

Nonlinear prediction of noisy time series with feedforward networks

Ramazan Gencay

Department of Economics, University of Windsor, Windsor, Ontario, Canada N9B 3P4

Received 15 September 1993; accepted for publication 24 February 1994

Communicated by A.R. Bishop

Abstract

The main focus of this study is the investigation of the noise filtering capabilities of the feedforward networks with small data sets. The first stage focuses on deterministic nonlinear time series estimation. The quality of the results with deterministic data will serve as a benchmark performance of the estimation techniques under study. The second stage involves the investigation of the out-of-sample performance of feedforward networks with noisy data sets. The noise component is investigated as a measurement noise. Basically, the in-sample and the out-of-sample mean square errors, sign predictions and the estimates of the Lyapunov exponents are used as the criteria of the fit and the quality of the forecasts. Although there has been some work done with feedforward networks within the context of nonlinear function approximation, the out-of-sample forecast capabilities of these are not yet investigated. This issue is addressed within the framework of the inverse problem. As compared to the other commonly used approximation techniques, the results of this study show that feedforward networks may prove to be an invaluable technique in the prediction of noisy time series data.

1. Introduction

The standard problem in dynamical system analysis involves the description of the asymptotic behaviour of the iterates of a given nonlinear system. The inverse problem, on the other hand, involves the construction of a nonlinear map from a sequence of its iterates. The constructed map can then be a candidate for a predictive model. Here, the inverse problem approach will be followed for estimation and prediction experiments. Consider a dynamical system, $f: \mathbb{R}^n \rightarrow \mathbb{R}^n$, with the trajectory

$$x_{t+1} = f(x_t), \quad t = 0, 1, 2, \dots \quad (1)$$

In practice, one rarely has the advantage of observing the true state of the system, let alone knowing the actual functional form, f , which generates the dynamical system.

The model that is widely used is the following: associated with the dynamical system in (1) there is a measurement function $h: \mathbb{R}^n \rightarrow \mathbb{R}$ which generates observations

$$y_t = h(x_t). \quad (2)$$

It is assumed that all that is available to us is the sequence $\{y_t\}$ to reconstruct f . Under certain regularity conditions, Takens' [1] embedding theorem indicates that this is feasible.

There are a variety of numerical techniques for solving the inverse problem such as Taylor series expansion, radial basis functions, nonparametric kernel and artificial neural networks. These techniques essentially involve interpolating or approximating unknown functions from scattered data points. The idea behind the Taylor series expansion is to increase

the order of the expansion to the point where a curved surface of that order can follow the curvature of the local data points closely. The trade off with this method is that the number of terms in a multidimensional Taylor series expansion increases quite rapidly with the order. Indeed, the number of parameters needed for Taylor series expansion of a given order grows multiplicatively as the order of the expansion is increased and this method involves a choice of an optimal order of expansion. Casdagli [2] points out that there are no known order of convergence results for $n > 1$, and polynomials of high degree have an undesirable tendency to oscillate wildly.

The nonparametric kernel estimation is a method for estimating the probability density functions from observed data. It is a generalization of histograms to continuously differentiable density estimators. The kernel density estimation involves the choice of a kernel function and a smoothing parameter. The idea behind this method is to determine the influence of each data point by placing a weight to each of the data points. The kernel function determines the shape of these weights and the window width determines their width. The approximation of an unknown function from data can be obtained by calculating the conditional mean of the regression function. Within the framework of the inverse problem, the kernel density estimator works well with a few number lags. However, as the number of lags gets larger the rate of convergence of the nonparametric kernel density estimator slows down considerably, which leads to the deterioration of the estimator of the conditional mean in finite samples. This deterioration gets worse in the partial derivatives of the conditional mean estimator.

Radial basis functions are related to the kernel density estimator such that in radial basis functions the contribution of each point is computed by least squares and these functions are easy to implement numerically. If standard algorithms for the solution of linear systems of equations are used, Casdagli [2] indicates that for large data sets, it becomes infeasible to implement in standard workstations.

Among the techniques mentioned above, the artificial neural networks are the least used technique in the inverse problem. This is partly due to the fact that the early learning algorithms for these networks such as backpropagation are very slow and computationally very expensive and the early developments do not

provide any guidance in terms of how to choose the number of neurons in a given layer and how many layers to construct in a given network. The recent developments in the artificial neural networks literature have provided the theoretical foundations for the universality of the feedforward networks as function approximators. The results of Refs. [3–6] indicate that feedforward networks with sufficiently many hidden units and properly adjusted parameters can approximate an arbitrary function arbitrarily well in useful spaces of functions. The results of Ref. [6] show that feedforward networks with as few as a single layer and an appropriately smooth hidden layer activation function are capable of arbitrarily accurate approximation to an arbitrary function and its derivatives. Hornik, Stinchcombe and White show that the conditions imposed on the hidden layer activation function are relatively mild. The requirement is that the activation function should be continuously differentiable with bounded derivatives. The first candidates which satisfy these conditions are the logistic and hyperbolic tangent squashers. A recent survey of this literature is presented in Ref. [7].

