad (25)$$

Grid Search Comparison

By comparison, the grid search over $81 \times 301 = 24381$ points spaced evenly in areas of size $(0.0005, 0.01)$ used to plot figures 7 and 8 produced the following results:

For the objective function in step 1 (i.e. the approximation to the risk neutral investor's loss function)

$$B_0^N(grid) = (0.0042, -0.55)$$

The value of the objective function at this point is (compare to (24)):

$$\int r \cdot \left(1 + \exp\left(-\frac{0.0042 + 0.1r_t - 0.55r_{t-1}}{m^0}\right)\right)^{-1} dP^N = 8.4325 * 10^{-4} \quad (26)$$

[25]39% of the recursions of this step resulted in improvements in the objective function.

For the loss function of the risk neutral investor:

$$B^N(grid) = (0.0035, -0.45)$$

The daily profits that would have been obtained (in sample) by a risk neutral investor using the parameters estimated by grid search would are (compare to (25)):

$$\int r_{t+1} \cdot \mathbf{1}\left[0.0036 + 0.1r_t - 0.4765r_{t-1} > 0\right] dP^N = 9.2226 * 10^{-4} \quad (27)$$

The figure below plots the sequence $\{B_i^N\}$ and the value for B^N obtained by grid search.

Insert Figure 9

We can draw the following conclusions from our results:

i. The proposed computational procedure for estimation of Step 1 is very accurate (more accurate than that obtained from a grid search with a fine grid).

ii. The recursions of Step 2 lead to significant improvements over the point estimated in Step 1 and the overall results of our computational procedure our better than that from the grid search.

iii. The procedure was approximately 10 times faster than the grid search.

iv. The parameters derived are not too different (scaled appropriately) to those estimated by OLS as described in Simulation 5.1.

It therefore seems that the results of Proposition 8 seem to apply and that the algorithm is very effective. Its effectiveness in comparison to the grid search will increase as the number of estimated parameters increases.

7 Conclusion

The purpose of this paper has been to develop techniques by which risk neutral investors can conduct Risk Neutral Forecasting - that is, estimation of their optimal point forecasts for financial returns. Risk Neutral Forecasting is a natural framework in which to fit estimation of optimal 'investment rules' such as technical trading rules or market timing rules. It provides the estimation counterpart to the literature which has evaluated the out of sample performance of returns models according to their 'economic value' as quantified by criteria equivalent to the risk neutral investor's loss function. When a model has economic value for a risk neutral investor, it also has value for other types of investors (Merton 1981, Skouras 1997).

The objective of Risk Neutral Forecasting is to use data to parametrically estimate the Risk Neutral Best Predictor. It is obviously easier to formulate a correctly specified model for some function that has the same sign as the conditional mean than for the conditional mean itself. Since any function that 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 it requires less stringent assumptions on the correctness of model specification.

Most of our analytical results are derived by exploiting the observation that Risk Neutral Forecasting can be seen as a generalisation of quantile regression of the sign of returns. This fact allows us to use some existing results to propose estimators for Risk Neutral Best Predictors that are asymptotically consistent. However, there are no available results on which we can build to derive further properties of our estimator so we explore these using some simulations.We find that our main estimator compares favourably with OLS procedures even when OLS is a maximum likelihood estimator.

There exists some empirical evidence indicating that it may be realistic to assume returns processes have some special features which make Risk Neutral Forecasting a standard exercise in quantile regression of a binary response. In this case, one of our estimators may be particularly

efficient and we may use existing results to test whether a parametric model is a correct specification for a Risk Neutral Best Predictor - or equivalently whether it contains an optimal investment rule.

One complication in implementing Risk Neutral Forecasting is the computational difficulty of the problem involved. We propose an optimisation algorithm that goes some way in overcoming this difficulty. The algorithm is justified using both simulations and theoretical results and should also be useful in quantile regression applications (which are known to be computationally demanding, see e.g. Koenker *et al.*, 1985).

There are a number of easy and interesting extensions, such as changing the risk neutral decision problem to include transaction costs and a second financial asset. There are also some difficult but very interesting questions that remain unanswered which relate to the properties of our estimators. In particular, a procedure for Risk Neutral Forecasting would be greatly improved by the derivation of an asymptotic distribution for the out of sample performance of estimated models since this could be used for model selection and validation. There is hope for the determination of such a distribution if the results of West (1996) can be extended to the case of discontinuous objective functions.

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. Such a hybrid model should provide improved understanding of those features of returns processes that are the most important determinants of investment behaviour.

8 Appendices

Appendix A: Proofs of lemmata and lengthy propositions Proof of Proposition 2

Since

$$\begin{aligned} g(x) &= E(r|x) = \int_{-\infty}^{0} r dP|x + \int_{0}^{\infty} r dP|x \\ &= \Pr(\delta = 0|x) E(r|\delta = 0, x) + \Pr(\delta = 1|x) E(r|\delta = 1, x) \\ &= -\Pr(\delta = 0|x) E(|r| \,|\delta = 0, x) + \Pr(\delta = 1|\xi) E(|r| \,|\delta = 1, x) \\ &= -\Pr(\delta = 0|x) \left[E(|r| \,|\delta = 0, x) + \Pr(\delta = 1|\xi) E(|r| \,|\delta = 1, x)\right] \\ &\quad + E(|r| \,|\delta = 1, x) \end{aligned}$$

Let:

$$z(x) \equiv \frac{1}{E(|r| \,|\delta = 0, x) + E(|r| \,|\delta = 1, x)}$$

Then $g(x) \cdot z(x) = A(x) - \Pr(\delta = 0|x)$

Since $z(x) > 0$, $g(x) \cdot z(x)$ is a sign-preserving transform of $g(x)$ so by Proposition 1b:

$$p(\xi) \in \arg\min_{\theta \in \mathbb{R}^1} -(A(\xi) - \Pr(\delta = 0|\xi)) \cdot \mathbf{1}[\theta > 0], \ \xi \in X \blacksquare$$

Proof of Proposition 4

(a)

$\Rightarrow$

Suppose $s(*, b)$ is a RNBP almost everywhere on P_X. Then:

$$s(\xi, b) \in \arg\min_{\theta \in \mathbb{R}^1} -\int r \cdot \mathbf{1}[\theta > 0] \, dP|\xi, \text{ a.e. } P_X$$

so it must be that:

$$-\int r \cdot \mathbf{1}[s(\xi, b) > 0] \, dP|\xi \ \square \ -\int r \cdot \mathbf{1}[s(x, c) > 0] \, dP|\xi, \ \forall c \in B \text{ a.e. } P_X$$

Summing over all inequalities for each $\xi \in X$:

$$-\int r \cdot \mathbf{1}\left[s\left(x,b\right)>0\right]dP \square -\int r \cdot \mathbf{1}\left[s\left(x,c\right)>0\right]dP,\ \forall c \in B,$$

So (15) is indeed a necessary condition.

