Entropy and Predictability of Stock Market Returns*

Esfandiar Maasoumi
Department of Economics
Southern Methodist University
Dallas, TX 75275-0496, USA
maasoumi@mail.smu.edu

Jeff Racine
Department of Economics
University of South Florida
Tampa, FL 33620, USA
jracine@coba.usf.edu

First version, July 2000
This version, December 2000

Abstract

We examine the predictability of stock market returns by employing a new metric entropy measure of dependence with several desirable properties. We compare our results with a number of traditional measures. The metric entropy is capable of detecting nonlinear dependence within the returns series, and is also capable of detecting nonlinear"affinity" between the returns and their predictions obtained from various models thereby serving as a measure of out-of-sample goodness-of-fit or model adequacy. Several models are investigated, including the linear and neural-network models as well as nonparametric and recursive unconditional mean models. We find significant evidence of small nonlinear *unconditional* serial dependence within the returns series, but fragile evidence of superior conditional predictability (profit opportunity) when using market-switching versus buy-and-hold strategies.

Keywords: Entropy, stock returns, nonparametric, neural-networks, prediction, dependence, nonlinear.

JEL Classification: C14 - Semiparametric and Nonparametric Methods.

*We would like to thank but not implicate Min Qi for her many useful comments and for providing both her data and the recursive residuals from the neural-network model used in Qi (1999). Special thanks to Amos Golan and two referees for extensive comments and suggestions.

1 Introduction

Much of the theoretical literature in finance is based on market efficiency arguments which imply unpredictability of returns (or, no "profit opportunity"). The empirical evidence, however, is mixed. In the existing literature the most common paradigm is a linear-in-mean model with searches for variables that may provide significant explanation for the returns (see, for example, Pesaran & Timmermann (1995)). Alternatively, given a set of variables one searches empirically for suitable functional forms (see, for example, Qi (1999)). The linear model evidence is somewhat inconclusive. A number of authors have pointed out that these inferences may be in reality an artifact of the linear filters, and have presented results which indicate the presence of nonlinear dependence in returns and other financial series. This is a vast literature; for some examples see Qi (1999) who considers both linear models and feed-forward neural-network models of monthly excess returns, Abhyankar, Copeland & Wong (1997), Campbell, Lo & MacKinlay (1997), and the references therein.

Given that findings such as those of Qi (1999) are based on linear parametric models and neural-network implementations, they are in effect concerned with the joint questions of predictability and functional specification. This latter issue raises the possibility that these findings may be contaminated by model misspecification. This is particularly worrisome since these findings may be further confounded by the failings of the currently dominant, correlation-type measures of fit, predictability, and dependence. The evidence on these failings is quite widespread (for example, see Granger, Maasoumi & Racine (2000), Skaug & Tjøstheim (1996), Hsieh (1989), Scheinkman & LeBaron (1989) and Liu & Stengos (1999) with their references in the area of "growth convergence"). Noting these difficulties, many authors have explored nonlinear, nonparametric, and semiparametric approaches. For instance White (1988), Stengos (1995), and Frank & Stengos (1988), *inter alia*, considered nonlinear models (including neural nets) and nonparametric regressions for returns on cer-

tain equities and precious metals as a means of accommodating generic nonlinearity in the conditional mean of returns.

We have not investigated GARCH models which may be viewed as attempts to capture some nonlinearities. Hong & White (2000) removed persistent GARCH effects from the S&P series but came to similar conclusions as ours. In other financial applications (see Hsieh (1989) for exchange rates) the presence of unaccounted nonlinear dependence in the residuals of such models has been detected. Pagan & Schwert (1990) studied the performance of parametric, nonparametric, as well as semiparametric models of conditional variances in the GARCH setting for monthly U.S. stock returns (1834-1925). Based on traditional criteria, they found parametric variants did "better" out of sample, with exponential GARCH being the overall best. But the better *in-sample* performance of their nonparametric and semiparametric techniques led them to advocate combining the two approaches. Given the sensitivity of parametric GARCH to misspecification of the mean and variances, its nonparametric implementations are worthy of further research but lie beyond the scope of our paper.

Notably, in most studies the burden of dealing with nonlinearities is placed on the successful modeling of the *conditional mean and conditional variance* (see Hong & White (2000) for an exception). A common conclusion from these studies is that the observed nonlinearities may be too complex to be exploitable for improved "predictability", especially with small samples (see Stengos (1995) and Hsieh (1989) for examples).

It is desirable to disentangle some of these issues. For instance, one can make an inquiry about any unconditional, or conditional (and possibly nonlinear) dependence structure in returns without requiring the specification of conditional mean-variance models. Here, nonparametric density and other functional estimation receives further attention in our work.

And, one may begin to question the traditional *concepts* as well as *measures* of "predictability". The latter are generally limited to moment-based measures of prediction error

in the conditional mean. Indeed, the *concept* of the conditional mean itself is an optimum predictor under criteria that may be inadequate in non-Gaussian, nonlinear, and non-symmetric circumstances found in financial processes. In this connection entropy measures offer opportunities which are explored here. Entropies are defined directly in terms of the actual distributions and not the variables and their moments. Partly for this reason they also offer a clearer view of the relation between total independence, conditional dependence, and causality relations in several directions. For our purposes, it suffices to note that enquiring about unconditional independence in the return series is logically prior to questions of conditional dependence/predictability. Gourieroux, Monfort & Renault (1986) provide an extensive account based on the Kullback information criterion and its related "causality measures". They obtain Geweke's causality tests and decompositions as a special case (see Geweke (1982)).

Accordingly, we propose to study excess returns for unconditional, nonparametric dependence using a normalization of the Hellinger-Bhattacharya-Matusita *metric entropy* measure which we have found to be successful in detecting generic and possibly nonlinear dependence (see Granger et al. (2000)). Being entropy-based, this measure is defined over the densities of the stock returns which we estimate nonparametrically. For a range of popular models the detection capability of this entropy has been found to be at least as good as the BDS test, and better than the traditional correlation and other moments measures, especially in nonlinear settings (see Skaug & Tjøstheim (1993), Skaug & Tjøstheim (1996), and Granger et al. (2000)). Other entropies, especially the Kullback-Leibler (KL) measure, are seeing increasing and welcome use in testing for independence and other hypotheses. For instance, Robinson (1991), Delgado (1994), Hong & White (2000), and Zheng (2000) are all concerned with the KL measure for testing independence. Most entropies, however, are non-metric since they violate the *triangularity rule* and *symmetry* required of *distance* measures (see Maasoumi (1993) for a synthesis, and Hirschberg, Maasoumi & Slottje (2000) for a cluster analysis

example wherein "divergence" needs to be distinguished from "distance").

We find a small but significant degree of nonlinear serial dependence in stock market returns *which is not model dependent*. We then propose and use the same entropy measure to investigate the predictive performance of linear, neural-network, and a nonparametric regression of stock returns all of which employ the same set of popular "predictive" variables used in Qi (1999) and Pesaran & Timmermann (1995), along with predictions generated simply using the *unconditional mean* of past returns. We weigh the evidence on the predictive performance of various models based on the recursive predictions generated by each model. This last use of our entropy as a goodness-of-fit, or predictability measure is similar to the use of Kullback-Leibler and relative entropy measures of fit as in, for instance, Joe (1989) and Cameron & Windmeijer (1997). Since potential "profit" is an additional and popular measure of "success" in this literature, we merely report it for a number of time periods.

Based on the results here and elsewhere, we infer that the evidence in favor of conditional predictability of stock returns is non-robust with respect to such things as period of analysis, data frequency, variable and functional form choices, as well as the predictability criteria. In particular, we find that the buy-and-hold strategy generated by using the *unconditional mean* of past returns can outperform the market-switching strategies generated by the linear, neural-network, and nonparametric models. The small unconditional serial dependence in the returns themselves may be too small, or too complex, for robust *conditional mean and variance* analysis. Similar conclusions were drawn by Diebold & Nason (1990) and Stengos (1995), and by Hong & White (2000) even after they removed persistent GARCH effects from the S&P series.

Section 2 defines the data to be studied along with an overview of our entropy measure, and then uses the entropy measure to detect serial dependence in the returns series itself. Section 3 introduces potential models for predicting excess returns and examines the relative performance of these various specifications. Section 3.1 considers the use of the entropy

measure for assessing the adequacy of the various conditional mean models based on the actual out-of-sample excess returns and the out-of-sample predictions generated by these models, and Section 3.2 offers further discussion. Section 3.3 reports the profits generated by these various models, and Section 4 concludes.

2 Assessing Dependence in Excess Returns

We use data found in the recent study by Qi (1999). These were also used in Pesaran & Timmermann (1995) and, in addition to being used in a number of other published studies, are publicly available. We briefly describe the data for the interested reader.