The first goal here is to utilize feedforward networks in the inverse problem for prediction purposes. One particular aspect of the problem is to analyse the performance of these networks with small data sets, as acquisition of large data sets in any field is difficult. The other aspect of our goal is to investigate how many data points are required for a reliable prediction.

The forecast experiments such as Refs. [2,8–10] have focused only on predicting the value of a map at a given point. The prediction errors of a map in almost all approximation techniques magnify in the predicted values of the partial derivatives of the map under study. The second goal here is to extend the analysis to the predicted partial derivatives of a map at given points. The accuracy of the forecast of the partial derivatives of a map is especially important in terms of the identification of the dynamical invariants of an underlying attractor such as Lyapunov exponents.

Basically, the in-sample and out-of-sample mean square errors as well as sign predictions are used as the criteria of the quality of the fit and the quality of the forecasts.

Feedforward networks are summarized briefly in

Section 2. Forecasting results are presented in Sections 3–5. Thereafter, concluding comments follow.

2. Feedforward networks

A rich class of nonlinear models studied in the artificial neural networks literature is the class of single hidden layer feedforward networks. Given inputs $x_t = (x_{t1}, \dots, x_{tr})'$, an output of a single layer feedforward network with q hidden units is written as

$$o_t = \beta_0 + \sum_{j=1}^q \beta_j h_{tj}, \quad h_{tj} = k\left(w_{j0} + \sum_{i=1}^r w_{ij} x_{ti}\right), \quad (3)$$

where $\beta = (\beta_0, \dots, \beta_q)'$, $w = (w'_1, \dots, w'_q)'$, $w_j = (w_{j0}, \dots, w_{jr})'$ are the parameters to be estimated and k is a known hidden unit activation function. In a feedforward network, hidden units are not dynamic as they do not depend on past values of h_j . For this reason, the network is called a feedforward network and it is a sum of simple univariate flexible functions.

A set of conditions under which the single layer feedforward networks are dense in general Sobolev spaces is analysed in Ref. [6]. The important part of its result is that both a function and its derivatives can be asymptotically approximated to any arbitrary degree of accuracy with a single layer feedforward network and with sufficiently many hidden units. Thus, the functions of the form (3) are in the sense of Gallant's [11] flexible Fourier form. Indeed, Gallant and White [12] show that a multiple input, single output, single hidden layer feedforward network with known connections from input to hidden layer, a sigmoid choice for h embeds as a special case a Fourier network which yields a Fourier series approximation to a given function as its output.

One important theoretical and practical issue is the degree of the accuracy of the approximation. How rapidly does the approximation to an arbitrary function improve as the number of hidden units q increases? Ref. [13] shows that the degree of approximation improves at root- q for continuously differentiable hidden unit functions. This part of the literature is still in progress and further results will provide insight into advantages and disadvantages of artificial network models to other flexible functional forms.

Another important question is how to decide on the number of hidden layers in a given feedforward network. For what classes of functions for instance, does a two hidden layer network with fewer parameters achieve a higher level of accuracy than a single layer feedforward network with more parameters? There are examples, such as in Ref. [14], which show that a single layer cannot exactly represent a class of piecewise constant functions exactly representable by a two layer network. Although the results of Ref. [6] carry over the multi-hidden layer feedforward networks, this area needs further research.

In Ref. [15] it is shown that feedforward networks can be used to consistently estimate both a function and its derivatives^{#1}. They show that the least squares estimates are consistent in Sobolev norm, provided that the number of hidden units increases with the size of the data set. This would mean that larger number of data points would require larger number of hidden units to avoid overfitting in noisy environments. Barron [16] has recently shown that for the least squares estimators of the conditional mean obtained through feedforward networks, the rate of convergence is slightly slower than $n^{1/2}$ for identically and independently distributed samples. Asymptotic distribution of neural network estimators has not been worked out and awaits for further research.

In this work the logistic function (which is a sigmoid^{#2} function)

$$k(x) = \frac{\beta}{1 + \exp(-wx)} \quad (4)$$

is used as the hidden layer activation function. The position and the slope of the curve are determined by w and the height of the function is determined by β . For small values of w the curve is more of a straight line whereas for large values of w the function is more like a step function. Skewed curves, sharp spikes, or bi-modal curves can be obtained by using various combinations of the parameters β , w and b . Some examples of these combinations are given in Ref. [17].

^{#1} A minimal property for any estimation procedure is that of consistency. A stochastic sequence $\{\theta_T\}$ is consistent for $\{\theta_0\}$ if the probability that $\{\theta_T\}$ exceeds any specified level of approximation error relative to $\{\theta_0\}$ tends to zero as the sample size T tends to infinity.