$\Leftarrow$

Suppose $s\left(*,b\right)$ is not a RNBP almost everywhere on P_X. Then for some $\overline{X}$ with non-zero measure:

$$s\left(\xi,b\right) \notin \arg\min_{\theta \in \mathbb{R}^1} -\int r \cdot \mathbf{1}\left[\theta>0\right]dP|\xi,\ \xi \in \overline{X}$$

Since by assumption $s\left(x,c'\right)$ is a RNBP:

$$\begin{aligned}
-\int r \cdot \mathbf{1}\left[s\left(x,b\right)>0\right]dP|\xi &> -\int r \cdot \mathbf{1}\left[s\left(x,c'\right)>0\right]dP|\xi,\ \xi \in \overline{X} \\
-\int r \cdot \mathbf{1}\left[s\left(x,b\right)>0\right]dP|\xi &\geq -\int r \cdot \mathbf{1}\left[s\left(x,c'\right)>0\right]dP|\xi,\ \xi \notin \overline{X}
\end{aligned}$$

Summing over all inequalities for each $\xi \in X$:

$$-\int r \cdot \mathbf{1}\left[s\left(x,b\right)>0\right]dP > -\int r \cdot \mathbf{1}\left[s\left(x,c'\right)>0\right]dP$$

But this contradicts (15), so it is also a sufficient condition.

(b) Using Proposition 2 $s\left(*,b\right)$ is a RNBP iff

$$s\left(*,b\right) \in \arg\min_{\theta \in \mathbb{R}^1} -\mathbf{1}\left[\theta>0\right] \cdot \left(A\left(\xi\right)-\Pr\left(\delta=0|\xi\right)\right),\ \xi \in X$$

Using this fact and applying the same logic as in (a) we get the desired result.■

Proposition 7, Lemma A

Lemma A: Let

$$\begin{aligned}
h(x,c) &\equiv v(x)\mathbf{1}\left[s(x,c)>0\right] \\
h_N(x,c) &\equiv v_N(x)\mathbf{1}\left[s(x,c)>0\right] \\
v(x) &\equiv \delta - A\left(x\right) \\
v_N(x) &\equiv \delta - A^N\left(x\right)
\end{aligned}$$

Under the assumptions imposed, the sample expected loss function using the uniformly convergent model $A^N(x)$ for $A(x)$ converges uniformly to the true loss function:

$$\lim_{N\to\infty} \sup_{c\in B} \left| \int h_N(x,c)dP^N - \int h(x,c)dP \right| = 0$$

Proof of Lemma A

This proof extends Theorem 3', Chapter 7, Manski 1988 to the case where v is replaced with a uniformly consistent estimate for it.

Its proof utilises the following Lemma:

Lemma B:

If $A_N(x) \to A(x)$ uniformly, then

$$\max_{c\in B} \left| \int h_N(x,c)dP^N - \int h(x,c)dP \right| \to 0 \text{ a.s.}$$

Proof of Lemma B:

Since $h_N(x,c)$ is a linear function of $A_N(x)$ uniform convergence of the latter implies uniform convergence of the former.

This means:

$$\forall\ \varepsilon > 0, \exists\ N_0 : |h_N(x,c) - h(x,c)| < \varepsilon,\ x \in X,\ N > N_0$$

Hence:

$$\frac{1}{N}\sum_{i=1}^{N} |h_N(x_i,c) - h(x_i,c)| < \frac{1}{N}\sum_{i=1}^{N} \varepsilon = \varepsilon$$

But since $\frac{1}{N}\sum_{i=1}^{N} |h_N(x_i,c) - h(x_i,c)| \geq \left| \frac{1}{N}\sum_{i=1}^{N} h_N(x_i,c) - h(x_i,c) \right|$ this implies:

$$\left| \frac{1}{N}\sum_{i=1}^{N} h_N(x_i,c) - h(x_i,c) \right| < \varepsilon$$

Letting $N_0 \to \infty$, ε can be arbitrarily close to zero so it follows that:

$$\left| \int h_N(x,c)dP^N - \int h(x,c)dP^N \right| \to 0 \text{ a.s.} \qquad (28)$$

An appropriate LLN ensures that:

$$\left| \int h(x,c) dP^N - \int h(x,c) dP \right| \to 0 \text{ a.s.} \tag{29}$$

So combining (28) and (29):

$$\left| \int h_N(x,c) dP^N - \int h(x,c) dP^N \right| + \left| \int h(x,c) dP^N - \int h(x,c) dP \right| \to 0 \text{ a.s.} \tag{30}$$

Now notice that since $|a-b| + |b-c| \geq |a-c|$,

$$\begin{aligned} & \left| \int h_N(x,c) dP^N - \int h(x,c) dP^N \right| + \left| \int h(x,c) dP^N - \int h(x,c) dP \right| \\ \geq & \left| \int h_N(x,c) dP^N - \int h(x,c) dP \right| \end{aligned}$$

Using this fact and (30), we obtain:

$$\left| \int h_N(x,c) dP^N - \int h(x,c) dP \right| \to 0 \text{ a.s.}$$

and the proof to Lemma B is complete since this holds for all $c \in B$■

To prove Lemma A, we follow the logic of Manski (1988), Lemmata 5 and 6, pp 104-108:

$$\begin{aligned} & \left| \int h_N(x,a) - h_N(x,c) dP^N \right| \qquad (31) \\ = & \left| \int v^N(x) [\mathbf{1}(s(x,a) < 0) - \mathbf{1}(s(x,c) < 0)] dP^N \right| \\ \square & \int \left| v^N(x) [\mathbf{1}(s(x,a) < 0) - \mathbf{1}(s(x,c) < 0)] \right| dP^N \\ = & \int_{X(a,c)} \left| v^N(x) \right| dP^N \end{aligned}$$

where $X(a,c) \equiv \{\xi \in X : s(\xi,a) \square 0 \square s(\xi,c) \text{ or } s(\xi,a) \geq 0 \geq s(\xi,c)\}$

For $c \in B$, $\alpha > 0$, by the equicontinuity assumption (A5.4) which for notational simplicity (but without loss of generality) we assume holds for the identity function $\tau : \tau(x) = x$, it follows that $\exists \delta_\alpha : |a - c| < \delta_\alpha \Rightarrow$

$$\left\{ \begin{array}{c} s(\xi, c) > \alpha \Rightarrow s(\xi, a) > 0 \\ s(\xi, c) < -\alpha \Rightarrow s(\xi, a) < 0 \end{array} \right\} \forall a \in B, \xi \in X$$

Hence

$$|a - c| < \delta_\alpha \Rightarrow X(a, c) \subset X_{c\alpha} \equiv \{\xi \in X : -\alpha < s(\xi, c) < \alpha\}$$

And using (31) previously established:

$$|a - c| < \delta_\alpha \Rightarrow \left| \int h_N(x, a) - h_N(x, c) dP^N \right| \square \int_{Xc\alpha} \left| v^N(x) \right| dP^N$$

By identical reasoning, this condition holds if we replace $h_N(x, *)$ with $h(x, *)$ and P^N with P. Hence,

$$\begin{array}{ll} & |a - c| < \delta_\alpha \Rightarrow \\ & \left| \int h(x, a) - h(x, c) dP \right| + \left| \int h_N(x, a) - h_N(x, c) dP^N \right| \\ \square & \int_{X_{c\alpha}} |v(x)| \, dP + \int_{X_{c\alpha}} \left| v^N(x) \right| dP^N \end{array} \tag{32}$$

