CHAPTER 4

The Prediction Test for Nonlinear Determinism

Ted Jaditz

One of the more interesting problems of modern empirical economics is what one might call the forecasting paradox: Standard linear statistical models of economic phenomena invariably fit very well in sample. However, results for out-of-sample prediction are typically much worse than one might expect given the in-sample fit. This problem has been known in the profession since the late 1940s; see, for example, the discussion in Malinvaud (1970, chap. 4) of some early efforts toward predicting postwar consumption patterns.

Given the difficulties that economists have in forecasting, many in the profession are quite receptive to the possibility of nonlinear determinism in economic data. Some claim to have found evidence of chaos. (The most notable examples are Barnett and Chen [1988] and DeCoster and Mitchell [1991].) We call this the chaos hypothesis. If true, the hypothesis could help to explain the forecasting paradox. If the data-generating mechanisms are nonlinear, then it is not surprising that linear models don't forecast very well. Second, if the data generator is chaotic, then long-run point prediction is hopeless anyway, for reasons that we discuss below.

Direct tests of the chaos hypothesis are problematic. Certain tests, such as the calculation of dimension and entropy numbers, are quite useful at identifying underlying nonlinear determinism in the experimental sciences. However, these tests are difficult to apply to economic data, and the results to date are highly controversial. Economic data sets tend to exhibit certain features, such as regime shifting and volatility clustering, that make it difficult to sort out whether the underlying data are nonlinear deterministic or nonlinear stochastic. See Jaditz and Sayers (1993a) for a brief overview.

A great deal of literature is emerging that suggests that an approach based on out-of-sample prediction may be able to tell us whether it is likely that the underlying data-generating mechanism is nonlinear. Even if the underlying process is chaotic, certain special nonlinear forecasting techniques ought to do significantly better at short-term forecasting than the linear techniques typically used by economists. We refer to this hunt for forecastable

nonlinear structure as the prediction test. This paper will discuss the basis for the prediction test and evaluate the prospects for its application to economic data. We illustrate by applying the test to monthly inflation data. We ask whether nonlinear methods offer statistically significant improvements in forecast performance over the standard linear models typically employed by economists.

To summarize the discussion, the prediction test for nonlinearity is extremely promising. There are certain difficulties in applying the test and interpreting the results: It is not at all obvious what measure of forecast accuracy should be used, and certain care must be taken in interpreting the statistical tests of forecast improvement. Despite these difficulties, out-of-sample forecast improvement is perhaps the most persuasive evidence that one can offer in favor of the hypothesis of nonlinearity.

Having said all that, the evidence for deterministic nonlinearity in economic data is weak. The methods that work very well at predicting even noisy chaos do not offer substantially improved forecasts of economic time series. This suggests that it is rather unlikely that low-dimensional chaos underlies economic time series.

Preliminaries: What Is the Chaos Hypothesis?

The prediction test works because nonparametric forecasting methods can exploit the subtle dependence that characterizes certain nonlinear deterministic systems. To explain why the prediction test works, we must digress somewhat to discuss more precisely the characteristics of nonlinear systems, and to point out the features of these systems that allow us to form accurate short-run forecasts. Following is a very brief discussion of what we mean by a chaotic, deterministic system. We refer the reader to Brock (1986) for a more technical introduction from an economic perspective, and Eckmann and Ruelle (1985) for an introduction from the perspective of physical systems.

DEFINITION 1. (Brock 1986) *The time series* $\{a_t\}_{t=1}^{\infty}$ *has a smoothly deterministic explanation if a system exists* (h, F, X_0) *such that* $h: \Re^n \rightarrow \Re$ *and* $F: \Re^n \rightarrow \Re^n$ *are "smooth" (i.e., twice differentiable almost everywhere) and*

$$X_t = F(X_{t-1}) \tag{1}$$

$$a_t = h(X_t) \tag{2}$$

To interpret, the relationship $X_t = F(X_{t-1})$ is an unknown law of motion involving the unobserved **n** vector of state variables. Given an initial starting

value X_0 in period 0, the state evolves to $F(X_0) = X_1$ in period 1, then $F(X_1) = X_2$ in period 2, and so on. The sequence of values of the state variable $\{X_0, X_1, X_2, \ldots\}$ defines an orbit of the system. We introduce some additional notation: let $F^t(x)$ be defined as the t^{th} iterate of F. For example, $F^2(x) = F(F(x))$, $F^3(x) = F(F(F(x)))$, and so on. Thus, $X_t = F^t(X_0)$. If we knew the precise initial state of the process and the transition function, F, then in principle we could predict the subsequent evolution of the process perfectly accurately. Note that the function h can be interpreted as a "measuring device," which provides information about the current values of the state variables. In this scheme, we need not observe the state variables directly.

Of particular interest is the case where $F()$ is bounded on a closed subspace of $\Re^n$, and admits an attractor: Heuristically, the attractor is a parsimonious description of the long-run behavior of the system. The long-run behavior of the system can be envisaged as orbits moving along the attractor.

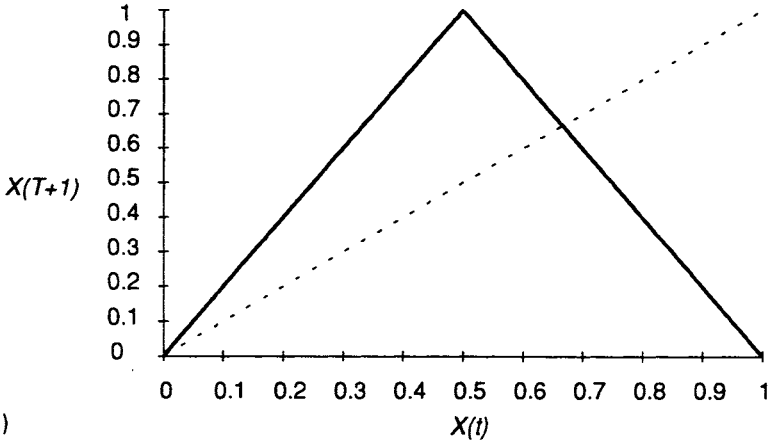
DEFINITION 2. *A set Λ is an attractor for the system (h, F, X_0) relative to a set U if the following hold.*