Following standard practice, we consider *monthly* excess returns on the S&P 500 index at time t defined as the capital gain, plus dividend yield, minus the one-month Treasury-bill rate:

$$r_t = \left[\frac{(P_t - P_{t-1}) + D_t}{P_{t-1}}\right] - I1_{t-1},, \tag{1}$$

where P_t is the stock price, D_t dividends, and $I1_{t-1}$ the return from holding a one-month Treasury bill from the end of month $t-1$ through the end of month t. The predictive variables are the unscaled dividend yield, the earnings-price ratio, the 1-month T-bill rate (1 & 2 lags), the 12-month T-bill rate (1 & 2 lags), the year-on-year inflation rate, the year-on-year rate of change in industrial output, and the year-on-year growth rate in the narrow money stock.

Qi (1999) applied a number of existing tests for nonlinear dependence to this series for the period 1954:1-1992:12[1], and results were somewhat mixed. Qi's tests were performed on the residuals of a prewhitening autoregressive model with order chosen by AIC or BIC (see also Abhyankar et al. (1997)). Interestingly, the BDS test of Brock, Deckert, Scheinkman

[1]We are indebted to Min Qi for allowing us full access to her data and for verifying that our Matlab code indeed should replicate the model she estimated.

& LeBaron (1996) was not significant (Qi (1999, page 422)), while other tests such as the Ljung & Box (1978) test were. However, limitations of a number of competing tests have been noted such as the significance of the Ljung-Box test potentially being due to ARCH effects (Diebold & Nason (1990)). Overall, Qi concludes that her results also "confirm the evidence of nonlinearity in returns on the S&P 500 index".

It would be useful to consider measures of dependence and association defined in the space of distribution functions. Granger & Maasoumi (1993) considered a normalization of the Matusita-Bhattacharya-Hellinger measure of dependence given by

$$S_\rho = \frac{1}{2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(f^{\frac{1}{2}} - f_1^{\frac{1}{2}} f_2^{\frac{1}{2}}\right)^2 dx\, dy, \tag{2}$$

where $f = f(x, y)$ is the joint density and $f_1 = f(x)$ and $f_2 = f(y)$ are the marginal densities of the random variables X and Y. If X and Y are independent, this metric will yield the value zero, and is otherwise positive and less than one. To see the relation of this normalized measure to entropy *divergence* measures, consider the k-class entropy family of Havrda & Charvat (1967) cited and discussed in Maasoumi (1993):

$$\begin{aligned} H_k(f) &= (k-1)^{-1}(1 - Ef^{k-1}), \ldots k \neq 1 \\ &= -E\log f, \quad (\textit{Shannon's entropy}) \quad \textit{for } k = 1, \end{aligned}$$

where E denotes expectation with respect to the distribution f. The k-class entropies satisfy axioms A1-A5 of Maasoumi (1993), and are similar to Renyi's entropy family which satisfy axioms A1-A4 as well as an "arithmetic mean value" property. Essentially they differ only slightly in their branching (aggregation) properties. For two density functions f_1 and f_2, the asymmetric (with respect to f_2) k-class entropy *divergence measure* is:

$$I_k(f_2, f_1) = \frac{1}{k-1}\left[\int (f_1^k / f_2^k) dF_2 - 1\right], \quad k \neq 1,$$

such that $\lim_{k\to 1} I_k(\cdot) = I_1(\cdot)$, the Shannon cross entropy (divergence) measure. Once the divergence in both directions of f_1 *and* f_2 are added, a symmetric measure is obtained which, for $k = 1$, is well known as the Kullback-Leibler measure. The symmetric $k - class$ measure at $k = \frac{1}{2}$ is of interest to us in this paper, and is given as follows:

$$I_{\frac{1}{2}} = I_{\frac{1}{2}}(f_2, f_1) + I_{\frac{1}{2}}(f_1, f_2) = 2M(f_1, f_2) = 4B(f_1, f_2)$$

$$where\ M(\cdot) = \int (f_1^{\frac{1}{2}} - f_2^{\frac{1}{2}})^2 dx \text{ , is known as the Matusita } distance\text{, and,}$$

$$B(\cdot) = 1 - \rho^* \text{ , is known as the Bhattacharya } distance \text{ with}$$

$$\rho^* = \int (f_1 f_2)^{\frac{1}{2}} \text{ being a measure of "affinity" between the two densities.}$$

$B(\cdot)$ and $M(\cdot)$ are rather unique among measures of divergence since they satisfy the triangular (distance) inequality and are, therefore, *metric.* While the other divergence measures are also quite capable of characterizing desired null hypotheses (such as independence) they are not appropriate when these distances are compared across models, sample periods, or agents. These comparisons are often made, and more often implicit in inferences.

Other than axioms A1-A5 alluded to above, the following useful properties are satisfied by this entropy measure (see Granger et al. (2000)):

1. It is well defined for both continuous and discrete variables.

2. It is normalized to zero if X and Y are independent, and lies between 0 and +1.

3. The modulus of the measure is equal to unity if there is a *measurable* exact (nonlinear) relationship, $Y = g(X)$ say, between the random variables.

4. It is equal to or has a simple relationship with the (linear) correlation coefficient in the case of a bivariate normal distribution.

5. It is metric, that is, it is a true measure of "distance" and not just of divergence.

6. The measure is invariant under continuous and strictly increasing transformations $h(\cdot)$. This is useful since X and Y are independent if and only if $h(X)$ and $h(Y)$ are independent. Invariance is important since otherwise clever or inadvertent transformations would produce different levels of dependence.

Note that,

$$S_\rho = 1 - \rho^*, \quad \text{where} \quad 0 \leq \rho^* \leq \int\int (f \times f_1 f_2)^{\frac{1}{2}} \, dx \, dy \leq 1.$$

Properties 1-3 are easily verified. Property 5 is established in the literature (see Maasoumi (1993)), and Property 6 was established by Skaug & Tjøstheim (1993). To verify Property 4, let $f(x,y) = N(0,0,1,1,\rho)$ and $f(x) = N(0,1) = f(y)$, then:

$$\begin{aligned} \rho^* &= \frac{(1-\rho^2)^{\frac{5}{4}}}{(1-\frac{\rho^2}{2})^{\frac{3}{2}}} = 1 \quad \text{if } \rho = 0 \\ &= 0 \quad \text{if } \rho = 1. \end{aligned}$$

As in Maasoumi (1993), the axiomatic characterizations that uniquely identify the k-class entropies, or divergence measures, reveal the extent to which our measure is unique for $k = 1/2$. We make no further claims. In particular, various normalizations are possible and must be justified within the application context. Ours is justified by a desire for analogy to the linear correlation coefficient. Also, as was commented by Hong & White (2000), it seems possible to obtain a more general proof of asymptotic Gaussianity of our measure than is given by Skaug & Tjøstheim (1993). Gaussianity, however, appears to be generally a poor approximation for this statistic. A final appealing property is worthy of mention. Entropy measures are *dimensionless* since they are defined over distribution functions. In multivariate settings, competing moment-based measures must be summarized by arbitrary scalar functions (such as trace or determinant of variance matrices).

Granger et al. (2000) consider a kernel implementation of this metric and demonstrate how critical values can be obtained under the null of serial independence in the case of time-series data. A few words on the computation of this metric are in order. For the estimation of the univariate and bivariate densities in S_ρ we use kernel density estimators. For the kernel function we employ the widely used second-order Gaussian kernel, while bandwidths are selected via likelihood cross-validation (Silverman (1986, page 52)). To compute the multivariate integral we employ numerical quadrature using the 'tricub()' algorithm of Lau (1995, pg 303).

We apply the metric S_ρ to the excess returns data in Equation (1) by setting $x = r_t$ and $y = r_{t-k}$ in Equation (2). Note that this test is applied directly to the series rather than indirectly via a prewhitening AR model as in Qi (1999), and is therefore not subject to pitfalls surrounding inappropriate specification of the prewhitening filters. The results are summarized in Figure 1 which depicts the value of the metric itself and the 90th percentile of its distribution under the null of independence.

Examining Figure 1 we see that there exists a small but significant nonlinear dependence present in this series at lag $K = 1$, but not thereafter. We applied the test for serial independence found in Granger et al. (2000) for lag $K = 1$ and obtained an empirical p-value of 0.001. This too suggests the presence of a small but highly significant, nonlinear serial dependence, and is consistent with the overall findings of Qi (1999) and others.

Having demonstrated how a metric entropy is capable of detecting the presence of significant nonlinear dependence for univariate time-series data, we now demonstrate how a metric entropy can be useful for assessing out-of-sample forecasting performance of several popular models.

