# ECONOMIC PREDICTION USING NEURAL NETWORKS: THE CASE OF IBM DAILY STOCK RETURNS

by

Halbert White
Department of Economics
University of California, San Diego

## ABSTRACT

This paper reports some results of an on-going project using neural network modelling and learning techniques to search for and decode nonlinear regularities in asset price movements. We focus here on the case of IBM common stock daily returns. Having to deal with the salient features of economic data highlights the role to be played by statistical inference and requires modifications to standard learning techniques which may prove useful in other contexts.

## I. INTRODUCTION

The value of neural network modelling techniques in performing complicated pattern recognition and nonlinear forecasting tasks has now been demonstrated across an impressive spectrum of applications. Two particularly interesting recent examples are those of Lapedes and Farber who in [1987a] apply neural networks to decoding genetic protein sequences, and in [1987b] demonstrate that neural networks are capable of decoding deterministic chaos. Given these successes, it is natural to ask whether such techniques can be of use in extracting nonlinear regularities from economic time series. Not surprisingly, especially strong interest attaches to the possibility of decoding previously undetected regularities in asset price movements, such as the minute-to-minute or day-to-day fluctuations of common stock prices. Such regularities, if found, could be the key to great wealth.

Against the optimistic hope that neural network methods can unlock the mysteries of the stock market is the pessimistic received wisdom (at least among academics) of the "efficient markets hypothesis." In its simplest form, this hypothesis asserts that asset prices follow a random walk (e.g. Malkiel [1985]). That is, apart from a possible constant expected appreciation (a risk-free return plus a premium for holding a risky asset), the movement of an asset's price is completely unpredictable from publicly available information such as the price and volume history for the asset itself or that of any other asset. (Note that predictability from publicly unavailable (insider) information is not ruled out.) The justification for the absence of predictability is akin to the reason that there are so few $100 bills lying on the ground. Apart from the fact that they aren't often dropped, they tend to be picked up very rapidly. The same is held to be true of predictable profit opportunities in asset markets: they are exploited as soon as they arise. In the case of a strongly expected price increase, market participants go long (buy), driving up the price to its expected level, thus quickly wiping out the profit opportunity which existed only moments ago. Given the human and financial resources devoted to the attempt to detect and exploit such opportunities, the efficient markets hypothesis is indeed an attractive one. It also appears to be one of the few well documented empirical successes of modern economic theory. Numerous studies have found little evidence against the simple efficient markets hypothesis just described, although mixed results have been obtained using some of its more sophisticated variants

(see e.g. Baillie [1986], Lo and MacKinley [1988], Malkiel [1985] and Shiller [1981]).

Despite the strength of the simple efficient markets hypothesis, it is still only a theory, and any theory can be refuted with appropriate evidence. It may be that techniques capable of finding such evidence have not yet been applied. Furthermore, the theory is realistically mitigated by bounded rationality arguments (Simon [1955, 1982]). Such arguments hold that humans are inherently limited in their ability to process information, so that efficiency can hold only to the limits of human information processing. If a new technology (such as neural network methods) suddenly becomes available for processing available information, then profit opportunities to the possessor of that technology may arise. The technology effectively allows creation of a form of inside information. However, the efficient markets hypothesis implies that as the new technology becomes publicly available, these advantages will dwindle (rapidly) and ultimately disappear.

In view of the relative novelty of neural network methods and the implications of bounded rationality, it is at least conceivable that previously undetected regularities exist in historical asset price data, and that such regularities may yet persist. The purpose of this paper is to illustrate how the search for such regularities using neural network methods might proceed, using the case of IBM daily common stock returns as an example. The necessity of dealing with the salient features of economic time series highlights the role to be played by methods of statistical inference and also requires modifications of neural network learning methods which may prove useful in general contexts.