- *(a) For all $x \in \Lambda$, $F(x) \in \Lambda$. This means that once an orbit enters the attractor, all subsequent motion of the system is on the attractor. (Formally, the set Λ is invariant with respect to the law of motion F.)*
- *(b) The attractor isn't a union of disjoint sets. (The set Λ is indecomposable.)*
- *(c) The orbits generated by F are dense in Λ. This means that the trajectory visits all neighborhoods of the attractor, so that a single realization of a time series will fully describe the system if that realization has enough observations. (The set Λ is topologically transitive.)*
- *(d) Finally, for the law of motion F, $\Lambda = \cap_{t=1}^{\infty} F^t(U)$. All orbits that start anywhere in U eventually converge to the attractor.*

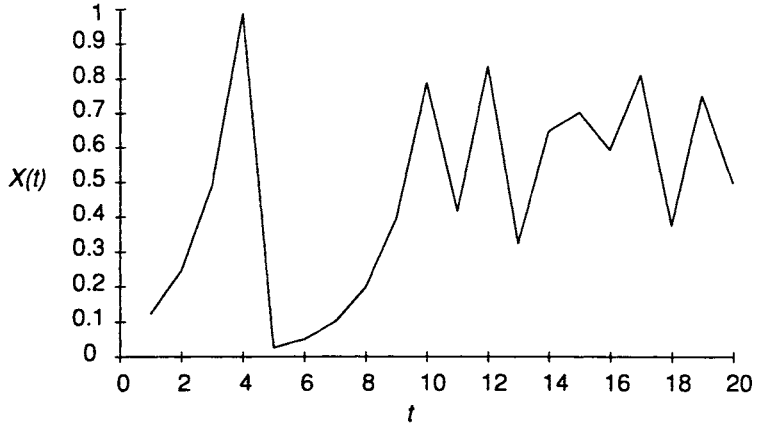
To illustrate these definitions, consider the tent map, whose properties were first described by Saki and Tokumaru (1980). The tent map is a simple nonlinear iterative map from [0, 1] to [0, 1]. The transition function F for the tent map is given by the equations

$$X_{t+1} = \begin{cases} 2X_t & \text{if } X_t < 0.5 \\ 2(1 - X_t) & \text{if } X_t > 0.5 \end{cases} \tag{3}$$

The phase curve and a sample time path are given in figure 4.1. The attractor for the tent map is just the phase curve, the kinked line segment in (X_t, X_{t+1})



(a)



(b)

Fig. 4.1. The tent map: (a) phase portrait; (b) typical time path

space from (0, 0) to (0.5, 1), to (1, 0). Starting from any point on the attractor X_t, the subsequent value $X_{t+1} = F(X_t)$ will also lie on the attractor. Thus, even though a typical orbit of the tent map appears haphazard to the eye (and to many statistical tests!), it is in fact generated by a very simple rule. If one knew the rule, one could predict subsequent observations to a high degree of accuracy.

Note that attractors need not be as simple as this example. One way to describe an attractor is to give a measure of its complexity. One such measure is the dimension. For the tent map, the attractor is a one-dimensional curve in a two-dimensional space. Hence we say that the attractor has dimension one. Attractors can be highly irregular; many interesting systems exhibit attractors of fractional dimension. The dimension is generally less than or equal to the minimum number of coordinates necessary to represent the state space of the process. See Eckmann and Ruelle 1985 for further discussion of this important concept.

Let us add a trivial complication. Suppose that our observer function is

$$a_t = h(X_t) = X_t^3 \tag{4}$$

In other words, we do not observe the state of the system, but we observe some function of the state of the system. Of course, if we knew $h(\cdot)$, we could invert it and thus observe the state of the system directly. However, we typically do not know enough about $h(\cdot)$ to be able to invert it. Can we infer anything about the true state of the system from the observed time series?

The answer to this question is given by the Takens Embedding Theorem. Takens's (1985) theorem states that we can learn about the underlying state variables, X_t, from "embeddings" of the observed a_t's. Given a sequence of observations, $\{a_t\}$, define the *embedding* or m-history as:

$$a_t^m \equiv (a_t, a_{t+1}, \ldots, a_{t+m-1}) \tag{5}$$

Each m-history consists of m consecutive observations, so that the first three points in a sequence of two-histories are (a_1, a_2), (a_2, a_3), and (a_3, a_4). There will be a total of $T - m + 1$ of these, or $T - 1$ histories in the two-history case. Takens (1985) shows that the mapping from the m-histories of a sequence a_t to the state vectors X_t given by

$$a_t^m = \{h(X_t), h(F(X_t)), h(F^2(X_t), \ldots, h(F^{m-1}(X_t)\} \equiv J(X_t) \tag{6}$$

is a diffeomorphism (i.e., a differentiable map with a differentiable inverse), provided that the embedding dimension, m, is sufficiently greater than dimension of the process, F. (If the process is n-dimensional, Takens shows that $m \geq 2n + 1$ will certainly work.) This means that knowing the m-histories, a_t^m, is essentially equivalent to knowing the current state of the system, X_t.

To illustrate, suppose we observe 250 iterates of the tent map filtered through the observer function $h(X_t) = X_t^3$. Plot the corresponding two-histories, and you get a picture that looks like figure 4.2: the two histories map out a one-dimensional curve in the plane. The plot of the two-histories of

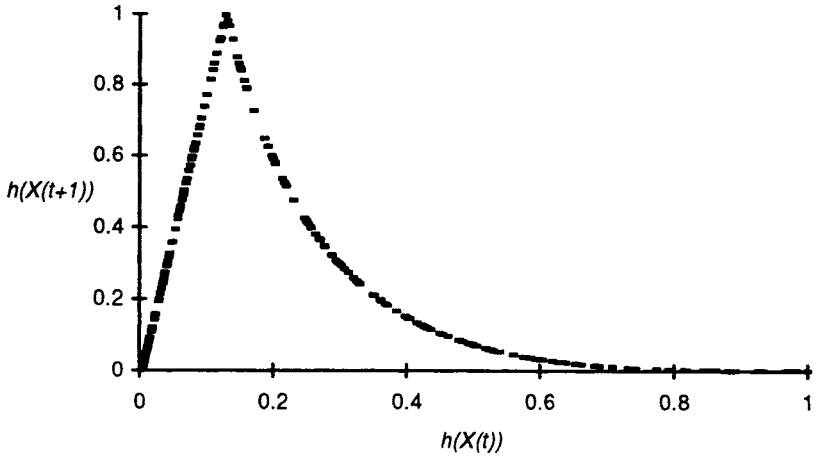


Fig. 4.2. Embedding of the observed values a_1

the observed values, a_t, is in many ways similar to the phase diagram for the unobserved variables, X_t. It is stretched somewhat due to the effects of the observer function, h, and perhaps it would be stretched or folded differently for a different observer function. However, Takens theorem states that many of the fundamental topological properties of the attractor in the unobserved state space of the process are preserved in and can be inferred from embeddings of the observed data. Takens theorem is the most important result in nonlinear science.