3 Predicting Excess Returns Using Economic Variables

Pagan & Schwert (1990) considered the forecasting performance of both parametric and nonparametric GARCH models for monthly stock returns. On the basis of their criteria they found frequent disagreement between the in-sample and out-of-sample results. Qi's (1999) study focused primarily on prediction and considered two models for recursive prediction of the monthly excess returns series, a linear regression model (LR) and a single hidden-layer, eight hidden-unit feed-forward neural-network (NN) trained via Bayesian Regularization. We use the same data and the LR model as a benchmark and focus on conditional *recursive prediction* of excess returns. Similarly, we generate recursive predictions for the entire period 1960:1-1992:12 and focus on forecasting performance during this period, as well as the three decades 1960:1-1969:12, 1970:1-1979:12, and 1980:1-1989:12.

In addition to the neural-network (NN) and linear regression (LR) models used in Qi (1999), we also consider the unconditional mean of past excess returns (MN), and a non-parametric kernel regression model (NP) having the same variables as the NN and LR models. For the NP model, bandwidths were selected via leave-one-out cross-validation for each sample upon which the recursive predictions were made. The models are therefore as follows:

$$
\begin{aligned}
\text{LR:} \quad & \hat{r}_{t+1} = g(x_t, \hat{\beta}_t) = \hat{\beta}_{0t} + \sum_{i=1}^{k} \hat{\beta}_{it} x_{it}, \\
\text{NN:} \quad & \hat{r}_{t+1} = g(x_t, \hat{\alpha}_t, \hat{\beta}_t) = \hat{\alpha}_{0t} + \sum_{j=1}^{n} \hat{\alpha}_{jt} \, \text{logsig} \left(\sum_{i=1}^{k} \hat{\beta}_{ijt} x_{it} + \hat{\beta}_{0jt} \right), \\
\text{MN:} \quad & \hat{r}_{t+1} = g(r_t, t) = t^{-1} \sum_{i=1}^{t} r_i, \\
\text{NP:} \quad & \hat{r}_{t+1} = g(x_t, \hat{h}_t) = \frac{\sum_{i=1}^{t} r_i K \left(\frac{x_{1i} - x_{1t}}{\hat{h}_{1t}}, \ldots, \frac{x_{ki} - x_{kt}}{\hat{h}_{kt}} \right)}{\sum_{i=1}^{t} K \left(\frac{x_{1i} - x_{1t}}{\hat{h}_{1t}}, \ldots, \frac{x_{ki} - x_{kt}}{\hat{h}_{kt}} \right)},
\end{aligned}
\tag{3}
$$

where $x_t' = (x_{1t}, \ldots, x_{kt})$ are the k (nine) financial and economic variables, r_t is the excess returns on the S&P 500 index at time t, and n is the number of hidden units used in the NN model.

We found that no NN model could come close to generating the level of *accumulated wealth* reported in Qi (1999) even though our NN models performed similarly in terms of their statistical in and out-of-sample performance. We therefore report both the results in Qi (1999) and our otherwise identical NN model estimated using 10 re-starts of the minimization algorithm from different random values. We believe that the replicated predictions and profits reported below more accurately reflect the performance of the NN model (see Racine (2001) for further discussion).

Figure 2 presents the predicted versus actual values for each model for the entire forecast period 1960:1-1992:12 where the predicted values are the one-step-ahead (out-of-sample) recursive forecasts generated by each model as described above. As is seen from the predictions generated by the MN plotted in Figure 2, this model always predicts that excess returns will be positive and leads to a 'buy-and-hold' strategy where the investor remains fully invested in stocks and purchases more when her budget permits. The other models, on the other hand, generate the 'market-switching' behavior described below in Section 3.3.

3.1 Assessing Model Adequacy

To assess models many have used the following traditional and moment-based measures defined as follows:

$$\begin{aligned}
\text{RMSE} &= \sqrt{T^{-1}\sum_{t=1}^{T}(r_t - \hat{r}_t)^2}, \\
\text{MAE} &= T^{-1}\sum_{t=1}^{T}|r_t - \hat{r}_t|, \\
\text{MAPE} &= T^{-1}\sum_{t=1}^{T}\left|\frac{r_t - \hat{r}_t}{r_t}\right|, \\
\text{CORR} &= \frac{\sum_{t=1}^{T}(r_t - \bar{r}_t)(\hat{r}_t - \bar{\hat{r}}_t)}{\sqrt{\sum_{t=1}^{T}(r_t - \bar{r}_t)^2}\sqrt{\sum_{t=1}^{T}(\hat{r}_t - \bar{\hat{r}}_t)^2}}, \\
\text{SIGN} &= T^{-1}\sum_{t=1}^{T} z_t,
\end{aligned} \tag{4}$$

where $z_t = 1$ if $r_{t+1} \times \hat{r}_{t+1} > 0$ and 0 otherwise. The first four are L_2 and L_1-norm summary measures of out-of-sample prediction error while the last is the proportion of times the sign of excess returns is correctly predicted. We note that SIGN is unusual, not continuous, and does not satisfy the conditions of Property 3 discussed earlier. Therefore, it does not have to agree with entropy nor any of the other measures.

This is a suitable setting in which to examine the S_ρ statistic as a measure of goodness-of-fit and as a test of predictive performance where we set $x = r_t$ and $y = \hat{r}_t$ in Equation (2), $\hat{r}_t$ being the recursive prediction generated by a model. An adequate model would be expected to produce a strong relationship between the actual returns and their predicted values. In particular, the value of the entropy metric for the nonparametric regression would be free of functional form misspecification (but not variable selection). We compute the entropy measure given in Equation (2), with $f = f(\hat{r}_t, r_t)$ as the joint density of the predicted and

actual excess returns, and $f_1 = f(\hat{r}_t)$ and $f_2 = f(r_t)$ as the respective marginal densities. If predicted and actual returns are independent, this metric will yield the value zero, and will increase as the model's predictive ability improves.

There is a complicated relationship between entropies, generally, and moment-based measures of a distribution. Ebrahimi, Maasoumi & Soofi (1999) explored this issue for Shannon's entropy. Based on a Legendre series expansion of entropy, they showed that it is a function of many moments of the underlying probability density function (PDF), much like a moment generating function. In view of the definition of entropy as a relatively unique function of the PDF, this result is quite revealing (if not surprising). Ebrahimi et al. (1999) examined ranking of distributions by variance and Shannon's entropy. Their findings are relevant to the comparison of our entropy measure with moment-based measures of forecasting and fit, but a full discussion is beyond the scope of this paper. We merely summarize three useful results. First, the class of distributions that can be characterized by a single moment (e.g., variance) is very large. Put simply, single moment measures are inadequate for ranking PDFs that are likely to be appropriate for financial processes. Secondly, nonlinear transformations of variables tend to increase "uncertainty" defined broadly by entropy or otherwise. For most relevant PDFs, single moment measures are not likely to reflect this increased uncertainty (unpredictability) as fully as entropy can. This does not bode well for the traditional measures of forecastability for nonlinear, non-Gaussian, and asymmetric processes. Thirdly, in univariate settings, when characterization of a PDF by variance is adequate (as for the Normal PDF), the entropy characterization is in agreement. This is important since it indicates limited risks associated with the use of the more general entropy measures even when "variance" measures may suffice. Of course, there is no general agreement on a scalar function of variance in multivariate cases (e.g., trace or determinant of the variance matrix), whereas *entropy is a dimensionless measure*. Since one does not know the "true" PDF in empirical settings, one may be inclined to regard the entropy guidance on predictability and

fit as controlling. We note that SIGN is very different from measures that characterize statistical distribution functions. Also, CORR is maximized by linear filters whether they are correct or not.

As was noted in Granger et al. (2000) and Skaug & Tjøstheim (1996), the asymptotic normal distribution of S_ρ is unreliable for practical inference. We therefore compute p-values by resampling the statistic under the null of independence to detect significant deviation from zero. We generate replications which are serially independent having marginal distributions identical to the original data simply by applying a random shuffle to the dataset. Randomly reordering the data leaves the marginal distributions intact while generating an independent bivariate distribution. This reshuffle is used to recompute the statistic using data generated under the null, and this can be repeated a large number of times to generate the empirical distribution of the statistic under the null. One could then use the empirical distribution of this resampled statistic to compute finite-sample critical values. The null distribution will be that for a given bandwidth and will therefore adapt to bandwidth choice as in Racine (1997).

There is another difficulty with the existing results in the literature. Whereas we can and do compute p-values for our entropy measure, results reported in previous studies are only the numerical estimates of the traditional fit criteria. Statistical significance can not be inferred. As is appreciated, exact sampling distributions of statistics such as the simple correlation-coefficient have been elusive. Following Efron & Tibshirani (1993, page 49) in which they bootstrap the sampling distribution of CORR, we choose to bootstrap the empirical distributions of RMSE, MAE, MAPE, CORR, and SIGN under the null of independence. We then can compute p -values under the null of independence to detect significant deviation from that which would occur if the actual and predicted values were in fact independent.