## II. DATA, MODELS, METHODS AND RESULTS

The target variable of interest in the present study is $r_t$, the one day rate of return to holding IBM common stock on day $t$, as reported in the Center for Research in Security Price's security price data file ("the CRSP file"). The one day return is defined as $r_t = (p_t - p_{t-1} + d_t)/p_{t-1}$, where $p_t$ is the closing price on day $t$ and $d_t$ is the dividend paid on day $t$. The one-day return $r_t$ is also adjusted for stock splits if any. Of the available 5000 days of returns data, we select a sample of 1000 days for training purposes, together with samples of 500 days before and after the training period which we use for evaluating whatever knowledge our networks have acquired. The training sample covers trading days during the period 1974:II through 1978:I. The evaluation periods cover 1972:II through 1974:I and 1978:II through 1980:I. The training set is depicted in Figure 1.

Stated formally, the simple efficient markets hypothesis asserts that $E(r_t \mid I_{t-1}) = r^*$, where $E(r_t \mid I_{t-1})$ denotes the conditional expectation of $r_t$ given publicly available information at time $t-1$, $I_{t-1}$ (formally $I_{t-1}$ is the $\sigma$–field generated by publicly available information), and $r^*$ is a constant (which may be unknown) consisting of the risk free return plus a risk premium. Because $I_{t-1}$ includes the previous IBM price history, the force of the simple efficient markets hypothesis is that this history is of no use in forecasting $r_t$.

In the economics literature, a standard way of testing this form of the efficient markets hypothesis begins by embedding it as a special case in a linear autoregressive model for asset returns of the form

$$r_t = w_o + w_1 r_{t-1} + \cdots + w_p r_{t-p} + \epsilon_t\,, \quad t = 1\,, 2\,, \ldots\,,$$

where $\underline{w} = (w_o\,,\ w_1\,, \ldots\,,\ w_p)'$ is an unknown column vector of weights, $p$ is a positive integer determining the order of the autoregression, and $\epsilon_t$ is a stochastic error assumed to be such that $E(\epsilon_t \mid I_{t-1}) = 0$.

The efficient markets hypothesis implies the restriction that $w_1 = \cdots = w_p = 0$. Thus, any empirical evidence that $w_1 \neq 0$ or $w_2 \neq 0$ ... or $w_p \neq 0$ is evidence against the efficient markets

hypothesis. On the other hand, empirical evidence that $w_1 = \cdots = w_p = 0$, while not refuting the efficient markets hypothesis, does not confirm it; numerous instances of deterministic nonlinear processes with no linear structure whatsoever are now well known (e.g. Sakai and Tokumaru [1980]; see also Eckmann and Ruelle [1985]). The finding that $w_1 = \cdots = w_p = 0$ is consistent with either the efficient markets hypothesis or the presence of linearly undetectable nonlinear regularities.

An equivalent implication of the simple efficient markets hypothesis that will primarily concern us here is that $var\ r_t = var\ \epsilon_t$, where *var* denotes the variance of the indicated random variable. Equivalently, $R^2 \equiv 1 - var\ \epsilon_t / var\ r_t = 0$ under the simple efficient market hypothesis. Thus, empirical evidence that $R^2 \neq 0$ is evidence against the simple efficient markets hypothesis, while empirical evidence that $R^2 = 0$ is consistent with either the efficient markets hypothesis or the existence of nonlinear structure.

Thus, as a first step, we examine the empirical evidence against the simple efficient markets hypothesis using the linear model posited above. The linear autoregressive model of order $p$ ($AR(p)$ model) corresponds to a very simple two layer linear feedforward network. Given inputs $r_{t-1}, \ldots, r_{t-p}$, the network output is given as $\hat{r}_t = \hat{w}_o + \hat{w}_1\, r_{t-1} + \cdots + \hat{w}_p\, r_{t-p}$, where $\hat{w}_o$, $\hat{w}_1, \ldots, \hat{w}_p$ are the network weights arrived at by a suitable learning procedure. Our interest then attaches to an empirical estimate of $R^2$, computed in the standard way (e.g. Theil [1971, p. 176]) as $\hat{R}^2 \equiv 1 - \hat{var}\ \epsilon_t / \hat{var}\ r_t$, where $\hat{var}\ \epsilon_t \equiv n^{-1} \sum_{t=1}^{n} (r_t - \hat{r}_t)^2$, $\hat{var}\ r_t \equiv n^{-1} \sum_{t=1}^{n} (r_t - \bar{r})^2$, $\bar{r} \equiv n^{-1} \sum_{t=1}^{n} r_t$, and $n$ is the number of training observations. Here $n = 1000$.