One more feature of the tent map deserves comment. The tent map is also chaotic. We will say that the deterministic system (h, F, X_0) is chaotic if it exhibits sensitive dependence on initial conditions.

DEFINITION 3. *A dynamical system exhibits sensitive dependence on initial conditions if for all x on the attractor, and for all* ε, *there exists a* δ *and a point y on the attractor such that for some t, we have* $\| x - y \| < \varepsilon$ *and* $\| F^t(x) - F^t(y) \| > \delta$.

In plainer words, the time paths of the process started from two nearby states will diverge eventually. Figure 4.3 illustrates two trajectories of the tent map, using starting values that differ by 10^{-4}. Even though the two starting values are quite close, by the eleventh iteration, the paths have diverged enough to lose all subsequent resemblance.

Sensitive dependence to initial conditions makes long-term prediction of their future states virtually impossible. Suppose we only know the current

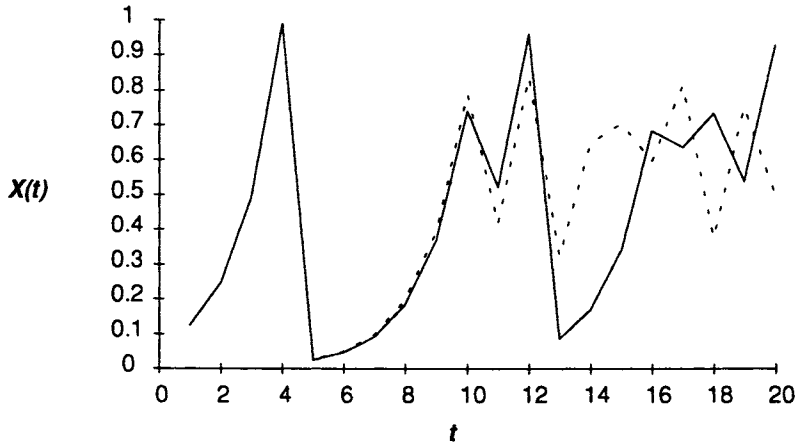


Fig. 4.3. Orbits of the tent map are sensitive to initial conditions

state of a chaotic process to some fixed degree of accuracy, due to the imprecision of our measuring device. Suppose we observe a noisy chaos.

DEFINITION 4. *A noisy chaos is a dynamical system (h^*, F, X_0) where F and X_0 follow definition 1 above, with the observer function "polluted" by the addition of observer noise:*

$$a_t = h^*(X_t) = h(X_t) + \varepsilon_t, \quad \varepsilon_t \text{ a white noise random variable.} \qquad (2')$$

Thus, even if we knew the transition function, F, forecasting would be complicated by the imprecision with which we observe the current state of the system. Suppose that we knew that the current state, X_t, was in an interval $[a, b]$. Then we would predict that X_{t+1} would lie in the interval $F([a, b])$. If the process is chaotic, uncertainty about the current state is rapidly magnified, and thus our long-term forecasts of a noisy chaotic process can do no better than to predict that the process will be somewhere on the attractor. On the other hand, if the interval $[a, b]$ is small enough—if we know the current state to a fair degree of precision—then our prediction of the near future states may have a reasonable degree of accuracy.

The chaos hypothesis is just the suggestion that economic data are generated by a noisy chaos. If this is true, then there is a nonlinear function, F, that is the data generator underlying economic time series. To implement the prediction test, we use embeddings of the observed data to try and identify features of the attractor for this function, F. If the attractor is simple enough,

if the observer function does not obscure the underlying structure, and if the amount of observer noise is small, we may be able to learn enough about the attractor to aid in short-term forecasts.

An Introduction to the Prediction Test

The underlying nonlinear structure of a deterministic system suggests that certain specialized nonlinear forecasting techniques may be able to outperform linear techniques at short horizons. Casdagli (1992) describes how near neighbor algorithms can be used to forecast nonlinear processes. Given the current m-history, a_t^m, we may be able to predict the next observation, a_{t+1}, by looking at past m-histories, a_s^m, $s < t$, that are sufficiently close to a_t^m. If there is a deterministic function underlying the data, and if the noise-to-signal ratio is small, then if a_s^m is sufficiently close to a_t^m, a_{t+1} ought to be close to a_{s+1}. Casdagli shows that if we have a simple nonlinear process on a bounded attractor, and if we have enough observations to get a complete picture of the attractor, this method can generate highly accurate forecasts of nonlinear systems.

We start with observations on a time series $\{x_t\}_{t=1,T}$ which we partition into a fitting set $\mathcal{F}$ and a prediction set $\mathcal{P}$, such as

$$\mathcal{F} = \{x_t : m + 1 < t \leq N_f\} \tag{7}$$

$$\mathcal{P} = \{x_t : N_f < t \leq T\}$$

for some $N_f < T$. The aim is to use the information in $\mathcal{F}$ to predict observations in $\mathcal{P}$.