Each of the models appearing in Equation (3) along with the replicated NN model were used to obtain recursive forecasts. Beginning with the data for 1954:1-1959:12, each model

was estimated and a forecast was made for 1960:1. Next, each model was estimated for 1954:1-1960:1 and a forecast was made for 1960:2 and so on. Following Qi (1999), we consider results for the entire period 1960:1-1992:12 and for the decade sub-periods 1960:1-1969:12, 1970:1-1979:12, and 1980:1-1989:12. We then computed S_ρ and the measures in Equation (4) for the realized series r_t and its out-of-sample forecast, $\hat{r}_t$. The results are presented in Table 1.

3.2 Discussion

Consider the out-of-sample performance reported in Table 1, and the plots in Figure 3. An examination of the plots of the predicted versus the actual values is suggestive of a lack of robustness in the predictive ability of all the conditional mean models considered, *including nonparametric regression*. Models having good predictive ability may be expected to have a plot with tight scattering around the diagonal, passing through the origin and with a slope of 1.

Turning now to the traditional measures found in Table 1, the values marked with asterisks are significant in that they differ from that which would occur if the predicted and actual returns were independent. It is clear that, except for the decade of the 70s, there is no model that performs well by any of the criteria. And for the 70s, only the Linear Regression and the NN models show some significant promise by CORR (which is maximized for LR), by SIGN and the entropy measures (the empirical p-values are less than 0.01). The findings are generally confirmed in the first panel of Table 1 for the entire sample period. RMSE, MAE, and MAPE detect no significant predictability for any model and in any period. Interestingly, the nonparametric regression model does not perform well at all. We remind the reader that we employ cross-validation for bandwidth selection. Naturally, different bandwidths will be selected through cross validation for the joint and marginal distribution of

Table 1: Out-of-sample forecast performance of alternative models for various periods using correlation and entropy-based measures. Entries marked with *** have empirical p-values < 0.01, ** $0.01 \leq p < 0.05$, and * $0.05 \leq p < 0.10$ under the null of independence of actual and predicted excess returns.

Model	RMSE	MAE	MAPE	CORR	SIGN	S_ρ
			Panel A: 1960:1-1992:12			
LR	0.0430	0.0329	2.1095	0.2081^{***}	0.5960^{***}	0.064
NN (Qi)	0.0429^{*}	0.0328^{**}	1.8289	0.2292^{***}	0.6237^{***}	0.065^{***}
NN (Rep)	0.0448	0.0337^{*}	2.1334	0.1527^{***}	0.6035^{***}	0.072
NP	0.0482	0.0353	1.7803	0.0154	0.5516	0.083
MN	0.0435	0.0328	1.5102	-0.0426	0.5581	0.123
			Panel B: 1960:1-1969:12			
LR	0.0361	0.0283	1.3726	0.0656	0.5667	0.099^{***}
NN (Qi)	0.0352	0.0274	1.1150	0.0661	0.6083	0.092^{**}
NN (Rep)	0.0351	0.0274	1.1202	0.0969	0.6000^{*}	0.088
NP	0.0387	0.0296	1.6515	-0.0061	0.5750	0.087
MN	0.0358	0.0276	1.5734	-0.0392	0.5917	0.097
			Panel C: 1970:1-1979:12			
LR	0.0451	0.0349	3.5492	0.2458^{***}	0.6250^{***}	0.088^{***}
NN (Qi)	0.0444	0.0345	2.5604	0.3067^{***}	0.6750^{***}	0.103
NN (Rep)	0.0485	0.0369	3.6540	0.1239	0.6250^{***}	0.081^{*}
NP	0.0515	0.0371	2.3849	0.1151	0.5667^{*}	0.093
MN	0.0466	0.0359	1.9231	-0.1062	0.4833	0.115^{*}
			Panel D: 1980:1-1989:12			
LR	0.0476	0.0359	1.3894	0.2287^{**}	0.5750	0.093
NN (Qi)	0.0487	0.0368	1.7134	0.2070^{**}	0.5750	0.084
NN (Rep)	0.0505	0.0375	1.6292	0.1742^{*}	0.5667	0.101
NP	0.0549	0.0406	1.4663	-0.1113	0.4917	0.126
MN	0.0477	0.0353	1.0458	-0.1251	0.5917	0.087

the predictions given by different models for different periods. In finite-samples, then, some of the variation in the statistic S_ρ is due to bandwidth differences.

It is not meaningful to compare how these criteria rank models when their estimates are not significant. Taken together these results suggest that these models possess little or no predictive ability. Given the plots in Figure 3, these findings support the conclusion of inadequacy in these models.

The specific instances of disagreement between the SIGN criterion and others, including the entropy, are not surprising *per se* in light of our earlier discussion. Entropy is a function of many moments of a distribution. When higher order moments are important for characterizing distributions, many of the traditional criteria fail to be adequate. To shed more light on this topic, in Appendix B we report a small Monte Carlo study of the sampling properties of the traditional criteria for *linear* models which are generally favorable to them. For these models the traditional criteria perform well and can detect dependence and independence, especially asymptotically. Our entropy measure, however, can detect dependence in nonlinear and nonstationary cases as well when the traditional measures may fail, as shown by Granger et al. (2000). These results suggest to us that the agreement between the entropy measure and the traditional measures observed in this study is evidence against the "mean models" and their conditioning variables, and/or predictability of the stock markets. This conclusion is in broad agreement with the evidence based on GARCH models (see Pagan & Schwert (1990) and Campbell et al. (1997)). Hong & White (2000) employed the Kullback-Leibler measure, removed GARCH effects, and still concluded that the daily S&P returns are white noise but with strong nonlinear dependence.

3.3 Comparing Profitability for Different Models and Time Horizons

In this paper our aim is to examine the link between entropy and predictability. In the literature predictability has also been associated with profitability. We also look at prof-

itability but do not intend to suggest any necessary association - indeed if one is interested in modeling profitability then this should directly enter the objective function in our opinion.

We consider profits generated by a *switching portfolio* based upon predictions from each of these models assuming that $100 was invested in either stocks or bonds as detailed in Pesaran & Timmermann (1995), and used by Qi (1999) who considered three levels of transactions costs, zero, low and high. We consider only the low/high cases following Qi (1999) who defined low transactions costs to be 0.5% on trading in stocks and 0.1% on trading in bonds, while high costs were defined to be 1% on stocks and 0.1% in bonds. Briefly, the investor generates a recursive prediction of next period's excess returns and predicts either positive or negative excess returns. Given this prediction, she makes a decision of whether to buy stocks or bonds giving rise to four possible decisions: for positive predictions, if currently holding stock, buy more stock if the budget permits, while if currently holding bonds, switch to stock; for negative predictions, if currently holding stock switch to bonds, while if currently holding bonds, buy more bonds if the budget permits. Below we report those results for the low transactions cost scenario, while Appendix A reports on the high cost scenario. We follow Pesaran & Timmermann (1995) and Qi (1999) who only considered the case of holding either stock or bonds, but not both simultaneously. Tables 2 through 5 present the profits and ranks of each model for each period considered.

Table 2: Final wealth for the entire period (Panel A: 1960:1-1992:12): low transactions costs.

Final Wealth and Rank of $100.00 Invested For Various Models: Entire Period										
Period	LR	Rank	NN (Qi)	Rank	NN (Rep)	Rank	NP	Rank	MN	Rank
1	$4631.13	(2)	$8420.43	(1)	$4204.96	(3)	$1148.36	(5)	$2432.78	(4)

Qi (1999) found that her neural-network model generated higher profits with lower risks than did the linear regression when the recursive predictions were used to form a switching

Table 3: Final wealth for ten year periods (Panel B: 1960:1-1969:12, Panel C: 1970:1-1979:12, and Panel D: 1980:1-1989:12): low transactions costs.

Final Wealth and Rank of $100.00 Invested For Various Models: Ten-Year Periods										
Period	LR	Rank	NN (Qi)	Rank	NN (Rep)	Rank	NP	Rank	MN	Rank
1	$177.94	(5)	$222.33	(3)	$224.12	(2)	$179.94	(4)	$227.65	(1)
2	$337.28	(2)	$395.96	(1)	$286.44	(3)	$257.81	(4)	$183.07	(5)
3	$452.25	(3)	$560.57	(1)	$383.88	(4)	$212.70	(5)	$452.54	(2)

Table 4: Final wealth for five year periods (1960:1-1964:12, 1965:1-1969:12, 1970:1-1974:12, 1975:1-1979:12, 1980:1-1984:12, 1985:1-1989:12): low transactions costs.