These quantities are readily determined once we have arrived at suitable values for the network weights. A variety of learning procedures is available. A common learning method for linear networks is the delta method

$$\underline{w}_{t+1} = \underline{w}_t - \eta\, \underline{x}_t' \,(r_t - \underline{x}_t\, \underline{w}_t) \qquad t = 1, \ldots, 1000$$

where $\underline{w}_t$ is the $(p+1) \times 1$ weight vector after presentation of $t-1$ target/input pairs, $\eta$ is the learning rate, and $\underline{x}_t$ is the $1 \times (p+1)$ vector of inputs $\underline{x}_t = (1, r_{t-1}, \ldots, r_{t-p})$. A major defect of this method is that because of the constant learning rate and the presence of a random component $\epsilon_t$ in $r_t$, this method will never converge to a useful set of weight values, but is doomed to wander eternally in the netherworld of suboptimality.

A theoretical solution to this problem lies in allowing $\eta$ to depend on $t$. As shown by White [1987a, b] an optimal choice is $\eta_t \propto t^{-1}$. Nevertheless, this method yields very slow convergence. A very satisfactory computational solution is to dispense with recursive learning methods altogether, and simply apply the method of ordinary least squares (OLS). This gives weights by solving the problem

$$\min_{\underline{w}} \sum_{t=1}^{n} (r_t - \underline{x}_t\, \underline{w})^2 .$$

The solution is given analytically as

$$\underline{w} = (X'X)^{-1}\, X'r ,$$

where $X$ is the $1000 \times (p+1)$ matrix with rows $\underline{x}_t$, $r$ is the $1000 \times 1$ vector with elements $r_t$, and the $-1$ superscript denotes matrix inversion.

Network learning by OLS is unlikely as a biological mechanism; however, our interest is not on learning per se, but on the results of learning. We are interested in the performance of "mature" networks. Furthermore, White [1987a, b] proves that as $n \to \infty$ both OLS and the delta method with

$\eta_t \propto t^{-1}$ converge stochastically to identical limits. Thus, nothing is lost and much computational effort is saved by using OLS.

When OLS is applied to the linear network with $p = 5$, we obtain $\hat{R}^2 = .0079$. By construction, $\hat{R}^2$ must lie between zero and one. The fact that $\hat{R}^2$ is so low suggests little evidence against the simple efficient markets hypothesis. In fact, under some statistical regularity conditions, $n\hat{R}^2$ is distributed approximately as $\chi^2_p$ when $w_1 = \cdots = w_p = 0$. In our case, $n\hat{R}^2 = 7.9$, so we have evidence against $w_1 = \cdots = w_p = 0$ at less than the 10% level, which is below usual levels considered to be statistically significant. The plot of $\hat{r}_t$ also reveals the virtual absence of any relation between $\hat{r}_t$ and $r_t$. (See Figure 2.)

Thus, standard methods yield standard conclusions, although nonlinear regularities are not ruled out. To investigate the possibility that neural network methods can detect nonlinear regularities inconsistent with the simple efficient markets hypothesis, we trained a three layer feedforward network with the same five inputs and five hidden units over the same training period. The choice of five hidden units is not entirely ad hoc, as it represents a compromise between the necessity to include enough hidden units so that at least simple nonlinear regularities can be detected by the network (Lapedes and Farber [1987b] detected the deterministic chaos of the logistic map using five hidden units with tanh squashing functions; we use logistic squashes, but performance in that case at least is comparable, even with only three or even two hidden units) and the necessity to avoid including so many hidden units that the network is capable of "memorizing" the entire training sequence. It is our view that this latter requirement is extremely important if one wishes to obtain a network which has any hope at all of being able to generalize adequately in an environment in which the output is not some exact function of the input, but exhibits random variation around some average value determined by the inputs. Recent results in the statistics literature for the method of sieves (e.g. Grenander [1981], Geman and Hwang [1982]) suggest that with a fixed number of inputs and outputs, the number of hidden units should grow only as some small power of the number of training observations. Over-elaborate networks are capable of data-mining as enthusiastically as any young graduate student.