For a given lag length, m, construct for each observation $\{x_t\}_{t=m+1,T}$, an m-history, x_{t-1}^m

$$x_{t-1}^m = (x_{t-1}, x_{t-2}, \ldots, x_{t-m}). \tag{8}$$

Thus we have a set of ordered pairs, $\{(x_t, x_{t-1}^m)\}_{t=m+1,T}$. For each x_t in the prediction set, we calculate the distance between x_{t-1}^m and x_{s-1}^m $\forall$ $s \in \mathcal{F}$. Casdagli suggests using the norm to calculate distances,

$$\| x \| = \max_i | x_i |. \tag{9}$$

We then select the k nearest pairs, (x_s, x_{s-1}^m), to estimate the parameters in the local regression,

$$x_s = \alpha_k + x_{s-1}^m \beta_k + \varepsilon_s. \tag{10}$$

The estimated parameters $\hat{\alpha}_{0,k}$ and $\hat{\alpha}_k$ are used to calculate the prediction,

$$\hat{x}_t = \hat{\alpha}_k + x_{t-1}^m \hat{\beta}_k. \tag{11}$$

The prediction is then used to calculate the prediction error, $x_t - \hat{x}_t = e_{t,k}$.

Citing the Takens theorem, Casdagli suggests setting the lag length, m, equal to $2n + 1$, where n is the dimension of the process. When the dimension of the process is not known, we can estimate it, using (to pick one possibility out of several) the procedure sketched in Brock and Baek 1991.

We should point out how this procedure compares to the global linear predictors typically used by economists. For the global linear predictor, we use every observation in the fitting set to estimate the parameters of the regression. If the underlying process is sufficiently nonlinear, the global linear predictor will do a poor job of approximating the relationship between the histories and the futures, and the resulting forecasts will be inaccurate.

The essence of the prediction test is to compare the forecasts generated by the linear methodologies preferred by economists to the forecasts generated by the nonlinear, nearest neighbor algorithms preferred by physicists. If the near neighbor algorithm gives more accurate forecasts, then we have evidence that the underlying process is nonlinear. To evaluate forecasting performance, we use the Root Mean Square Error (RMSE) of the forecast. Given a time series, $\{y_1, y_2, \ldots, y_T\}$, and a set of forecasts for that series, $\{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_T\}$, we define the RMSE as

$$\text{RMSE} = \left[\frac{1}{T} \sum_{i=1}^{T} (y_i - \hat{y}_i)^2 \right]^{0.5}. \tag{12}$$

Many authors further normalize the RMSE by dividing by the standard deviation of the data that we are trying to predict: Define the normalized RMSE as

$$\text{NRMSE} = \text{RMSE}/\sigma_y.$$

This normalized RMSE attains a value of 1 if the method of prediction is no more accurate than forecasting the unconditional mean of the prediction set. This normalization is quite useful as an aid to the interpretation of the results. We will follow custom and report only the normalized root mean square error.

Another example is in order. To evaluate this procedure on a short data set, we generated 600 observations from the tent map, using a starting value of 0.1234567891. For our example, we set the fitting set to be the first 500 observations, and the prediction set to be the last 100 observations of the

process. We first attempted to forecast the process with the global linear predictor. We estimate an AR(3) model,

$$x_t = \alpha + \sum_{i=1}^{3} \beta_i x_{t-1} + \varepsilon_t. \tag{13}$$

We use the estimated parameters from the fitting set to generate predictions for the data in $\mathcal{P}$. The results for the linear model are summarized in table 4.1 and are quite poor. First, the linear model does not fit the data very well. The regression R^2 is only about 0.01; the regression explains only about one-half of 1 percent of the total variation of the process. The only significant coefficient is the regression intercept. Out-of-sample forecast performance is dismal. The normalized RMSE for the AR(3) model is 1.024, which is worse than the result obtained if we just use the unconditional mean of the fitting set as our forecast.

If we apply the near neighbor methodology, we obtain far superior results. In figure 4.4, we plot the normalized RMSE as a function of the number of near neighbors used in the regression. For a fitting set of 500 observations, we can set the number of near neighbors used in the regression from one to 500. The minimum of the NRMSE plot is the best near neighbor predictor. The plot is minimized at $k = 5$: our best prediction of an element in $\mathcal{P}$ is obtained when we comb through the fitting set and calculate our prediction based on the five observations in the fitting set that are most similar to the history of the observation that we are trying to predict. A regression calculated using only those five observations will, on average, explain nearly 98

TABLE 4.1. Summary of Regression Results for the Tent Map

Coefficient	Value	Test Statistic
Constant	0.560	12.99
X_{t-1}	−0.059	−1.31
X_{t-2}	−0.019	−0.42
X_{t-3}	−0.029	−0.65
Regression R^2		0.012
Fitting Set Mean Forecast NRMSE		1.008
Regression Forecast RMSE		1.024
Number of Observations		500

Note: The test statistic is the standard OLS test statistic for the hypothesis that the corresponding coefficient is significantly different from zero.

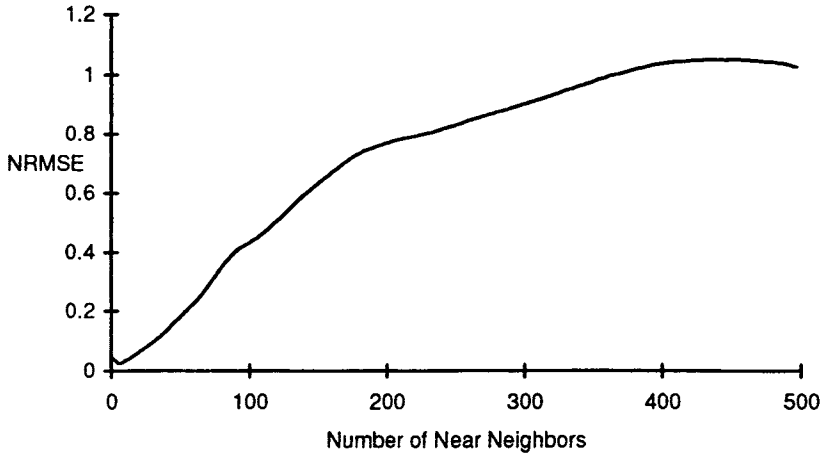


Fig. 4.4. NRMSE plot for the near neighbor forecasts of the tent map

percent of the variation in the process. The forecasts generated by the near neighbor methodology are stunningly more accurate than the forecasts generated by the global linear predictor, and this provides very convincing evidence of the nonlinearity present in the tent map.