Final Wealth and Rank of $100.00 Invested For Various Models: Five Year Periods										
Period	LR	Rank	NN (Qi)	Rank	NN (Rep)	Rank	NP	Rank	MN	Rank
1	$150.00	(5)	$164.69	(2)	$162.67	(3)	$153.39	(4)	$175.16	(1)
2	$112.69	(4)	$128.24	(2)	$130.87	(1)	$111.43	(5)	$123.46	(3)
3	$169.51	(2)	$179.97	(1)	$156.39	(3)	$147.15	(4)	$96.03	(5)
4	$175.73	(2)	$194.32	(1)	$161.76	(4)	$154.74	(5)	$170.13	(3)
5	$181.02	(3)	$213.73	(1)	$194.60	(2)	$145.14	(5)	$179.97	(4)
6	$223.62	(3)	$234.76	(1)	$176.57	(4)	$145.01	(5)	$225.08	(2)

strategy between stocks and bonds. As mentioned, we were unable to replicate Qi's results and the replicated NN model does not perform nearly as well from a profit standpoint as she has reported. But we include both of them here for the interested reader. We examine the number of times that the switching strategy based upon a given model yields the highest profit, and we must allow for ties given the nature of the data and period-length. Results appear in tables 2 through 5.

Turning next to profits and to the relative ranking of the models reported in Tables 2 through 5 for the low transaction cost scenario and 6 through 9 for the high cost scenario (Appendix A), it becomes evident that no model dominates throughout. Furthermore, we observe that the buy-and-hold strategy generated by the simple mean predictions model (MN) performs comparatively well for a number of sub-periods, often out-performing the

Table 5: Final wealth for one year periods (1960:1-1960:12, 1961:1-1961:12 and so on): low transactions costs.

Final Wealth and Rank of $100.00 Invested For Various Models: Annual Periods										
Period	LR	Rank	NN (Qi)	Rank	NN (Rep)	Rank	NP	Rank	MN	Rank
1	$96.02	(5)	$102.05	(1)	$100.80	(4)	$102.03	(2)	$101.81	(3)
2	$117.12	(1)	$117.12	(1)	$117.12	(1)	$117.12	(1)	$117.12	(1)
3	$101.94	(1)	$87.27	(4)	$87.27	(4)	$88.26	(3)	$92.11	(2)
4	$106.84	(5)	$112.42	(1)	$112.42	(1)	$112.42	(1)	$112.42	(1)
5	$111.18	(1)	$111.18	(1)	$111.18	(1)	$111.18	(1)	$111.18	(1)
6	$106.18	(1)	$106.18	(1)	$106.18	(1)	$103.69	(5)	$106.18	(1)
7	$91.74	(4)	$102.84	(1)	$102.84	(1)	$93.63	(3)	$88.21	(5)
8	$101.51	(5)	$105.72	(3)	$105.72	(3)	$108.98	(2)	$110.30	(1)
9	$111.85	(3)	$110.79	(4)	$113.06	(2)	$108.69	(5)	$119.30	(1)
10	$106.76	(1)	$104.36	(2)	$104.36	(2)	$92.60	(4)	$92.60	(4)
11	$107.86	(2)	$103.98	(4)	$103.98	(4)	$110.96	(1)	$104.85	(3)
12	$99.62	(2)	$99.62	(2)	$99.62	(2)	$104.87	(1)	$99.62	(2)
13	$111.63	(2)	$107.91	(4)	$107.91	(4)	$108.49	(3)	$113.76	(1)
14	$104.94	(1)	$104.94	(1)	$104.94	(1)	$97.90	(4)	$83.99	(5)
15	$105.59	(2)	$121.51	(1)	$105.59	(2)	$105.59	(2)	$74.42	(5)
16	$124.68	(2)	$131.92	(1)	$119.36	(4)	$113.32	(5)	$121.51	(3)
17	$103.37	(1)	$98.37	(5)	$99.27	(4)	$103.37	(1)	$103.37	(1)
18	$95.56	(3)	$98.24	(2)	$99.96	(1)	$95.56	(3)	$95.56	(3)
19	$120.21	(1)	$115.06	(3)	$115.06	(3)	$120.21	(1)	$109.64	(5)
20	$107.04	(3)	$115.58	(1)	$107.04	(3)	$98.34	(5)	$109.86	(2)
21	$122.22	(4)	$127.39	(1)	$127.39	(1)	$106.82	(5)	$127.11	(3)
22	$111.64	(1)	$111.64	(1)	$105.06	(3)	$99.61	(5)	$100.68	(4)
23	$122.13	(2)	$127.74	(1)	$119.52	(4)	$121.04	(3)	$119.52	(4)
24	$117.51	(1)	$117.51	(1)	$117.51	(1)	$117.51	(1)	$117.51	(1)
25	$94.12	(5)	$102.96	(2)	$102.96	(2)	$102.44	(4)	$102.97	(1)
26	$114.76	(4)	$115.70	(1)	$114.89	(3)	$102.91	(5)	$115.33	(2)
27	$119.82	(1)	$119.82	(1)	$119.82	(1)	$112.37	(5)	$119.82	(1)
28	$107.88	(1)	$107.88	(1)	$85.27	(3)	$85.27	(3)	$85.27	(3)
29	$104.88	(3)	$107.29	(2)	$104.65	(4)	$98.97	(5)	$108.50	(1)
30	$105.62	(3)	$101.66	(5)	$105.62	(3)	$107.90	(2)	$118.27	(1)
31	$109.26	(1)	$109.26	(1)	$109.26	(1)	$91.58	(5)	$99.82	(4)
32	$110.91	(2)	$110.91	(2)	$110.91	(2)	$114.30	(1)	$110.91	(2)
33	$107.07	(1)	$107.07	(1)	$107.07	(1)	$107.07	(1)	$107.07	(1)

linear, neural-network, and nonparametric predictions. We see that, under the low cost scenario, for ten-year periods the buy-and-hold strategy dominates the LR and NN switching

portfolios for 2 out of 3 decades (Table 3), and is equal to or better for 3 out of 6 five year periods (Table 4). For the high cost scenario, we see that, under the low cost scenario, for ten-year periods the buy-and-hold strategy again dominates the LR and replicated NN switching portfolios for 2 out of 3 decades (Table 7), and is better for 4 out of 6 five year periods relative to LR and 3 relative to the replicated NN (Table 8).

The relatively good performance of a buy-and hold strategy using a model which does not use predictive variables, but simply forecasts using the historical average of excess returns available when the prediction is being made, deserves further interpretation. One interpretation is that this finding may not be that remarkable given recent studies which highlight the mean-reverting nature of some markets (Balvers, Wu & Gilliland (2000)). As our graphs show, the mean of excess returns is almost zero (but positive) over the entire sample. All of the conditional specifications may therefore be attempts to model this almost zero mean. These attempts may produce some apparently large or "interesting" residuals which may reflect estimator properties rather than model characteristics.

4 Conclusions

Occasional findings of predictability here and elsewhere seemingly derive from a certain, small degree of *unconditional* nonlinear serial dependence in the returns. But, any finding of linear/nonlinear predictability involves difficult questions of *conditional dependence.* Our and other results indicate that the latter type of inference is sensitive to the period of analysis, frequency of data observations, conditioning variables, functional forms for conditional mean and variance, and predictability criteria. The nonlinear memory effect detected by us using both entropy-based and traditional measures may be too small to allow any robust and durable finding of conditional predictability in stock market returns. Indeed nonlinearities may inherently lead to poorer predictability in a statistically deeper sense. We conclude

that empirical evidence in favor of market-switching strategies over simple buy-and-hold strategies is fragile at best. While our data do not extend to the decade of the 1990s, given the well-known recent market performance it is likely that these findings would be further reinforced.

References

Abhyankar, A., Copeland, L. & Wong, W. (1997), 'Uncovering nonlinear structure in real-time stock-market indexes: The s&p 500, the dax, the nikkei 225, and the ftse-100', *Journal of Business and Economic Statistics* **15**(1), 1–14.

Balvers, R., Wu, Y. & Gilliland, E. (2000), 'Mean reversion across national stock markets and parametric contrarian investment strategies', *Journal of Finance* .

Brock, W., Deckert, W., Scheinkman, J. & LeBaron, B. (1996), 'A test for independence based on the correlation dimension', *Econometric Reviews* **15**(3), 197–236.

Cameron, A. & Windmeijer, F. A. G. (1997), 'An r-squared measure of goodness of fit for some common nonlinear regression models', *Journal of Econometrics* **77**, 329–342.

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

Delgado, M. (1994), 'Testing for serial correlation using sample distribution function', *Journal of Time series Analysis* **17**(3), 271–285.

Diebold, F. X. & Nason, J. A. (1990), 'Nonparametric exchange rate prediction', *Journal of International Economics* **28**, 315–332.