The network architecture used in the present exercise is the standard single hidden layer architecture, with inputs $\underline{x}_t$ passed to a hidden layer (with full interconnections) and then with hidden layer activations passed to the output unit. Our analysis was conducted with and without a logistic squash at the output; results were comparable, so we discuss the results without an output squash.

The output of this network is given by

$$\tilde{r}_t = \hat{\beta}_o + \sum_{j=1}^{5} \Psi(\underline{x}_t \, \hat{\underline{\gamma}}_j) \, \hat{\beta}_j \equiv f(\underline{x}_t \, , \hat{\underline{\theta}})$$

where $(\hat{\beta}_o \, , \hat{\beta}_1 \, , \ldots , \hat{\beta}_5)$ are a bias and weights from the hidden units to the output and $\hat{\underline{\gamma}} \equiv (\hat{\underline{\gamma}}_1 \, , \ldots , \hat{\underline{\gamma}}_5)$ are weights from the input units, both after a suitable training procedure; and $\Psi$ is the logistic squashing function. The function $f$ summarizes the dependence of the output on the input $\underline{x}_t$ and the vector of all connection strengths, $\hat{\underline{\theta}}$.

As with the preceding linear network, the efficient markets hypothesis implies that $\tilde{R}^2 \equiv 1 - \hat{var}\, \tilde{\epsilon}_t / \hat{var}\, r_t$ should be approximately zero, where now $\hat{var}\, \tilde{\epsilon}_t \equiv n^{-1} \sum_{t=1}^{n} (r_t - \tilde{r}_t)^2$ and $\hat{var}\, r_t = n^{-1} \sum_{t=1}^{n} (r_t - \bar{r})^2$ as before. This result will be associated with values for $\hat{\beta}_1 \, , \ldots , \hat{\beta}_5$ close to zero, and random values for $\hat{\underline{\gamma}}_j$. A value for $\tilde{R}^2$ close to zero will reflect the inability of the network to extract nonlinear regularities from the training set.

As with the linear network, a variety of training procedures is available. One popular method is the method of back propagation (Parker [1982], Rumelhart et. al. [1986]). In our notation, it can be

represented as

$$\underline{\theta}_{t+1} = \underline{\theta}_t - \eta_t \nabla_{\theta} f(\underline{x}_t\ ,\ \underline{\theta}_t)'\ (r_t - f(\underline{x}_t\ ,\ \underline{\theta}_t))\ ,$$

where $\underline{\theta}_t$ is the vector of all connection strengths after $t-1$ training observations have been presented, $\eta_t$ is the learning rate (now explicitly dependent on $t$) $\nabla_{\theta}$ represents the gradient with respect to $\underline{\theta}$ (a row vector) and the other notation is as before.

Back propagation shares the drawbacks of the delta method previously discussed. With $\eta_t$ a constant, it fails to converge, while with $\eta_t \propto t^{-1}$, it converges (in theory) to a local minimum. Unfortunately, the random component of $r_t$ renders convergence extremely difficult to obtain in practice. In fact, running on an IBM RT at well over 4 mips, convergence was not achieved after 36 hours of computation.