Now notice that:

$$\begin{array}{ll} & \left| \int h_N(x, a) dP^N - \int h(x, a) dP \right| \\ = & \left| \begin{array}{c} \int (h_N(x, a) - h_N(x, c)) \, dP^N - \int (h(x, a) - h(x, c)) \, dP + \\ + \int h_N(x, c) dP^N - \int h(x, c) dP \end{array} \right| \\ \square & \left| \int h_N(x, a) - h_N(x, c) dP^N \right| + \left| \int h(x, a) - h(x, c) dP \right| + \\ & + \left| \int h_N(x, c) dP^N - \int h(x, c) dP \right| \end{array} \tag{33}$$

Hence combining (32) and (33):

$$
\begin{aligned}
&|a-c| < \delta_\alpha \Rightarrow \\
&\left| \int h_N(x,a) dP^N - \int h(x,a) dP \right| \\
\square \quad &\int_{Xca} |v(x)|\, dP + \int_{Xca} \left|v^N(x)\right| dP^N + \left| \int h_N(x,c) dP^N - \int h(x,c) dP \right|
\end{aligned}
$$

Now the assumption of compactness of B (A5.1) implies directly that $\exists$ $B_\alpha \subset B$ s.t. $card(B_\alpha) < \infty$ and $c \in B_\alpha$ satisfies $|a-c| < \delta_\alpha$ $\forall a \in B$.

Hence $\forall$ $a \in B$,

$$
\begin{aligned}
&\left| \int h_N(x,a)\, dP^N - \int h(x,a)\, dP \right| \qquad (34) \\
\square \quad &\max_{c \in B_\alpha} \int_{X_{c\alpha}} |v(x)|\, dP + \max_{c \in B_\alpha} \int_{X_{c\alpha}} \left|v^N(x)\right| dP^N + \\
&+ \max_{c \in B_\alpha} \left| \int h_N(x,c) dP^N - \int h(x,c) dP \right|
\end{aligned}
$$

Now notice that by assumption $A_N(x) \to A(x)$ uniformly so $v^N(x) \to v(x)$ uniformly. Hence there is a N_0 s.t. for all $N > N_0$, $\left|v^N(x) - v(x)\right| \square \epsilon$

Since also $\left|v^N(x) - v(x)\right| \geq \left|v^N(x)\right| - |v(x)|$, it follows that:

$$\left|v^N(x)\right| - |v(x)| \square \epsilon$$

Hence for all $c, a \in B$

$$\int_{Xca} \left|v^N(x)\right| - |v(x)|\, dP^N \square \epsilon$$

Which implies that for all $c, a \in B$

$$\int_{X_{c\alpha}} \left|v^N(x)\right| dP^N \to \int_{X_{c\alpha}} |v(x)|\, dP^N \text{ a.s.}$$

The strong law of large numbers implies (given A4.2) that:

$$\int_{X_{c\alpha}} |v(x)|\, dP^N \to \int_{X_{c\alpha}} |v(x)|\, dP \text{ a.s.}$$

Hence for all $c, a \in B$

$$\int_{X_{c\alpha}} \left|v^N(x)\right| dP^N \to \int_{X_{c\alpha}} |v(x)|\, dP \text{ a.s.}$$

Which implies

$$\max_{c\in B_\alpha} \int_{X_{c\alpha}} \left|v^N(x)\right| dP^N \to \max_{c\in B_\alpha} \int_{X_{c\alpha}} |v(x)|\, dP \text{ a.s.}$$

Using Lemma B to ensure $\max_{c\in B_\alpha} \left|\int h_N(x,c)dP^N - \int h(x,c)dP\right| \to 0$ and (34), it follows that $\forall\alpha$, $\eta > 0$, $\exists\ N_{\alpha\eta} < \infty$ s.t.

$$\begin{aligned} & & N &> N_{\alpha\eta} \Rightarrow \\ \sup_{a\in B}\left|\int h_N(x,a)dP^N - \int h(x,a)dP\right| & \square & 2[&\max_{c\in B_\alpha} \int_{Xca} |v(x)|\, dP + \eta \\ & \square & 2[&\sup_{c\in B_\alpha} \int_{Xca} |v(x)|\, dP + \eta \end{aligned}$$

Now the boundary assumption (A5.5) implies that as $(\alpha, \eta) \to 0$ the required result is obtained thus completing the proof to Lemma A.■

Proof of Proposition 8

Assumption (2) implies that $\exists N_1, N_2 : \forall\varepsilon_1, \varepsilon_2$

$$\begin{aligned} \left|B_i^N - b_i\right| &\leqslant \varepsilon_1 \text{ for } N > N_1 \\ \left|B_{i+1}^N - b_i\right| &\leqslant \varepsilon_2 \text{ for } N > N_2 \end{aligned}$$

Hence for $N > \max(N_1, N_2)$

$$\left|B_i^N - b_i\right| + \left|B_{i+1}^N - b_{i+1}\right| \leqslant \varepsilon_1 + \varepsilon_2 \tag{35}$$

Since also,

$$\left|B_i^N - b_i\right| + \left|B_{i+1}^N - b_{i+1}\right| = \left|B_i^N - b_i\right| + \left|b_{i+1} - B_{i+1}^N\right| \tag{36}$$

and

$$\left|B_i^N - b_i\right| + \left|b_{i+1} - B_{i+1}^N\right| \geq \left|B_i^N - b_i + b_{i+1} - B_{i+1}^N\right| \tag{37}$$

Letting $\varepsilon_3 = b_{i+1} - b_i$, it follows from (35), (36) and (37) that:

$$\varepsilon_1 + \varepsilon_2 \geq \left|B_i^N - b_i\right| + \left|B_{i+1}^N - b_{i+1}\right| \geq \left|B_i^N - B_{i+1}^N + \varepsilon_3\right|$$

Since:

$$\left|B_i^N - B_{i+1}^N + \varepsilon_3\right| \geq \left|B_i^N - B_{i+1}^N\right| - |\varepsilon_3|$$

It follows that:

$$\varepsilon_1 + \varepsilon_2 + |\varepsilon_3| \geq \left|B_i^N - B_{i+1}^N\right|$$

Let $\varepsilon = \varepsilon_1 + \varepsilon_2 + |\varepsilon_3|$.

By assumption (1) and the theorem of the maximum it follows that $b(m)$ is continuous. Hence, $\forall \varepsilon$:

$$\exists\ N_0,\ \overline{m} > 0 : \forall N > N_0, i, \left|B_i^N - B_{i+1}^N\right| < \varepsilon \tag{38}$$

Using this fact and assumptions 3(i),(ii) we find that:

$\exists\ \overline{m} > 0 : \forall N > N_0,$

$$\boxed{B_i^N} = B_i^N$$

Since $B_I^N = B^N \to b$, for this $\overline{m}$, as $N \to \infty$,

$$\boxed{B_I^N} \to b \blacksquare$$

Appendix B: Sufficient conditions for required assumptions Sufficient conditions for equicontinuity (A5.3)

By Manski 1988, Lemma 7, pp. 109-110:

For some $\tau \in \mathbb{T}$, *at least one* of (a), (b) or (c) hold:

a) $X \times B$ is a compact metric space and $\tau\left(s\left(*,*\right)\right)$ is continuous on it.

b) $\tau\left(s(*,*)\right)$ is bounded on $X \times C$ and $s(\xi,*)$ is convex on C for all $\xi \in X$, where $B \subset C \subset \mathbb{R}^K$ and C is an open convex set.

c) $\tau\left(s(*,c)\right) = w(*)'c$, $(\xi,c) \in (X,B)$, $w : X \to \mathbb{R}^K$.