It bears mentioning that the NRMSE plot for the tent map is quite typical of the shape of the plot when significant nonlinearity is present. Here, we observe that the plot is minimized for a relatively small number of near neighbors, and the plot is smoothly upward sloping from the minimum out to the maximum number of near neighbors, where we are essentially replicating the global linear predictor.

For real economic data, the difference in forecast performance between the linear and nonlinear predictors is seldom so clear cut. The normalized RMSE for the nonlinear predictor is typically much closer to the RMSE of the linear predictor. We are then left to ask whether the forecast improvement generated by the nonlinear models is sufficient to justify the inference that the data generator is nonlinear. It turns out that answering this question is surprisingly difficult.

The Problem of Forecast Evaluation

There are a number of conceptual problems that must be addressed before one can sensibly compare and evaluate forecasts generated by different methods.

The first problem is: how should we measure forecast accuracy? Almost all authors focus on a mechanical measure of forecast accuracy, such as Root

Mean Square Error (RMSE) or Mean Absolute Deviation (MAD). Each method of measuring forecast accuracy has its aficionados, and one can make a purely technical statistical argument for either. However, in economic contexts, it is very often useful to measure forecast accuracy in terms of an underlying objective function. For example, if one is forecasting the movements of stock prices, the forecast RMSE or MAD are poor yardsticks of forecast accuracy. There are two reasons for this. The first reason is that one isn't interested in stock prices per se, but rather in the profits that one could make by following an optimal trading rule based on the forecasts. Measured in these terms, an accurate forecast is one that allows the trader to make large (risk-adjusted) rates of return. Indeed, the empirical finance literature is full of anomalies that are in some sense "statistically significant," but economically negligible. It is well known, for example, that there is an element of mean reversion in stock prices: a stock that experiences an especially large loss in one week can be expected to rebound somewhat the next. What is debatable is whether the effect is large enough to allow one to make profits in excess of transactions costs by trading on this information (see, for example, Conrad et al. 1990). If one cannot design a trading rule that uses this information to generate profits, then the anomaly is not economically significant, no matter how statistically significant it appears to be.

The second reason is that there are types of forecast improvements that these measures are poorly suited to identify. RMSE and MAD are global measures of forecast accuracy. It is possible to have a method that forecasts poorly most of the time, but very well some of the time. The method may still be useful, provided that one can tell when it is going to work and when it will not. Global measures of forecast accuracy may not be sensitive enough to identify these occasional successes.

Having said all of that, we will continue to focus on the RMSE measure of forecast performance. We will follow the usual practice of measuring forecast performance by normalized RMSE.

The second and much more difficult problem in forecast evaluation is: how much better does the forecast have to be before the improvement is "significant"? The term *significant improvement* is thrown about quite frequently in papers on forecasting, and almost all of the time the use of the term is not justified. One step toward turning a heuristic procedure into a statistical test is to utilize some recently developed tools for comparing the accuracy of competing forecast methods.

A recent paper by Mizrach (1992) has derived a statistic to test the hypothesis that one forecast has a smaller mean square error than another. The test is a refinement of an approach due to Granger and Newbold (1986) and Meese and Rogoff (1988). To review the test, let $\{\varepsilon_{1,t}\}$ be the forecast residuals from method one, and let $\{\varepsilon_{2,t}\}$ be the forecast residuals from method two.

Granger and Newbold test whether $\text{var}(\varepsilon_{1,t}) = \sigma_1^2$ is significantly less than $\text{var}(\varepsilon_{2,t}) = \sigma_2^2$ by looking at the orthogonalized residuals $u_t = \varepsilon_{1,t} - \varepsilon_{2,t}$, $v_t = \varepsilon_{1,t} + \varepsilon_{2,t}$. The correlation between u_t and v_t is just $\sigma_1^2 - \sigma_2^2$. Thus, if the correlation between u_t and v_t is significantly greater (less) than zero, then σ_1^2 is significantly greater (less) than σ_2^2.

Mizrach's refinement is to generalize the test to biased, heteroscedastic forecast residuals, using Newey-West (1987) hardware. Mizrach shows that given that the two sequences are α mixing, the statistic is distributed standard normal asymptotically

$$R = \sqrt{T}\left(\frac{\frac{1}{T}\sum_{i=1}^{T} u_i v_i}{\left[\sum_{i=k}^{k} \text{w}(i)\, S_{uvuv}(i)\right]^{1/2}}\right) \tag{14}$$

where

$$w(i) = 1 - \frac{|i|}{k+1}$$

$$S_{uvuv}(i) = \frac{1}{T-i}\sum_{t=i+1}^{T} u_t v_t u_{t-i} v_{t-i}, \qquad i \geq 0$$

$$= \frac{1}{T+i}\sum_{t=-i+1}^{T} u_{i+t} v_{i+t} u_t v_t, \qquad i < 0$$

$$k = k(T), \text{ with } \lim_{T\to\infty} \frac{k(T)}{T^{1/2}} = 0, \qquad = \text{int}(T^{1/3}) + 1,$$

for example.

The α-mixing assumption is worth explaining. A sequence of random variables is independently distributed if no single observation or group of observations contains any information that can be used to forecast any aspect of any other observation in the sequence. A sequence of random variables is α mixing if (loosely speaking) the degree of dependence between observations declines down the sequence. That is, while information in observation x_t (say) can be useful in predicting some aspect of x_{t+1} (say), observation x_t contains very little information about x_{t+s} for sufficiently large s. This is essentially an assumption that anomalies or quirks in the data have only a transitory impact.

To use an economic example, it is essentially the assumption that anything that happened to inflation in (say) 1950 has very little effect on the inflation rate of (say) 1990. It is essentially the assumption that there is little long-run persistence in our data.

We now have a procedure to test the hypothesis that method one (RMSE) generates more accurate forecasts on a given data set than method two (MAD). The test involves applying the Mizrach statistic to test the null hypothesis that the MSE for forecast method number one is equal to the MSE for forecast method number two. If the test statistic is sufficiently large in absolute value, we reject the hypothesis that the two forecasts are of equal accuracy, and thus we have a winner in the forecasting competition.