Rather quick convergence was obtained using a variant of the method of nonlinear least squares described in White [1988]. The method of nonlinear least squares (NLS) uses standard iterative numerical methods such as Newton-Raphson and Davidson-Fletcher-Powell (see e.g. Dennis [1983]) to solve the problem

$$\min_{\underline{\theta}} \sum_{t=1}^{n} (r_t - f(\underline{x}_t\ ,\ \underline{\theta}))^2\ .$$

Under general condition, both NLS and back-propagation with $\eta_t \propto t^{-1}$ convergence stochastically to the same limit, as shown by White [1987a, b].

Our nonlinear least squares method yields connection strengths $\hat{\theta}$ which imply $\tilde{R}^2 = .175$. At least superficially, this is a surprisingly good fit, apparently inconsistent with the efficient markets hypothesis and consistent with the presence of nonlinear regularities. Furthermore, the plot of fitted $(\tilde{r}_t)$ values shows some very impressive hits. (See Figure 3.)

If for the moment we imagine that $\hat{\underline{\gamma}}$ is given, and not the result of an optimization procedure, then $n\tilde{R}^2 = 175$ is $\chi^2_6$ under the simple efficient markets hypothesis, a highly significant result by any standards. Unfortunately, $\hat{\underline{\gamma}}$ is the result of an optimization procedure, not given *a priori*. For this reason $n\tilde{R}^2$ is in fact not $\chi^2_6$; indeed, its distribution is a complicated non-standard distribution. The present situation is similar to that considered by Davies [1977, 1987] in which certain parameters ($\underline{\gamma}$ here) are not identified under the null hypothesis. A theory applicable in the present context has not yet been developed and constitutes an important area for further research.

Given the unknown distribution for $n\tilde{R}^2$, we must be cautious in claiming that the simple efficient markets hypothesis has been statistically refuted. We need further evidence. One way to obtain this evidence is to conduct out-of-sample forecasting experiments. Under the efficient markets hypothesis, the out-of-sample correlation between $r_t$ and $\tilde{r}_t$ (or $\hat{r}_t$), where $\tilde{r}_t(\hat{r}_t)$ is computed using weights determined during the training (sample) period and inputs from the evaluation (out-of-sample) period, should be close to zero. If, contrary to the simple efficient markets hypothesis, our three layer network has detected nonlinear structure, we should observe significant positive correlation between $r_t$ and $\tilde{r}_t$.

This exercise was carried out for a post sample period of 500 days, and a pre-sample period of 500 days. For the post-sample period we observe a correlation of -.0699; for the pre-sample period, it is .0751 (for comparison, the linear model gives post-sample correlation of -.207 and pre-sample correlation of .0996). Such results do not constitute convincing statistical evidence against the efficient markets hypothesis. The in-sample (training period) results are now seen to be over-optimistic, being either the result of over-fitting (random fluctuations recognized incorrectly as nonlinearities) or of

learning evanescent features (features which are indeed present during the training period, but which subsequently disappear). In either case the implication is the same: the present neural network is not a money machine.

## III. CONCLUDING REMARKS

Although some might be disappointed by the failure of the simple network considered here to find evidence against the simple efficient markets hypothesis, the present exercise suggests some valuable insights: (1) finding evidence against efficient markets with such simple networks is not going to be easy; (2) even simple networks are capable of misleadingly overfitting an asset price series with as many as 1,000 observations; (3) on the positive side, such simple networks are capable of extremely rich dynamic behavior, as evidenced by time-series plots of $\tilde{r}_t$ (Figure 3).

The present exercise yields practical benefits by fostering the development of computationally efficient methods for obtaining mature networks (White [1988]). It also highlights the role to be played by statistical inference in evaluating the performance of neural network models, and in fact suggests some interesting new statistical problems (finding the distribution of $n\bar{R}^2$). Solution of the latter problem will yield statistical methods for deciding on the inclusion or exclusion of additional hidden units to a given network.