Sufficient conditions for Identifiability (A5.4):

We give such conditions on the basis of the following Proposition which is a simple extension of a result in Manski (1985).

Proposition 9: Sufficient conditions for A5.4 are:

1. For some $\tau \in \mathbb{T}, \tau\left(s(\xi,c)\right) = w(\xi)'c$, $(\xi,c) \in (X,B)$, $w : X \to \mathbb{R}^K$, and

2. The support of P_x is not contained in any proper linear subspace of $\mathbb{R}^K$, and

3. $b_k \neq 0$ for some k and $\forall x_{-k} \equiv (x_1, x_2, ..., x_{k-1}, x_{k+1}, ...x_K)$ the distribution of $x_k|x_{k-1}$ has everywhere positive Lebesgue density.

Proof: The conditions we have assumed imply by Lemma 2, Manski 1985, p. 317 that $\forall c \neq b$,

$$\begin{aligned} \int_{X_c} dP_x &> 0 \\ X_c &\equiv \left\{\xi \in \mathbb{R}^K : \mathbf{sign}\left[w(\xi)'c\right] \neq \mathbf{sign}\left[w(\xi)'b\right]\right\} \end{aligned}$$

Let

$$X_c' \equiv \left\{x \in \mathbb{R}^K : \mathbf{1}\left[w(\xi)'c > 0\right] \neq \mathbf{1}\left[w(\xi)'b > 0\right]\right\}$$

Clearly, $X_c' = X_c$. Therefore $\forall c \neq b$:

$$\int r \cdot \mathbf{1}\left[x'c > 0\right] dP \neq \int r \cdot \mathbf{1}\left[x'b > 0\right] dP$$

and hence the minimum of the r.h.s. must be unique ensuring identifiability.■

Sufficient conditions for the boundary condition (A5.5)

According to Lemma 8, Manski 1988, pp. 110-111, the following three conditions must hold

1.For some $\tau \in \mathbb{T}$, $\tau\left(s(*,c)\right) = w(*)'c$, $(\xi,c) \in (X,B)$, $w : X \to Z \subset \mathbb{R}^K$

2. $\forall\,(c,\omega) \in B \times V$, where V is the range space of $|r|$, the probability measure $P_{w(\xi)'c}|\omega$ is absolutely continuous w.r.t. the Lebesgue measure μ and also $\forall\,\eta \in \mathbb{R}^1$, $\exists\,\lambda < \infty$ s.t. $\phi_\mu(\eta, P_{w(\xi)'c}|\omega) < \lambda$,

3. $\int |r|\, dP_x$ exists

Appendix C: The distribution of predictor profits in a special case. In this Appendix we derive general expressions for the distribution of a risk neutral investor's profits (the negative of the loss function we have used throughout the text) when a random variable that is jointly normal with returns is used as a prediction[26]. (These results are used in simulations 5.1 and 5.2).

Let
$$\begin{bmatrix} y_1' \\ y_2' \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_{12}^2 \\ \sigma_{12}^2 & \sigma_2^2 \end{bmatrix} \right)$$

Suppose y_2' is a forecast for y_1'

Let x be the returns obtained from the use of this forecast. Then,

$$x = y_1' \cdot \mathbf{1}\,(y_2' > 0)$$

Let
$$y_1 \equiv \frac{y_1' - \mu_1}{\sigma_1}, y_2 \equiv \frac{y_2' - \mu_2}{\sigma_2}, \rho = \frac{\sigma_{12}}{(\sigma_1\sigma_2)^{\frac{1}{2}}}$$

Then
$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho^2 \\ \rho^2 & 1 \end{bmatrix} \right)$$

And
$$x = \mu_1 \cdot \mathbf{1}\left(y_2 > -\frac{\mu_2}{\sigma_2}\right) + \sigma_1 y_1 \cdot \mathbf{1}\left(y_2 > -\frac{\mu_2}{\sigma_2}\right)$$

So
$$x|y_{1,}y_2 = \left\{ \begin{array}{c} \mu_1 + \sigma_1 y_1 \text{ if } y_2 > -\frac{\mu_2}{\sigma_2} \\ 0 \text{ otherwise} \end{array} \right\}$$

and therefore the p.d.f. f_x of x is:

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

$$f_x(x) =$$

$$\left\{ \begin{array}{c} f_{y_1|y_2>-\frac{\mu_2}{\sigma_2}}\left(\frac{x-\mu_1}{\sigma_1}\right) * \Pr\left(y_2 > -\frac{\mu_2}{\sigma_2}\right) \text{ for } x \neq 0 \\ f_{y_1|y_2>-\frac{\mu_2}{\sigma_2}}\left(\frac{x-\mu_1}{\sigma_1}\right) * \Pr\left(y_2 > -\frac{\mu_2}{\sigma_2}\right) + \left(1 - \Pr\left(y_2 > -\frac{\mu_2}{\sigma_2}\right)\right) \text{ for } x = 0 \end{array} \right\}$$

Which may be written as

$$f_x(x) =$$

$$\int_{-\frac{\mu_2}{\sigma_2}}^{\infty} f_{y_1,y_2}\left(\frac{x-\mu_1}{\sigma_1}\right) dy_2 * \Pr\left(y_2 > -\frac{\mu_2}{\sigma_2}\right) + \mathbf{1}(w=0)\left(1 - \Pr\left(y_2 > -\frac{\mu_2}{\sigma_2}\right)\right) \tag{39}$$

Letting Φ be the cdf of the standard normal, this implies:

$$\begin{aligned} E(x) &= \mu_1 \Pr\left(y_2 > -\frac{\mu_2}{\sigma_2}\right) + \sigma_1 E\left(y_1 \mathbf{1}\left(y_2 > -\frac{\mu_2}{\sigma_2}\right)\right) \\ &= \mu_1\left(1 - \Phi\left(-\frac{\mu_2}{\sigma_2}\right)\right) + \sigma_1 \int_{-\infty}^{\infty}\int_{-\frac{\mu_2}{\sigma_2}}^{\infty} y_1 f(y_1,y_2)\, dy_2 dy_1 \end{aligned}$$

Johnson and Kotz (1972) report results (p. 113) that imply:

$$\int_{-\infty}^{\infty}\int_{-\frac{\mu_2}{\sigma_2}}^{\infty} y_1 f(y_1,y_2)\, dy_2 dy_1 = \frac{\rho}{\sqrt{2\pi}} \exp\left(-0.5\left(\frac{\mu_2}{\sigma_2}\right)^2\right)$$

Hence the Expectation of x is:

$$E(x) = \mu_1\left(1 - \Phi\left(-\frac{\mu_2}{\sigma_2}\right)\right) + \sigma_1 \frac{\rho}{\sqrt{2\pi}} \exp\left(-\left(\frac{\mu_2}{\sigma_2}\right)^2\right) \tag{40}$$

The Variance of the strategy can also be calculated if it is desired, by using the fact that $Var(x) = E(x^2) - E(x)^2$, (40) and an expression for $E(x^2)$ provided by Johnson and Kotz (p.113).