However, there is a conceptual difficulty with this approach that we should not pass over without comment. The difficulty has to do with certain inescapable problems with economic data. To explain: suppose that we have two competing methods of forecasting, method one and method two. How should one decide whether forecast method one is better than forecast method two? To correctly answer this question, one would like to have access to repeated sampling from the data generator, and then look at the distribution of one's preferred measure of forecast performance for each method. The answer would depend on the expected values of the measures of forecast performance for a typical time series from the data generator.

In principle, one could do studies of the statistical size of the test under the null hypothesis of no nonlinearity, and the statistical power of the test at detecting various types of nonlinearities. This has not yet been done, and I would venture to guess that one reason for the lack of results is that the near neighbor forecasting methodology is very computationally intensive. Calculating the near neighbor forecast for a prediction set of 500 observations and a fitting set of 1,500 observations at an embedding dimension of twenty could easily take a day on a personal computer based on the Intel 80486 chip running at 33 MHz. A convincing Monte Carlo study could require 1,000 iterations for several different data generators. However, given enough computing power, we could conduct such a study of the reliability of the method on simulated data sets.

On the other hand, the major implication of the forecasting paradox is that we do not yet have reliable models of economic phenomena. Economic and financial data exhibit exotic dependencies that are often very difficult to model. There are, for example, controversies over the appropriate model for volatility clustering in economic data. A number of approaches have been tried, including Auto-Regressive Conditional Heteroscedasticity (ARCH) models, unconditional mixture models, and regime-shifting models. No single model seems to dominate the others for all applications. Worse, even if

we confine our attention to a single class of models, one must still demonstrate that the parameterization chosen for the simulation is sensible.

These difficulties all occur because, unfortunately, economics is not (for the most part) an experimental science. Economists (particularly macroeconomists) typically observe only a single, short realization of their data-generating process. A time series collected monthly since World War II (like the Consumer Price Index) has fewer than 600 observations. Data sets collected quarterly (like GNP figures) have less than 200 observations.[1] We cannot do controlled experiments to assess how the economy would respond to controlled shocks. This means that there is, in general, no way to determine whether the economic data we observe are typical or representative of the underlying process. In terms of the prediction test, this means that the best we can expect is for the test to tell us whether method one outforecast method two on the given sample. We have no way to determine how the methods would compare on alternative samples from the underlying data generator. Therefore, assessing the reliability of our measure of forecast performance on real data is a deep problem. These are important caveats that must be kept in mind when assessing our test results.

An Illustrative Example: The Inflation Rate

To illustrate the use of these techniques, we apply them to a data set generated from monthly observations on the Consumer Price Index for the period January 1947 to February 1993. I take the first differences of the logs of the CPI, which results in the monthly inflation rate.

Setting aside the last 20 percent of the data, we estimate an AR(12) model on the first 442 observations. For the observations s in our filling set, we obtain estimates of the parameters in the regression

$$y_s = \alpha + \sum_{i=1}^{12} y_{s-1}\, \beta_{s-1} + \varepsilon_s. \tag{15}$$

The lag length, while plausible, is selected for convenience. In sample fit of the model is quite good. The R^2 for the regression is 0.56, indicating that the regression explains a sizable fraction of the total variation of the process. This R^2 would correspond to an in-sample RMSE of 0.44. In the regression, the standard measures of parameter variability indicate that many of the coefficients are significantly different from zero at the 95 percent confidence level. Summary statistics for the regression are given in table 4.2.

Moving to out-of-sample forecasting, we use the estimated coefficients from the AR(12) to forecast the rest of the series. For observations of y_t in our prediction set, and their corresponding histories, we calculate forecasts using

TABLE 4.2. Summary of Regression Results for the Inflation Rate Regression

Coefficient	Value	Test Statistic
Constant	0.00046	2.13
X_{t-1}	0.36696	7.50
X_{t-2}	0.13934	2.69
X_{t-3}	0.13432	2.59
X_{t-4}	0.04699	0.93
X_{t-5}	−0.05750	−1.19
X_{t-6}	−0.05846	−1.21
X_{t-7}	0.05797	1.20
X_{t-8}	0.09294	1.91
X_{t-9}	0.12793	2.66
X_{t-10}	0.10052	2.19
X_{t-11}	−0.07539	−1.65
X_{t-12}	−0.02392	−0.54
Regression R^2		0.558
Fitting Set Mean Forecast NRMSE		1.011
Regression Forecast RMSE		1.005
Number of Observations		442

Note: The test statistic is the standard OLS test statistic for the hypothesis that the corresponding coefficient is significantly different from zero.

the OLS estimates (a, b) of (α, β) from equation 15 above to form the forecasts

$$\hat{y}_t = a + \sum_{i=1}^{12} y_{t-i}\, b_{t-i} + \varepsilon_t. \tag{16}$$

These results are an illustration of the forecasting paradox. Even though the model fits reasonably well in-sample, the out-of-sample forecasts are spectacularly inaccurate.The normalized RMSE for the fitting set is 1.005. To compare, if we just use the unconditional mean of the fitting set as our predictor, our normalized RMSE is 1.011. The regression estimate offers a reduction in RMSE of less than one-half of 1 percent out-of-sample, over the unconditional mean predictor. The conclusion is that the global linear AR(12) model is a very poor representation of the underlying data-generating process.

One then might cast about for a reason why performance is so poor. A useful start would be to estimate the dimension of the inflation data, to look

for evidence that there may be a low-dimensional attractor underlying the data. Table 4.3 gives the estimated dimension of the process, calculated at varying lag lengths, following the procedure given in Brock and Baek 1991. To remove spurious linear dependence in the data, we calculate the estimates of dimension from the residuals of the AR model. Also in table 4.3, we have two other common measures of nonlinear structure, the BDS statistic (Brock, Dechert, and Scheinkman 1986) and the Kolmogorov Entropy (calculated following the method of Brock and Baek 1991). All three measures of dependence indicate significant evidence of nonlinear structure in the residuals of the linear model.