Of course, the scope of the present exercise is very limited; indeed, it is intended primarily as a vehicle for presenting the relevant issues in a relatively uncomplicated setting, and for illustrating relevant approaches. Expanding the scope of the search for evidence against the efficient markets hypothesis is a high priority. This can be done by elaborating the network to allow additional inputs (e.g., volume, other stock prices and volume, leading indicators, macroeconomic data, etc.) and by permitting recurrent connections of the sort discussed by Jordan [1986]. Any of these elaborations must be supported with massive infusions of data for the training period: the more connections, the greater the danger of overfitting. There may also be useful insights gained by permitting additional network outputs, for example, returns over several different horizons (two day, three day, etc.) or prices of other assets over several different horizons, as well as by using within rather than between day data.

Another important limitation of the present exercise is that the optimization methods used here are essentially local. Although the final weight values were determined as giving the best performance over a range of different starting values for our iterations, there is no guarantee that a global maximum was found. A global optimization method such as simulated annealing or the genetic algorithm would be preferable.

Finally, it is extremely important to point out that while the method of least squares (equivalently, back-propagation) is adequate for testing the efficient markets hypothesis, it is not necessarily the method that one should use if interest attaches to building a network for market trading purposes. Such networks should be evaluated and trained using profit and loss in dollars from generated trades, not squared forecast error. Learning methods for this criterion are under development by the author.

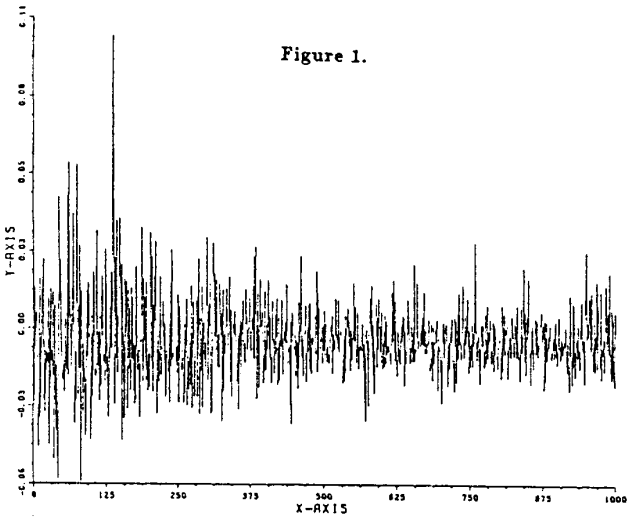


Figure 1.

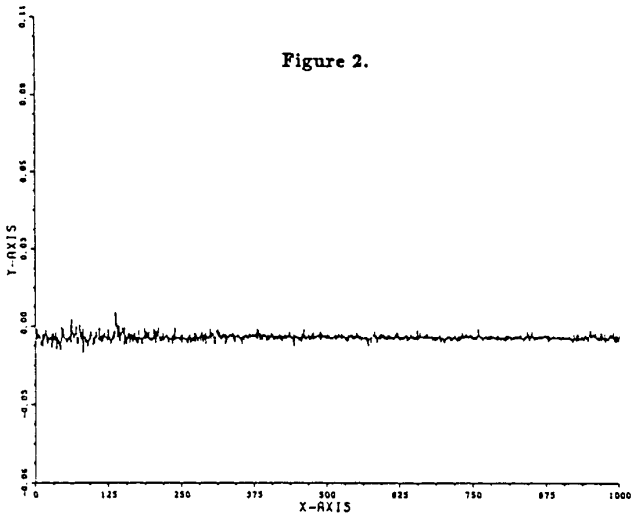


Figure 2.

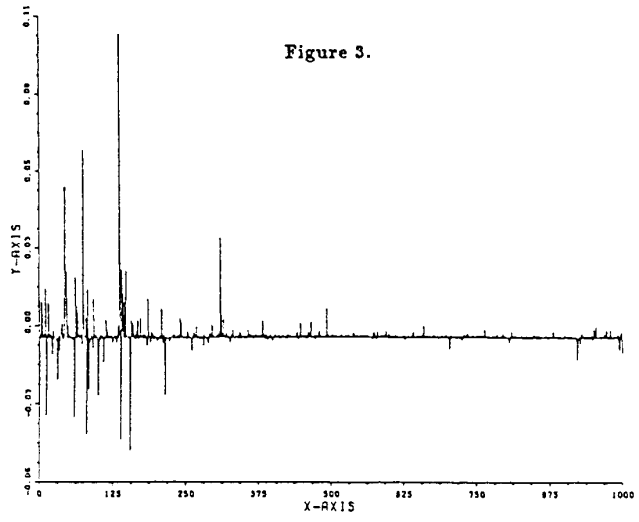


Figure 3.