Example

Suppose that:

$$\begin{aligned} y_1' &= r_{t+1} = b_0 + b_1 r_t + u_t \\ u_t &\sim N(0, \sigma_u) \\ y_2' &= c_0 + c_1 r_t \end{aligned}$$

Then it follows that:

$$\square \begin{bmatrix} y_1' \\ y_2' \end{bmatrix} \sim N\left(\begin{bmatrix} \frac{b_0}{1-b_1} \\ c_0 + c_1 \frac{b_0}{1-b_1} \end{bmatrix}, \begin{bmatrix} \frac{\sigma_u^2}{1-b_1^2} & c_1 b_1 \frac{\sigma_u^2}{1-b_1^2} \\ c_1 b_1 \frac{\sigma_u^2}{1-b_1^2} & c_1^2 \frac{\sigma_u^2}{1-b_1^2} \end{bmatrix} \right)$$

Substituting this back into (**??**,40), we can obtain exact values for the p.d.f., mean and even the variance of profits obtained from using an AR(1) forecast for an AR(1) series.

For the mean this becomes:

$E(x) =$

$$\frac{b_0}{1-b_1} q + \frac{\sigma_u b_1}{\sqrt{2\pi(1-b_1^2)}} \exp\left(-\left(\frac{c_0}{|c_1|} + \frac{b_0}{1-b_1} \right)^2 \frac{b_1^2 - 1}{\sigma_u^2} \right)$$

where $q\left(b_0, b_1, \sigma_u, \frac{c_0}{c_1}\right) =$

$$1 - \Phi\left(-\left(\frac{c_0}{|c_1|} + \frac{b_0}{1-b_1} \right) \frac{\sqrt{1-b_1^2}}{\sigma_u} \right)$$

Notice that the first order condition w.r.t. $\frac{c_0}{c_1}$ may be used to confirm that optimally, $\frac{c_0}{c_1} = \frac{b_0}{b_1}$.

For the parameters in Simulation 5.1

$$\begin{aligned} b_0 &= 0.0015 \\ b_1 &= 0.0330 \\ \sigma_u &= 0.0108 \\ c_0 &= b_0 \\ c_1 &= b_1 \end{aligned}$$

We find that: $E(x) = 0.0234$

9 Figures

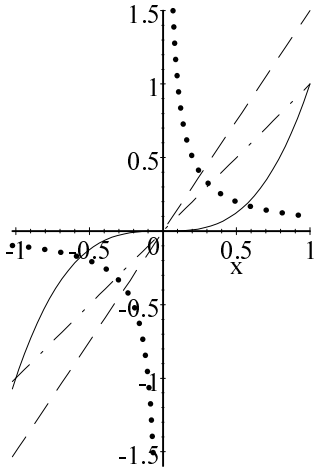

Figure 1: The mappings displayed are sign preserving transforms.

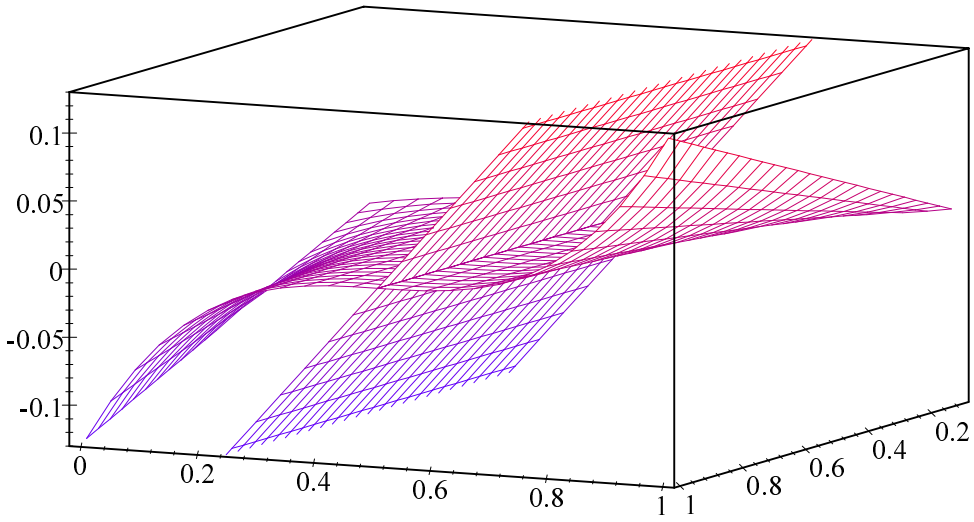

Figure 2: This figure illustrates that the parametric model is a sign-preserving transform of the DGP in Example 4.1. The curved surface is the DGP and the linear hyperplane is the model for it.

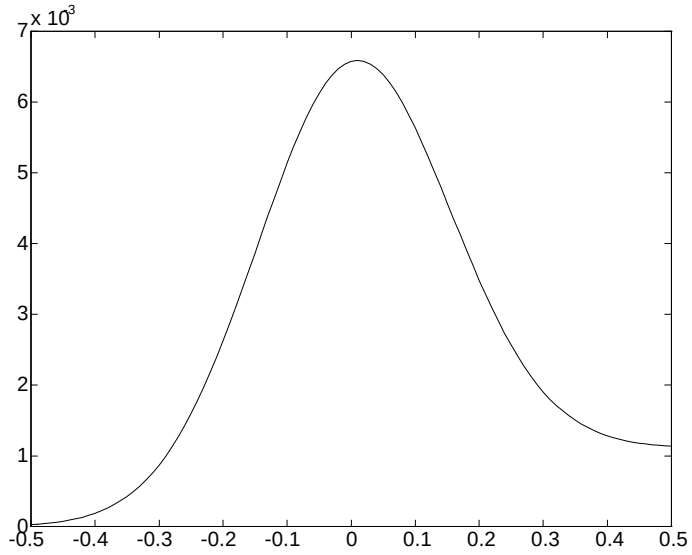

Figure 3: This plots the asymmetric profit function as described in Example 4.2

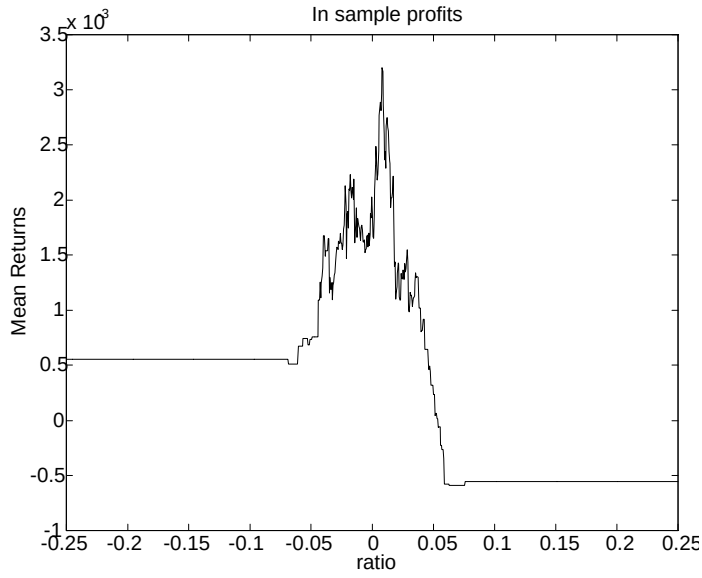

Figure 4: This figure depicts returns obtained for various choices of c_0 in the sample of Example 5.1.