To serve as a comparison, we repeat the procedure on simulated data from noisy chaos. We generated 600 observations from the Henon process (500 from the fitting set and 100 from the prediction set), which is a simple, low-dimensional chaos (Henon 1976). We "polluted" the Henon data with normally distributed pseudorandom numbers with a standard deviation of 50 percent of the standard deviation of the Henon data. This corresponds to a noise-to-signal ratio of 50 percent, which is fairly high by the standards of physical systems. When we estimate a linear model to the Henon data, we observe results that are somewhat similar to the results we obtained for the CPI data. In-sample fit is acceptable, with a regression R^2 of about 0.202. (This corresponds to an in-sample RMSE of 0.798.) Out-of-sample forecast performance is worse than the in-sample fit, with an out-of-sample nor-

TABLE 4.3. Nonlinearity Diagnostics for the Regression Residuals from Inflation Rate Data

			$\varepsilon = 0.81$			$\varepsilon = 0.9$		
Embedding Dimension	Estimated Dimension	Test Statistic	BDS	Entropy	Test Statistic	BDS	Entropy	Test Statistic
2	1.569	−2.47	5.71	0.584	−2.13	6.05	0.506	−2.15
3	2.312	−2.51	6.66	0.580	−2.11	7.01	0.502	−2.13
4	3.049	−2.47	6.75	0.569	−2.19	7.08	0.490	−2.24
5	3.794	−2.39	6.88	0.547	−2.39	7.22	0.467	−2.50
6	4.553	−2.27	7.25	0.553	−2.19	7.65	0.476	−2.26
7	5.284	−2.19	7.47	0.550	−2.12	7.85	0.480	−2.09
8	5.945	−2.17	7.72	—	—	8.00	—	—

Note: The dimension of the process is estimated as the elasticity of the correlation integral with respect to the link scale ε. Under the null of independent and identically distributed data (IID), the expected value of this statistic is equal to the embedding dimension. The test statistic should be distributed as a standard normal random variable under the null of IID. The entropy statistic is a measure of sensitive dependence to initial conditions. Under the null of IID, the expected value of the entropy is minus the log of the correlation integral at embedding dimension 1. The test statistic should be distributed as a standard normal under the null. The BDS test is a portmanteau test of IID. Again, the test statistic should be distributed as a standard normal under the null of IID.

malized RMSE of 0.913. This is still a bit better than the RMSE of the unconditional mean forecast, which is 0.996. When we estimate the BDS, dimension, and entropy numbers for the residuals of the linear model applied to the Henon data, we observe results that indicate that, if anything, the Henon residuals exhibit less nonlinear dependence than the inflation data. (These results are summarized in tables 4.4 and 4.5.)

When we forecast the Henon data using the near neighbor method, we obtain substantial forecast improvements. The RMSE plot for the polluted Henon data is given in figure 4.5. The RMSE plot is minimized at $k = 51$. Regressions based on 51 near neighbors yield normalized RMSE of 0.78. Note that this RMSE is about 15 percent lower than the RMSE for forecasts based on the linear model. When we apply Mizrach's test of forecast performance we observe a test statistic of 2.13, which is significant at the 95 percent level of confidence. In this case, the nonlinear forecast algorithm generates predictions that are significantly more accurate than the forecasts generated by the linear model. This is evidence that there is significant nonlinear determinism underlying these data.

TABLE 4.4. Summary of Regression Results for the Noisy Henon Data Regression

Coefficient	Value	Test Statistic
Constant	0.399	5.63
X_{t-1}	−0.120	−2.45
X_{t-2}	0.094	1.92
X_{t-3}	−0.287	−5.84
X_{t-4}	−0.059	−1.16
X_{t-5}	−0.107	−2.12
X_{t-6}	0.001	0.01
X_{t-7}	−0.006	−0.12
X_{t-8}	−0.084	−1.66
X_{t-9}	0.095	1.87
X_{t-10}	−0.012	−0.25
X_{t-11}	0.042	0.85
X_{t-12}	0.038	0.78
Regression R^2		0.798
Fitting Set Mean Forecast NRMSE		0.996
Regression Forecast RMSE		0.913
Number of Observations		500

Note: The test statistic is the standard OLS test statistic for the hypothesis that the corresponding coefficient is significantly different from zero.

TABLE 4.5. Nonlinearity Diagnostics for the Regression Residuals from Noisy Henon Map Data

			$\varepsilon = 0.81$			$\varepsilon = 0.9$		
Embedding Dimension	Estimated Dimension	Test Statistic	BDS	Entropy	Test Statistic	BDS	Entropy	Test Statistic
2	1.728	−1.39	4.52	0.796	−0.60	4.69	0.701	−0.68
3	2.627	−1.22	3.17	0.782	−0.84	3.45	0.690	−0.87
4	3.499	−1.19	2.90	0.787	−0.70	3.14	0.706	−0.53
5	4.269	−1.33	2.57	0.771	−0.96	2.60	0.696	−0.71
6	4.984	−1.48	2.54	0.762	−1.06	2.40	0.688	−0.81
7	5.687	−1.57	2.57	0.705	−1.90	2.32	0.650	−1.40
8	6.204	−1.80	3.07	—	—	2.57	—	—

Note: The dimension of the process is estimated as the elasticity of the correlation integral with respect to the link scale ε. Under the null of independent and identically distributed data (IID), the expected value of this statistic is equal to the embedding dimension. The test statistic should be distributed as a standard normal random variable under the null of IID. The entropy statistic is a measure of sensitive dependence to initial conditions. Under the null of IID, the expected value of the entropy is minus the log of the correlation integral at embedding dimension 1. The test statistic should be distributed as a standard normal under the null. The BDS test is a portmanteau test of IID. Again, the test statistic should be distributed as a standard normal under the null of IID.