Ebrahimi, N., Maasoumi, E. & Soofi, E. (1999), 'Comparison of entropy and variance orderings', *J. of Econometrics* .

Efron, B. & Tibshirani, R. (1993), *An Introduction to the Bootstrap*, Chapman & Hall.

Frank, M. & Stengos, T. (1988), 'Chaotic dynamics in economic time series', *Journal of Economic Surveys* **2**(2), 103–133.

Geweke, J. (1982), 'Measurement of linear dependence and feedback', *Journal of American Statistical Association* **77**, 304–313.

Gourieroux, C., Monfort, A. & Renault, E. (1986), 'Kullback causality measures', *Mimeo, Ecole Nationale de la Statistique et de l'Administration Economique* .

Granger, C. & Maasoumi, E. (1993), 'A metric entropy measure', *unpublished notes, Department of Economics, UCSD and SMU* .

Granger, C., Maasoumi, E. & Racine, J. (2000), 'A metric entropy and a new test of independence', *UCSD Working Paper* .

Havrda, J. & Charvat, F. (1967), 'Quantification method of classification processes: concept of structural α-entropy', *Kybernetika Cislo I. Rocnik* **3**, 30–34.

Hirschberg, D., Maasoumi, E. & Slottje, D. (2000), 'Clusters of attributes and well-being in the us', *Journal of Applied Econometrics, Forthcoming* .

Hong, Y. & White, H. (2000), 'Asymptotic distribution theory for nonparametric entropy measures of serial dependence', *Mimeo, Department of Economics, Cornell University, and UCSD.* .

Hsieh, D. (1989), 'Testing for nonlinear dependence in foreign exchange rates: 1974-1983', *Journal of Business* **62**, 339–68.

Joe, H. (1989), 'Relative entropy measures of multivariate dependence', *Journal of the American Statistical Association* **84**, 157–164.

Lau, H. T. (1995), *A Numerical Library in C for Scientists and Engineers*, CRC Press, Tokyo.

Liu, Z. & Stengos, T. (1999), 'Nonlinearities in cross-country growth regressions: A semiparametric approach', *Journal of Applied Econometrics* **14**, 527–538.

Ljung, G. & Box, G. (1978), 'On a measure of lack of fit in time series models', *Biometrica* **65**, 297–303.

Maasoumi, E. (1993), 'A compendium to information theory in economics and econometrics', *Econometric Reviews* pp. 137–181.

Pagan, A. & Schwert, G. (1990), 'Alternative models for conditional stock volatility', *Journal of Econometrics* **45**, 267–90.

Pesaran, M. H. & Timmermann, A. (1995), 'Predictability of stock returns: Robustness and economic significance', *Journal of Finance* **50**, 1201–1228.

Qi, M. (1999), 'Nonlinear predictability of stock returns using financial and economic variables', *Journal of Business and Economic Statistics* **17**(4), 419–429.

Racine, J. (1997), 'Consistent significance testing for nonparametric regression', *Journal of Business and Economic Statistics* **15**(3), 369–379.

Racine, J. (2001), 'On the nonlinear predictability of stock returns using financial and economic variables', *Journal of Business and Economic Statistics (forthcoming)* .

Robinson, P. (1991), 'Consistent nonparametric entropy based testing', *Review of Economic Studies* **58**, 437–453.

Scheinkman, J. & LeBaron, B. (1989), 'Nonlinear dynamics and stock returns', *Journal of Business* **62**, 311–37.

Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis*, Chapman and Hall.

Skaug, H. & Tjøstheim, D. (1993), Nonparametric tests of serial independence, *in* S. Rao, ed., 'Developments in Time Series Analysis', Chapman and Hall, pp. 207–229.

Skaug, H. & Tjøstheim, D. (1996), Testing for serial independence using measures of distance between densities, *in* P. Robinson & M. Rosenblatt, eds, 'Athens Conference on Applied Probability and Time Series', Springer Lecture Notes in Statistics, Springer.

Stengos, T. (1995), Nonparametric forecasts of gold rates of return, *in* A. K. Barnett & M. Salmon, eds, 'Nonlinear Dynamics and Economics', Cambridge University Press.

White, H. (1988), 'Economic prediction using neural networks: The case of ibm daily stock returns', *Paper no. 88-20, UCSD* .

Zheng, J. (2000), 'A specification test of conditional parametric distribution using kernel estimation methods', *Econometric Theory* .

A Comparing Profitability for Different Models and Time Horizons - High Transactions Costs

Table 6: Final wealth for the entire period (Panel A: 1960:1-1992:12): high transactions costs.

Final Wealth and Rank of $100.00 Invested For Various Models: Entire Period										
Period	LR	Rank	NN (Qi)	Rank	NN (Rep)	Rank	NP	Rank	MN	Rank
1	$3345.86	(2)	$5962.83	(1)	$3292.49	(3)	$735.30	(5)	$2393.67	(4)

Table 7: Final wealth for ten year periods (Panel B: 1960:1-1969:12, Panel C: 1970:1-1979:12, and Panel D: 1980:1-1989:12): high transactions costs.

Final Wealth and Rank of $100.00 Invested For Various Models: Ten-Year Periods										
Period	LR	Rank	NN (Qi)	Rank	NN (Rep)	Rank	NP	Rank	MN	Rank
1	$156.01	(5)	$207.02	(3)	$212.93	(2)	$156.14	(4)	$225.02	(1)
2	$317.22	(2)	$347.22	(1)	$258.85	(3)	$232.91	(4)	$180.88	(5)
3	$404.43	(3)	$496.24	(1)	$357.34	(4)	$182.75	(5)	$447.09	(2)

Table 8: Final wealth for five year periods (1960:1-1964:12, 1965:1-1969:12, 1970:1-1974:12, 1975:1-1979:12, 1980:1-1984:12, 1985:1-1989:12): high transactions costs.

Final Wealth and Rank of $100.00 Invested For Various Models: Five Year Periods										
Period	LR	Rank	NN (Qi)	Rank	NN (Rep)	Rank	NP	Rank	MN	Rank
1	$141.15	(5)	$161.31	(2)	$159.34	(3)	$145.78	(4)	$173.28	(1)
2	$103.94	(4)	$120.69	(3)	$125.68	(1)	$100.73	(5)	$122.13	(2)
3	$166.08	(2)	$176.34	(1)	$154.79	(3)	$138.49	(4)	$95.00	(5)
4	$168.69	(2)	$173.91	(1)	$147.70	(5)	$148.54	(4)	$168.25	(3)
5	$165.26	(4)	$199.08	(1)	$183.08	(2)	$133.86	(5)	$177.96	(3)
6	$216.87	(3)	$220.89	(2)	$172.97	(4)	$135.09	(5)	$222.64	(1)

Table 9: Final wealth for one year periods (1960:1-1960:12, 1961:1-1961:12 and so on): high transactions costs.