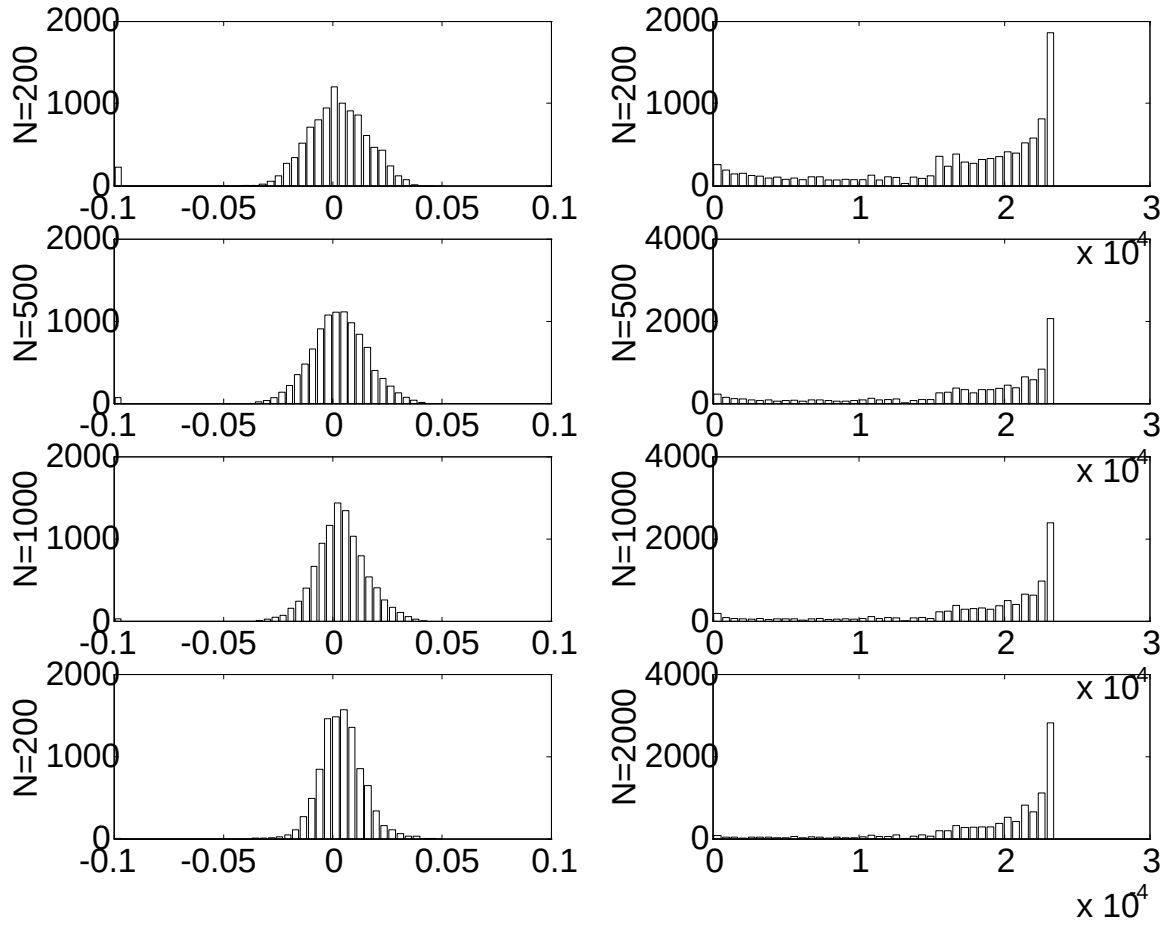

Figure 5. The left column is a histogram of the parameters estimated by the Risk Neutral Forecasting estimator and the right column is a histogram of the expected profits from these estimated parameters. Each row corresponds to a sample size equal to N.