## REFERENCES


Baillie, R.T. [1986]: "Econometric Tests of Rationality and Market Efficiency," Michigan State University Department of Economics Working Paper.

Davies, R.B. [1977]: "Hypothesis Testing When a Nuisance Parameter is Present Only Under the Alternative," *Biometrika* 64, 247-54.

Davies, R.B. [1987]: "Hypothesis Testing When a Nuisance Parameter is Present Only Under the Alternative," *Biometrika* 74, 33-43.

Dennis, J.E. [1983]: Numerical Methods for Unconstrained Optimization and Nonlinear Equations. Englewood Cliffs: Prentice-Hall.

Eckmann, J.-P. and D. Ruelle [1985]: "Ergodic Theory of Chaos and Strange Attractors," *Review of Modern Physics* 57, 617-656.

Geman, S. and C.H. Hwang [1982]: "Nonparametric Maximum Likelihood Estimation by the Method of Sieves," *Annals of Statistics* 70, 401-414.

Grenander, U. [1981]: *Abstract Inference.* New York: Wiley.

Jordan, M. [1986], "Serial Order: A Parallel Distributed Processing Approach," UCSD Institute of Cognitive Science Report 86-04.

Lapedes, A. and R. Farber [1987a]: "Genetic Data Base Analysis with Neural Nets," paper presented to the IEEE conference on Neural Information Processing Systems-Natural and Synthetic.

Lapedes, A. and R. Farber [1987b]: "Nonlinear Signal Processing Using Neural Networks," paper presented to the IEEE Conference on Neural Information Processing System-Natural and Synthetic.

Lo, A. and A.C. MacKinley [1988]: "Stock Market Prices do not Follow Random Walks: Evidence From a Simple Specification Test," *Review of Financial Studies* (forthcoming).

Malkiel, B.G. [1985]: *A Random Walk Down Wall Street.* New York: Norton.

Parker, D.B. [1982]: "Learning Logic," Invention Report, S81-64, File 1, Office of Technology Licensing, Stanford University.

Rumelhart, D.E., G.E. Hinton and R.J. Williams [1986]: "Learning Internal Representations by Error Propagation," in D.E. Rumelhart and J.L. McClelland eds., *Parallel Distributed Processing: Explorations in the Microstructures of Cognition*, Vol. 1, Cambridge: MIT Press, 318-362.

Sakai, H. and H. Tokumaru [1980]: "Autocorrelations of a Certain Chaos," *IEEE Transactions on Acoustics, Speech and Signal Processing* ASSP-28, 588-590.

Shiller, R.J. [1981]: "The Use of Volatility Measures in Assessing Market Efficiency," *Journal of Finance* 36, 291-304.

Simon, H. [1955]: "A Behavioral Model of Rational Choice," *Quarterly Journal of Economics* 69, 99-118.

Simon, H. [1982]: *Models of Bounded Rationality* (2 vols). Cambridge: MIT Press.

Theil, H. [1971]: *Principles of Econometrics.* New York: Wiley.

White, H. [1987a]: "Some Asymptotic Results for Learning in Single Hidden Layer Feedforward Network Models," UCSD Department of Economics Discussion Paper 87-13.

White, H. [1987b]: "Some Asymptotic Results for Back-Propagation," *Proceedings of the First Annual IEEE Conference on Neural Networks.*

White, H. [1988]: "A Performance Comparison for some On-Line and Off-Line Learning Methods for Single Hidden Layer Feedforward Nets," UCSD Department of Economics Discussion Paper.