Final Wealth and Rank of $100.00 Invested For Various Models: Annual Periods										
Period	LR	Rank	NN (Qi)	Rank	NN (Rep)	Rank	NP	Rank	MN	Rank
1	$93.16	(5)	$101.01	(1)	$99.78	(4)	$99.99	(3)	$100.77	(2)
2	$115.93	(1)	$115.93	(1)	$115.93	(1)	$115.93	(1)	$115.93	(1)
3	$100.92	(1)	$86.39	(4)	$86.39	(4)	$86.50	(3)	$91.18	(2)
4	$104.71	(5)	$111.28	(1)	$111.28	(1)	$111.28	(1)	$111.28	(1)
5	$110.05	(1)	$110.05	(1)	$110.05	(1)	$110.05	(1)	$110.05	(1)
6	$105.11	(1)	$105.11	(1)	$105.11	(1)	$101.62	(5)	$105.11	(1)
7	$89.01	(4)	$102.84	(1)	$102.84	(1)	$91.76	(3)	$87.31	(5)
8	$100.48	(5)	$104.65	(3)	$104.65	(3)	$105.74	(2)	$109.18	(1)
9	$110.72	(3)	$107.49	(4)	$111.93	(2)	$106.52	(5)	$118.09	(1)
10	$105.69	(1)	$104.36	(2)	$104.36	(2)	$91.66	(4)	$91.66	(4)
11	$106.78	(2)	$103.98	(3)	$103.98	(3)	$108.75	(1)	$103.78	(5)
12	$98.61	(2)	$98.61	(2)	$98.61	(2)	$103.81	(1)	$98.61	(2)
13	$109.40	(2)	$106.82	(4)	$106.82	(4)	$107.40	(3)	$112.61	(1)
14	$104.94	(1)	$104.94	(1)	$104.94	(1)	$95.95	(4)	$83.14	(5)
15	$105.59	(2)	$120.29	(1)	$105.59	(2)	$105.59	(2)	$73.66	(5)
16	$120.96	(2)	$129.28	(1)	$115.81	(4)	$111.05	(5)	$120.28	(3)
17	$102.32	(1)	$96.41	(4)	$96.31	(5)	$102.32	(1)	$102.32	(1)
18	$94.59	(3)	$95.31	(2)	$96.98	(1)	$94.59	(3)	$94.59	(3)
19	$119.00	(1)	$112.76	(3)	$112.76	(3)	$119.00	(1)	$108.53	(5)
20	$105.97	(3)	$113.28	(1)	$105.97	(3)	$97.36	(5)	$108.74	(2)
21	$120.99	(4)	$126.11	(1)	$126.11	(1)	$104.69	(5)	$125.81	(3)
22	$108.32	(1)	$108.32	(1)	$101.93	(3)	$98.61	(5)	$99.66	(4)
23	$119.69	(2)	$126.43	(1)	$118.30	(4)	$118.62	(3)	$118.30	(4)
24	$116.31	(1)	$116.31	(1)	$116.31	(1)	$116.31	(1)	$116.31	(1)
25	$91.32	(5)	$101.93	(1)	$101.93	(1)	$101.41	(4)	$101.92	(3)
26	$112.46	(4)	$113.38	(2)	$112.59	(3)	$101.88	(5)	$114.15	(1)
27	$118.60	(1)	$118.60	(1)	$118.60	(1)	$110.12	(5)	$118.60	(1)
28	$105.72	(1)	$105.72	(1)	$84.41	(3)	$84.41	(3)	$84.41	(3)
29	$103.83	(3)	$105.15	(2)	$103.60	(4)	$96.02	(5)	$107.40	(1)
30	$105.62	(3)	$100.64	(5)	$105.62	(3)	$105.74	(2)	$117.07	(1)
31	$108.16	(1)	$108.16	(1)	$108.16	(1)	$88.85	(5)	$98.81	(4)
32	$109.79	(2)	$109.79	(2)	$109.79	(2)	$110.90	(1)	$109.79	(2)
33	$105.99	(1)	$105.99	(1)	$105.99	(1)	$105.99	(1)	$105.99	(1)

B Finite-Sample Behavior of the Resampled Correlation-Based Measures

We investigate the power properties of the resampled moment-based measures of dependence used in our paper. To do this, $1,000$ samples were drawn from independent data, $\{Y, X\}$, and from dependent data defined by

$$y_i = x_i + u_i \tag{5}$$

where $x \sim N(0,1)$ and $u \sim N(0,1)$, while for the independent data

$$y_i = u_i \tag{6}$$

For each of the 1,000 resamples we compute the values of each of CORR, SIGN, MAE, MAPE, and RMSE, and we bootstrap their distribution under the null of independence. The null distribution is obtained by applying a random shuffle to x to 'break' any dependence between x and y, and we then compute the values of CORR, SIGN, MAE, MAPE, and RMSE for this independent data and repeat this 1,000 times obtaining the percentiles under the null. Sample sizes considered were $n = 100$, 250, and 500.

The mean values of the resampled statistics were computed and their percentiles over the 1,000 Monte Carlo draws are reported in Table 10. We do not intend this to be an exhaustive examination of the finite-sample properties of this method, rather, we intend this to be helpful to the interested reader who wishes to examine the finite-sample properties of our approach under the scenarios of independent versus (linearly) dependent processes.

It is clear from this modest experiment that when the data are truly linearly independent our resampling approach is capable of picking this up (asymptotically). It is also clear

from this experiment that if there exists *linear* dependence then our resampling approach is capable of detecting this asymptotically.

However, we note that the moment-based measures are indeed point estimates of a model's performance, thus they are subject to sampling variability. It is desirable to differentiate between models in a statistically meaningful manner when using these measures. That is, while one might be tempted to state that 'Model A' is preferable to 'Model B' since its RMSE is lower, it could be the case that there is no *statistically significant* difference between these models. Alternatively, as we suggest here, a relevant question would be whether or not the out-of-sample predictions differ significantly from those from a model with no predictive power. We therefore caution users against comparison of models on the basis of any of the standard measures of model performance *unless* one also constructs their sampling distributions in the manner implemented herein. We also caution readers to note that MAPE does not appear to be well-suited to this task. We are currently examining this issue as the measure enjoys widespread use in the economics and finance literature.

We refer readers interested in the performance of $\hat{S}_\rho$ in a wide variety of settings to Granger et al. (2000). Of particular interest is the fact that traditional measures are shown to fail for many models, sometimes badly. By way of example, Granger et al. (2000) demonstrate that their model of a "chaotic process" (Model 10) has an autocorrelation function that is indistinguishable from a white noise series, while $\hat{S}_\rho$ readily detects strong and significant dependence when using the resampling approach applied here for traditional measures.

Table 10: Linear Measures of Out-of-Sample Goodness of Fit and Bootstrap Percentiles Under the Null of Independence.

Measure	Value	$\text{Pct}_{0.005}$	$\text{Pct}_{0.025}$	$\text{Pct}_{0.05}$	$\text{Pct}_{0.095}$	$\text{Pct}_{0.975}$	$\text{Pct}_{0.995}$
			$n = 100$				
			Independent Data				
CORR:	-0.0043	-0.2596	-0.1973	-0.1659	0.1652	0.1958	0.2527
SIGN:	0.4978	0.3702	0.4019	0.4178	0.5817	0.5975	0.6260
MAE:	1.1356	0.9200	0.9706	0.9959	1.2736	1.3014	1.3547
MAPE:	8.4230	1.9681	2.2627	2.4518	19.1644	22.4119	29.5108
RMSE:	1.4191	1.1632	1.2235	1.2539	1.5770	1.6083	1.6688
			Dependent Data				
CORR:	0.7059	-0.2642	-0.2027	-0.1715	0.1568	0.1877	0.2452
SIGN:	0.7487	0.3726	0.4042	0.4204	0.5844	0.6001	0.6293
MAE:	0.8002	1.0998	1.1644	1.1973	1.5659	1.6035	1.6752
MAPE:	7.3249	2.6448	3.0571	3.3205	27.6059	32.0466	41.7175
RMSE:	0.9993	1.3903	1.4680	1.5076	1.9390	1.9814	2.0624
			$n = 250$				
			Independent Data				
CORR:	0.0003	-0.1647	-0.1243	-0.1045	0.1040	0.1235	0.1602
SIGN:	0.5004	0.4179	0.4380	0.4479	0.5518	0.5614	0.5800
MAE:	1.1270	0.9924	1.0245	1.0408	1.2148	1.2319	1.2647
MAPE:	7.5983	2.4581	2.7639	2.9617	16.7074	19.0975	24.3081
RMSE:	1.4112	1.2520	1.2902	1.3094	1.5127	1.5323	1.5696
			Dependent Data				
CORR:	0.7067	-0.1662	-0.1266	-0.1066	0.1008	0.1204	0.1573
SIGN:	0.7514	0.4189	0.4387	0.4487	0.5527	0.5626	0.5810
MAE:	0.7973	1.2012	1.2433	1.2646	1.4963	1.5192	1.5633
MAPE:	8.7442	3.3268	3.7667	4.0467	29.1595	33.6599	43.1593
RMSE:	0.9985	1.5148	1.5653	1.5905	1.8636	1.8904	1.9415
			$n = 500$				
			Independent Data				
CORR:	-0.0009	-0.1165	-0.0881	-0.0739	0.0734	0.0872	0.1135
SIGN:	0.5001	0.4419	0.4561	0.4632	0.5368	0.5437	0.5568
MAE:	1.1305	1.0333	1.0567	1.0684	1.1918	1.2038	1.2264
MAPE:	11.7687	2.8632	3.1875	3.3929	33.8989	40.0101	54.6691
RMSE:	1.4161	1.3018	1.3294	1.3432	1.4875	1.5014	1.5277
			Dependent Data				
CORR:	0.7064	-0.1177	-0.0893	-0.0751	0.0719	0.0858	0.1118
SIGN:	0.7497	0.4423	0.4565	0.4636	0.5371	0.5440	0.5571
MAE:	0.7986	1.2539	1.2844	1.2998	1.4641	1.4799	1.5100
MAPE:	9.5904	3.8867	4.3369	4.6210	28.4413	32.8146	41.6754
RMSE:	0.9999	1.5786	1.6149	1.6334	1.8272	1.8458	1.8809

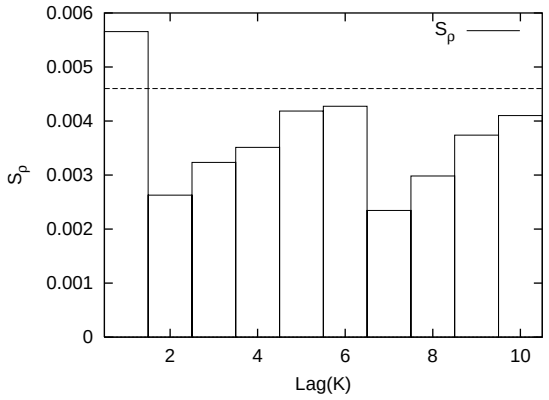

Figure 1: $\hat{S}_\rho$ for lags $K = 1, 3, \ldots, 10$ and their 90th percentile under the null of serial independence.