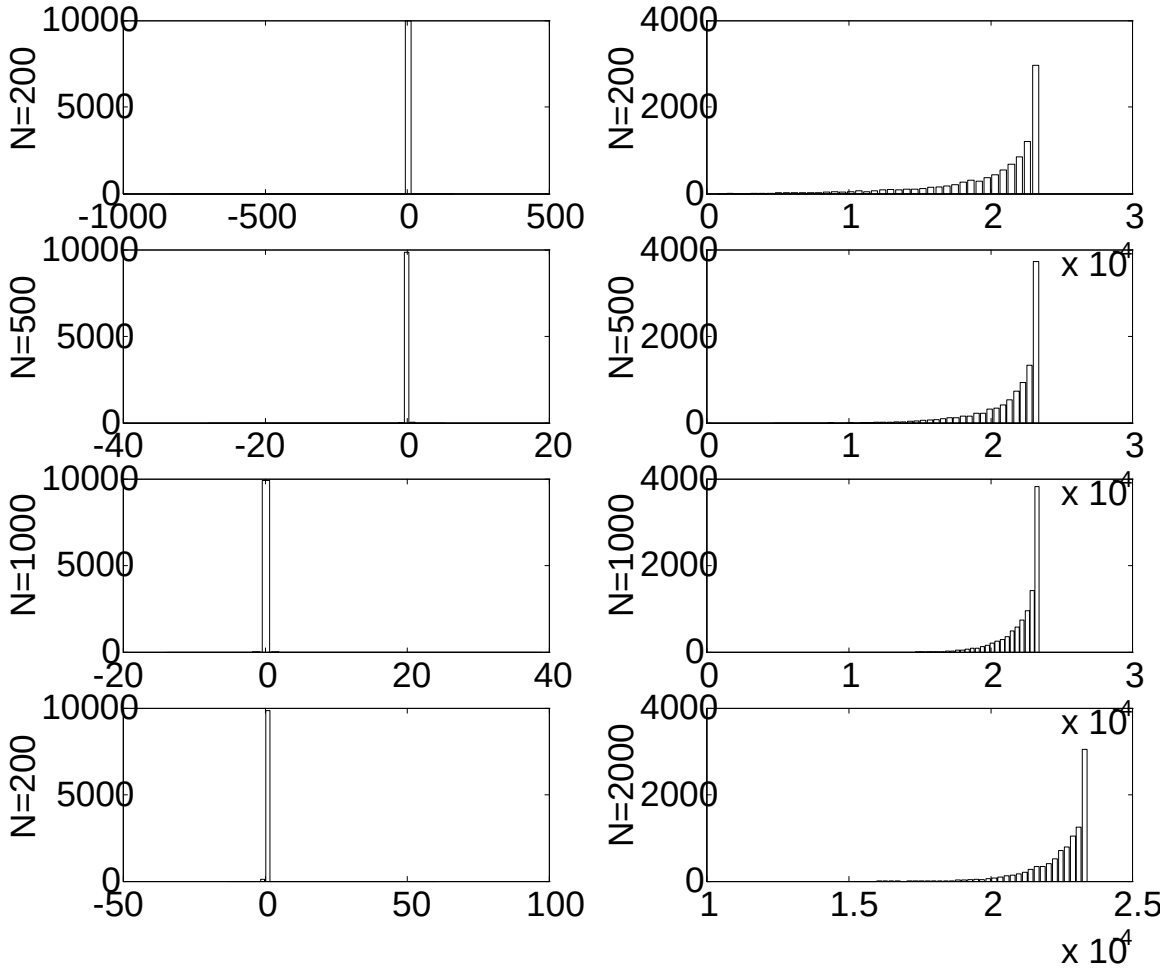

Figure 6. The left column is a histogram of the parameters estimated by OLS and the right column is a histogram of the expected profits from these estimated parameters. Each row corresponds to a sample size equal to N.

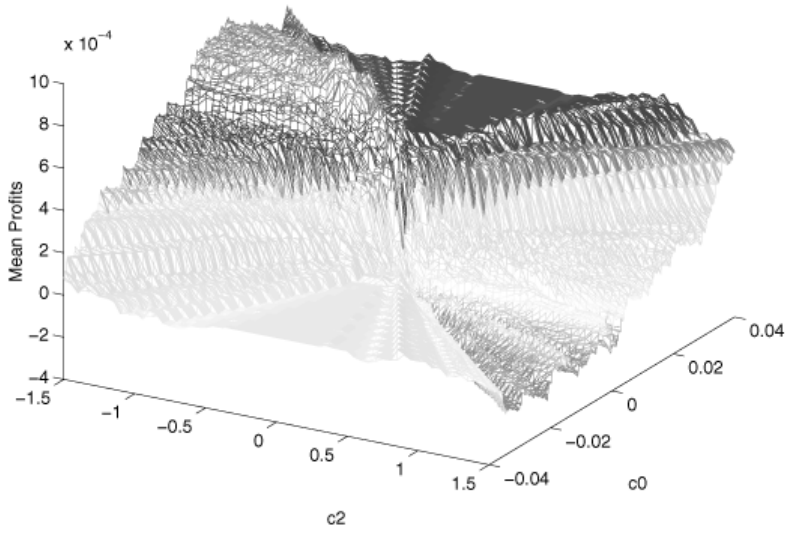

Figure 7. This figure plots profits obtained when a linear predictor with parameters (c_0, c_2) is used to forecast returns.

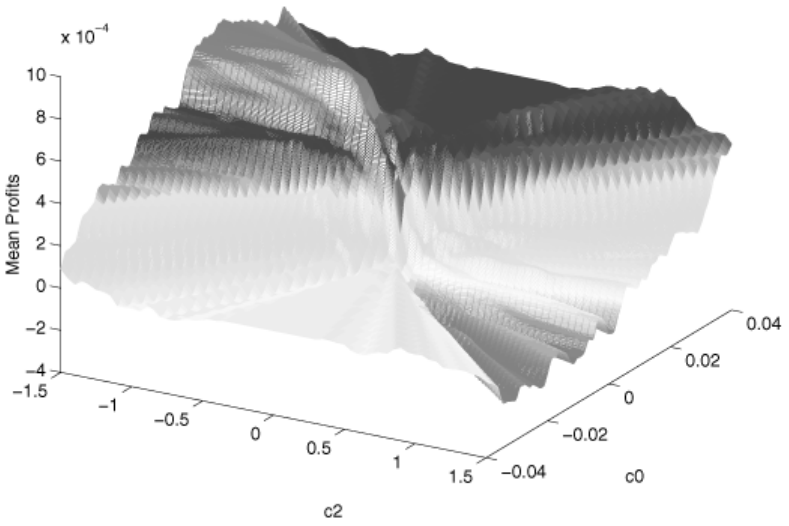

Figure 8. This figure plots an approximation to the profits obtained when a linear predictor with parameters (c_0, c_2) is used to forecast returns. The approximation is designed to make profits a smooth function of these parameters.