Compared to these results, the results for the inflation data are rather disappointing. The RMSE plot for the inflation data is given in figure 4.6. The NRMSE is minimized at $k = 85$, where NRMSE = 0.975. This is approximately a 2.5 percent improvement over the NRMSE for the AR(12) model. We check whether this improvement in forecast performance is signifi-

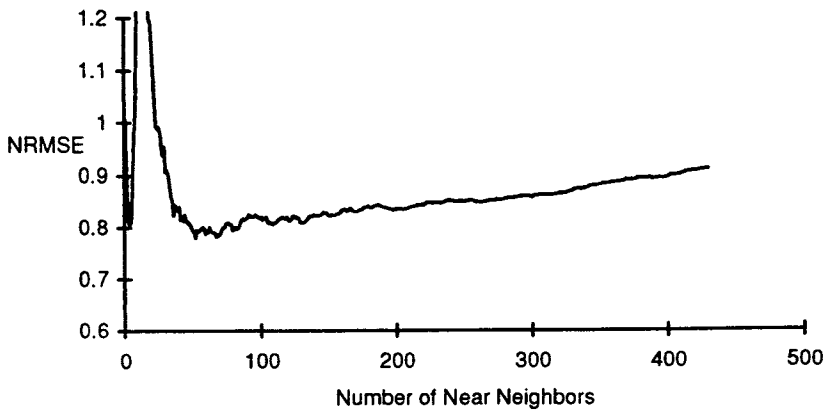


Fig. 4.5. NRMSE plot for the near neighbor forecasts of the noisy Henon data

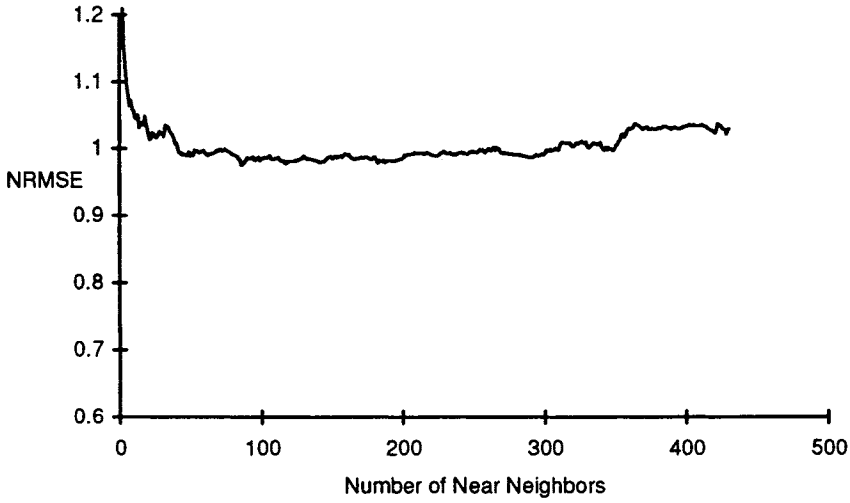


Fig. 4.6. NRMSE plot for the near neighbor forecasts of the inflation data

cant, using the Mizrach test. The test statistic for this comparison is 1.77. The improvement in forecast performance is not statistically significant by the usual criteria.

The lack of forecast improvement is especially troubling when we consider the computational burden of the near neighbor approach. The AR(12) model fits the data with 13 parameters, counting the regression intercept. For a fitting set of 500 and a prediction set of 100 observations, the near neighbor methodology requires on the order of $500 \cdot 100 = 50{,}000$ regressions, and the calculation of 750,000 parameters, all to yield an insignificant forecast improvement.

Discussion

The prediction test is a very enticing approach to testing for nonlinear determinism in a data set. For data from systems known to be chaotic, near neighbor techniques offer spectacular improvements in forecast performance, relative to the performance of the global linear predictors favored by economists. The near neighbor methodology is also useful for predicting noisy chaos. These developments are well documented in Casdagli (1992) and references therein.

Numerous authors have applied near neighbor algorithms to economic data sets, with no significant forecasting success in evidence in the literature.

Recent efforts include Diebold and Nason 1990, LeBaron 1992, Meese and Rose 1990, Mizrach 1992b, and Jaditz and Sayers 1993b. When the method is applied to economic data, the forecast improvements are generally not statistically significant. Even when there appears to be ample evidence of nonlinear dependence in a data set, nonlinear methods are unable to exploit the dependence to offer improved forecast performance. The question is, why does this occur?

In our previous discussions, we have touched upon some possible reasons why we may be failing to identify nonlinearities. One could question our measure of forecast performance or our method for testing whether our forecast improvement is significant. Other explanations are also possible. For example, the data set available may be too short to allow us to conclusively identify whether nonlinear determinism is present. It would be very difficult to detect the presence of a moderately high-dimensional process in samples as short as the one at hand. High-dimensional chaos may be observationally equivalent to random behavior in time series this short.

However, one could just as easily reject the hypothesis that economic data generators are chaotic, in favor of stochastic alternatives. In their study of near neighbor forecasting of current exchange rates, Diebold and Nason make two points on the apparent paradox of dependence and lack of forecastability. The first point is that in economic time series, the conditional variance tends to change over time. The standard tests for nonlinear determinism are known to be sensitive enough to pick up this dependence. A discussion of this point is carried out in Brock, Hsieh, and LeBaron 1991 in the context of the BDS statistic. While volatility clustering is easily detected, dependence in variance is not useful for the prediction of the mean. Thus there may be substantial dependence in the time series that cannot be exploited to improve level prediction.

A second possible explanation for the poor forecast performance is that tests for nonlinear dependence may be picking up regime shifts, statistical outliers, or other underlying structural instability. Many economic time series are subject to regime shifts. For example, in our inflation data there is generally conceded to be a regime shift occurring about 1980, when the Federal Reserve Board switched from a policy of pegging nominal interest rates to a policy of pegging money supply growth. Occasional regime shifts or low-probability outliers may be of limited utility for short-horizon forecasting by any pure time-series method.

Hence we reach a paradox. Conventional in-sample measures of nonlinear dependence suggest that economic data exhibit substantial nonlinear structure. Forecasting algorithms seem to be unable to utilize any of that structure to improve forecasting. Hence, claims of nonlinear determinism in economic data are still controversial.

The prediction test does, however, offer hope for a possible resolution of the dilemma. If nonlinear methods generate forecast improvements in economic time series on the order of the improvements observed in the forecasts of physical systems, the case for nonlinear determinism will be irrefutable. Such improvement has not yet been demonstrated. It is certainly worthwhile to continue the search.

NOTES

1. It is possible to argue that the problem is less severe in financial economics. Tick-by-tick price and volume of sales data are available for thousands of stocks, bonds, and related financial instruments. On the other hand, these many time series are not independent observations from the same data-generating process, nor are they independent in cross section. Thus, many of the same problems remain.