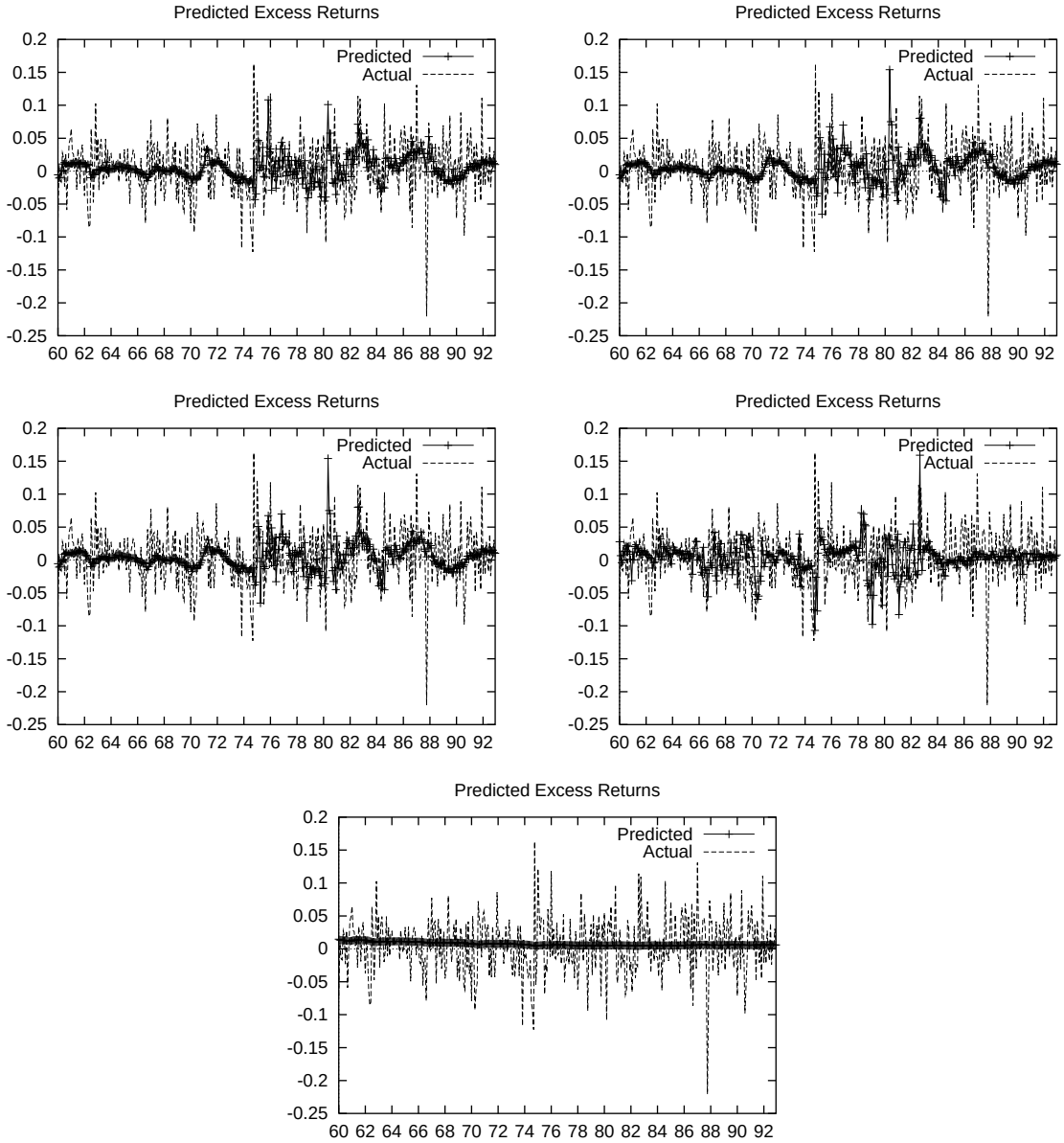

Figure 2: Recursive one-step-ahead predicted and actual returns for the period 1960:1-1992:12. The top left figure is the NN in Qi (1999), the top right the replicated NN, the middle left LR, middle right NP, and the bottom MN.

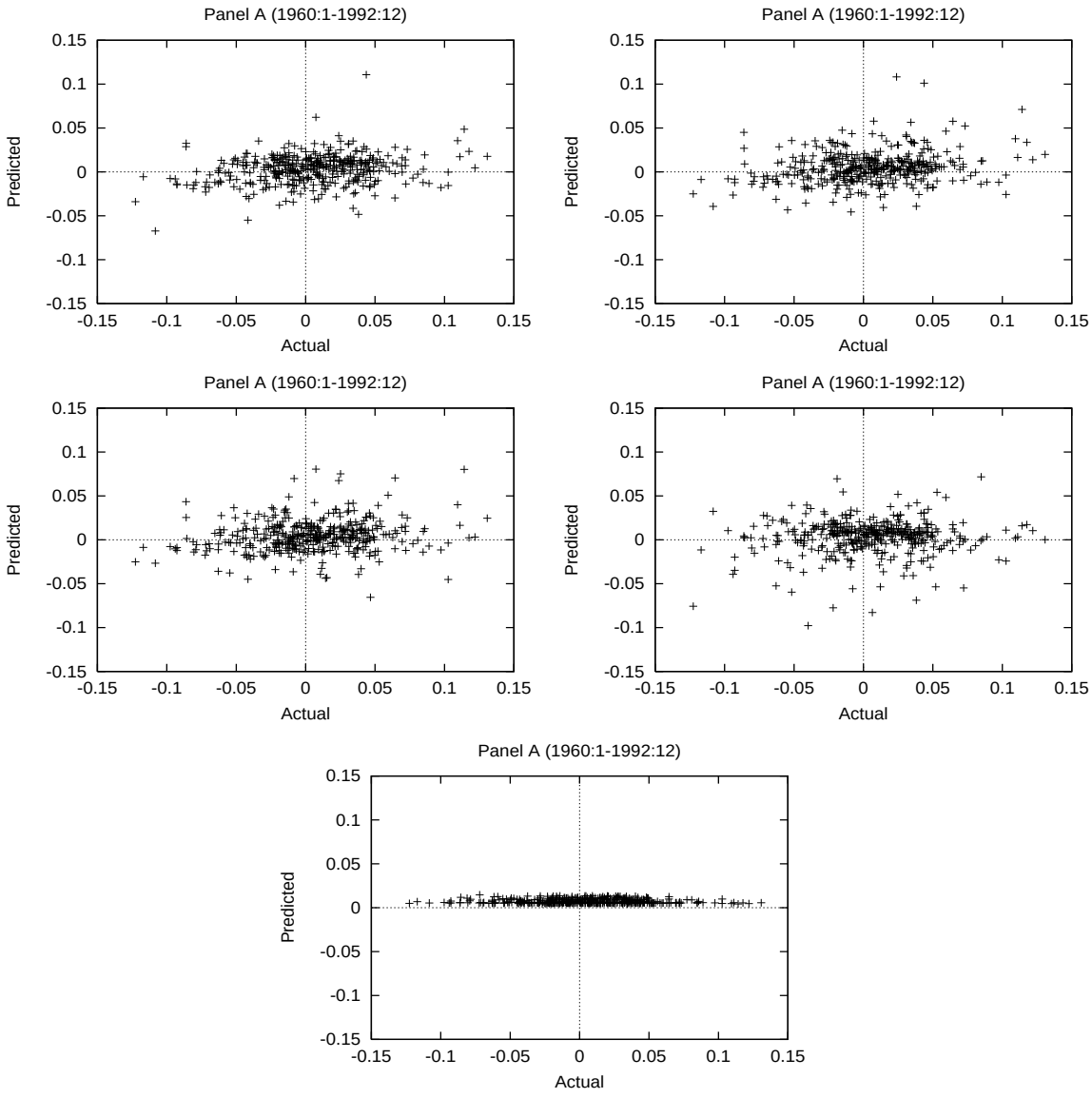

Figure 3: Predicted versus actual returns for the period 1960:1-1992:12. The top left figure is the LR model, the top right the NN in Qi (1999), the middle left the replicated NN, middle right NP, and the bottom MN.