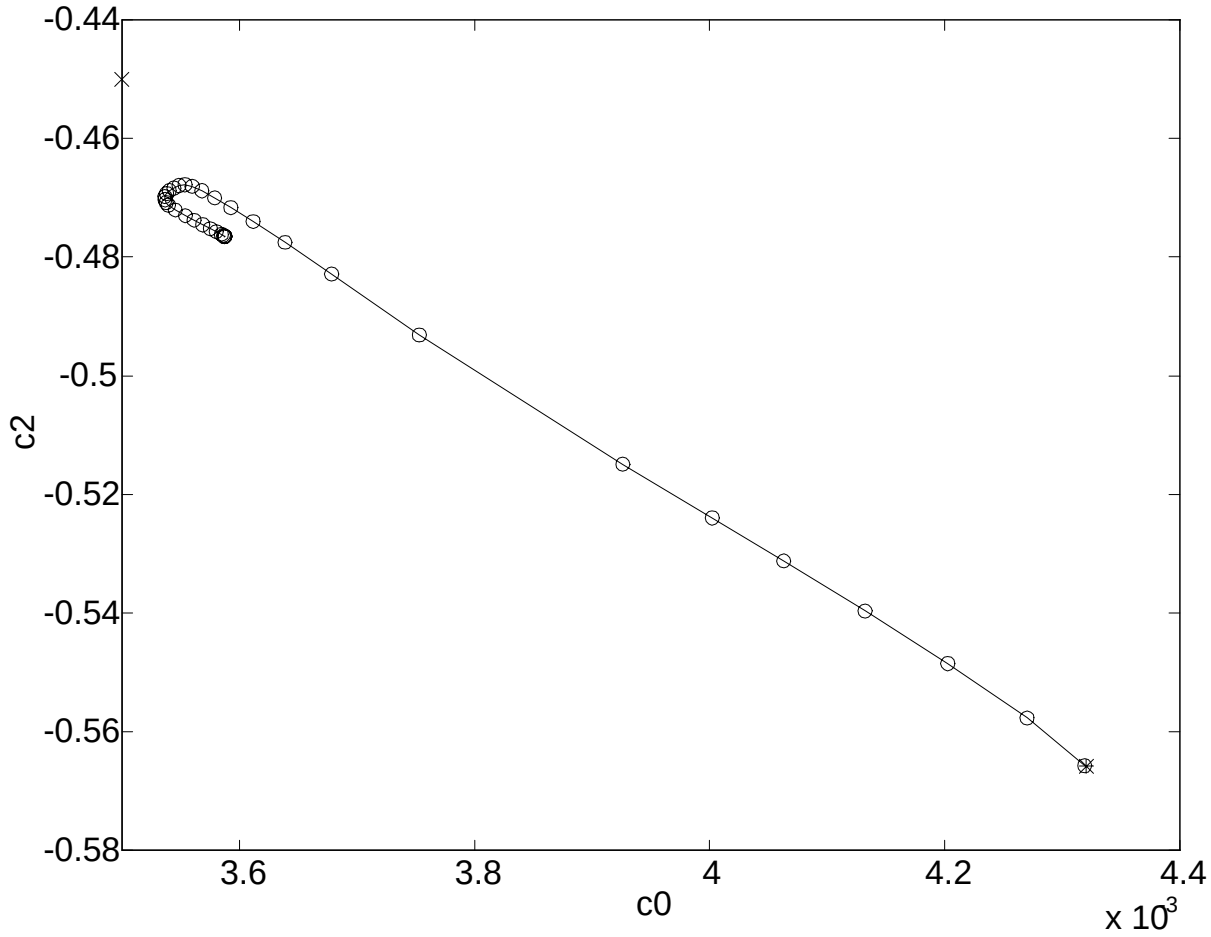

Figure 9: The circles display parameters computed at each recursion of Step 2 of the computational algorithm. The * is the starting point computed by the genetic algorith and the X is the grid-search parameter. Notice that the final parameter of the computational algorithm is a more accurate estimate of the optimum than that attainable with the grid search used.

10 References

Acar E., 1998, 'Expected returns of directional forecasters', in Acar E. and Satchell S. (eds), Advanced Trading Rules, Butterworth Heinemann.

Breen William, Lawrence R. Glosten and Ravi Jagannathan, 1989, 'Economic Significance of Predictable Variations in Stock Index Returns', Journal of Finance Vol XLIV, No. 5, 1177-1189.

Campbell J. Y., Lo A. W., MacKinlay A. C., 1997, *The Econometrics of Financial Markets*, Princeton University Press.

Campbell J.Y. and L.M. Viceira, 1996, Consumption and portfolio decisions when expected returns are time varying, NBER w.p. 5857.

Chamberlain G., 1986, 'Asymptotic Efficiency in semi-parametric models with censoring', Journal of Econometrics 32, 189-218.

Christofferson P. F., Diebold F. X., (1996), 'Further results on Forecasting and Model Selection Under Asymmetric Loss', Journal of Applied Econometrics, 11, 561-571.

Dorsey R. E. and W. J. Mayer, 1995, 'Genetic Algorithms for estimation problems with multiple optima, nondifferentiability and other irregular features', Journal of Business and Economic Statistics, 13, 1.

Fornari F., A. Mele, 1994, 'A stochastic variance model for absolute returns', Economics Letters, 46, 211-214.

Granger C.W.J., 1993, 'On the limitations of comparing mean squared forecast errors: comment', Journal of Forecasting, 12, 651-652.

Granger C. W. J., Z. Ding, 1994,a 'Stylized facts on the temporal and distributional properties of daily data from speculative markets', UCSD 93-38.

Granger C. W. J., Z. Ding, 1994b, 'Modeling volatility persistence of speculative returns: A new approach', Journal of Econometrics 73, 185-215.

Granger C. W. J., Z. Ding, 1993, 'Some properties of absolute return: an alternative measure of risk', UCSD 93-38.

Granger C.W.J. and H. Pesaran, 1996, 'A decision theoretic approach to forecast evaluation', UCSD working paper.

Henriksson and Merton R. C., 1981, 'On market timing and investment performace II: Statistical procedures for evaluating forecasting skills', Journal of Business, 54, 513-533.

Koenker R.W., Vasco d'Orey, 1985, 'Computing Regression Quantiles' U. of Illinois at Urbana-Champaign Bureau of Economic and Business Research Faculty Working Paper: 1141.

LeBaron B., 1998, 'An evolutionary bootstrap method for selecting dynamic trading strategies', U. of Wisconsin w.p.

Manski Charles F., 1988, *Analog Methods in Econometrics*, Monographs on Statistics and Applied Probability 39, Chapman and Hall.

Manski Charles F., 1988b, 'Identification of binary response models', Journal of the American Statistical Association, 83, No. 403.

Manski Charles F., 1985, 'Semiparametric analysis of discrete response', Journal of Econometrics, 313-333.

Manski C. F., T. S. Thompson, 1989, Estimation of best predictors of binary response, Journal of Econometrics 40, 97-123

Merton R. C., 1981, 'On market timing and investment performance, I: An equilibrium theory of market forecasts', Journal of Business 54, 363-406.

Mills Terence C., 1996, 'Non-linear forecasting of financial time series: An overview and some new models', Journal of Forecasting, Vol. 15, 127-135.

Moody J., Saffell M., Liao Y., Wu L., 1998, 'Reinforcement Learning for Trading Systems and Portfolios: Immediate vs. Future Rewards', mimeo, Oregon Graduate Institute of Science and Technology.

Papoulis A., 1984, *Probability, Random Variables and Stochastic Processes*, McGraw-Hill.

Pesaran M. Hashem and Alan Timmerman, 1995, 'Predictability of Stock Returns: Robustness and Economic Significance', Journal of Finance Vol.L, No. 4, 1201-1228.

Pictet O. V., Dacorogna M. M., Dave R. D., Chopard B, Schirru R., and Marco Tomassini, 1996, 'Genetic Algorithms with collective sharing for robust optimization in financial applications', Olsen Associates mimeo.

Pictet O.V., Dacorogna M.M., Muller U.A., Olsen R.B., and Ward J.R., 1992, 'Real time trading models for foreign exchange rates', Neural Network World 2(6), 713-744.

Satchell Steve and Alan Timmerman, 1995, 'An Assesment of the Economic Value of Non-linear Foreign Exchange Rate Forecasts', Journal of Forecasting, Vol. 14, 477-497.

Schwert William G., 1989, 'Why does Stock market volatility change over time?', Journal of Finance, XLIV 5.

Skouras Spyros, 1997, 'Analysing Technical Analysis', European University Institute Working Paper, 97-36.

Sullivan Ryan, Timmerman Allan and Halbert White, 1997, 'Data-snooping, technical trading rule performance and the bootstrap', UCSD working paper 97-31.

Sullivan Ryan, Timmerman Allan and Halbert White, 1998, 'Danger of data driven inference', UCSD working paper 98-16.

Tsybakov 1988, 'Passive Stochastic approximation', U. of Bonn A-207.

Tvede Lars, 1997, *Business Cycles: From John Law to Chaos Theory*, Harwood Academic Publishers.

Weiss Andrew A., 1996, 'Estimating Time Series Models Using the Relevant Cost Function', Journal of Applied Econometrics, 11, 539-560.

West K. D., 1996, 'Asymptotic inference about predictive ability', Econometrica, 64, 3, pp. 1067-1084.

West K. D., Edison H. J. and Dongchul Cho, 1993, A utility-based comparison of some models of exchange rate volatility, Jounal of International Economics 35 (1993), pp. 23-45.

White Halbert, 1984, *Asymptotic Theory for Econometricians*, Academic Press.

Zheng.X., 1998, 'A consistent nonparametric test of parametric regression models under conditional quantile restrictions', Econometric Theory, 14: 123-38.