^{#2} k is sigmoid function if $k: \mathbb{R} \rightarrow [0, \beta]$, $k(a) \rightarrow 0$ as $a \rightarrow -\infty$, $k(a) \rightarrow \beta$ as $a \rightarrow \infty$ and k is monotonic.

For a single layer network, the least squares criterion for a data set of length T is

$$L(\beta, w) = \sum_{t=1}^T (y_t - \hat{o}_t)^2, \quad (5)$$

where \hat{o}_t is the output of the network at time t . This is a straightforward multivariate minimization problem. Conjugant gradient routines given in Ref. [18] work very well for this problem.

3. Deterministic system estimation

For simulation purposes the Hénon map will be used as it is a widely used example in the literature. The Hénon map is given as

$$x_{t+1} = 1 - 1.4x_t^2 + y_t, \quad y_{t+1} = 0.3x_t. \quad (6)$$

Note that there is only one nonlinear term, so the Hénon map is one of the simplest nonlinear maps in higher dimensions.

The in-sample and out-of-sample mean square errors are calculated by

$$\begin{aligned} \text{MSE}^s &= \frac{1}{T} \sum_{t=1}^T (y_t - \hat{o}_t)^2, \\ \text{MSE}^p &= \frac{1}{T} \sum_{t=1}^T (y_t - \hat{o}_t^p)^2, \end{aligned} \quad (7)$$

where \hat{o}_t is the output of a network at time t and \hat{o}_t^p is the predicted value of a network with true inputs and estimated network weights at time t . Here the true values of the inputs are used for one-step ahead prediction. In addition to the mean square error calculation, we calculate the sign predictions and report it as a percentage of correct signs obtained in in-sample and out-of-sample experiments.

One important question in neural network estimation is the choice of the network complexity. How many neurons should be placed in a given hidden layer and how many hidden layers are needed to build an arbitrarily accurate approximation of the dynamical system under study? Following the results of Ref. [6] a single layer feedforward network is constructed. Therefore, the task is reduced to choosing an appropriate number of hidden units in a single layer feedforward network. Either cross-validation or information-theoretic methods can be used to deter-

mine the optimal number of hidden units for a given sample. Information-theoretic methods in which one optimizes a complexity-penalized quasi-log likelihood, similar to the Schwartz information criterion (SIC) have been shown to have desirable properties by Barron [19]. Here the SIC is used to determine the network complexity.

A thousand observations are generated from the Hénon map. The first eight hundred are discarded as transients and the remaining two hundred are kept for estimation and prediction purposes. In the estimation, the complexity of the network is chosen by comparing the SICs of feedforward networks with different numbers of hidden units. The feedforward network with four hidden units provided the smallest SIC which led the choice of this network for estimation and prediction purposes. This network contains a total of 17 parameters to be estimated.

Initially, the Hénon map is estimated with 200 observations. In the subsequent steps the number of observations is reduced for in-sample estimation and the remaining observations are used for out-of-sample predictions. This experiment is done for sample sizes of 190, 180, 150 and 17. With the 17 observations, we just provide the number of degrees of freedom to the network to interpolate the data. This experiment is important in terms of observing the consistency of the least squares learning techniques by approximating a function and its derivatives with just enough information. The results are presented in Table 1, which provides the in-sample mean square errors (IMSEs), out-of-sample mean square errors (OMSEs) of the predictions, percentages of the predicted signs and the SICs. IMSEs of the Hénon map estimates are fairly small as they are no greater than 0.8592×10^{-6} . IMSEs of the estimated derivatives are also quite accurate as the largest MSE is 0.2688×10^{-3} and are presented in Table 2.

As the OMSEs and the predicted signs indicate, the feedforward network provides very satisfactory in-sample fits and one period out-of-sample forecasts.

4. Noisy system estimation

The type of noise added to the Hénon map,

$$x_{t+1} = 1 - 1.4x_t^2 + y_t, \quad y_{t+1} = 0.3x_t, \quad (8)$$

Table 1
Conditional mean

Sample size	SIC	In-sample fit	Prediction interval	Predicted sign	Prediction
200	-14.264	0.407×10^{-6}			
190	-13.589	0.816×10^{-6}	10	1.00	0.893×10^{-6}
180	-13.477	0.859×10^{-6}	20	1.00	0.111×10^{-5}
150	-13.969	0.485×10^{-6}	50	1.00	0.472×10^{-6}
17	-12.243	0.409×10^{-6}	183	1.00	0.189×10^{-5}

Table 2
First and second partial derivatives

	Sample size	In-sample fit	Prediction interval	Predicted sign	Prediction
First partial derivative	200	0.848×10^{-4}			
	190	0.178×10^{-3}	10	1.00	0.806×10^{-4}
	180	0.187×10^{-3}	20	1.00	0.892×10^{-4}
	150	0.112×10^{-3}	50	1.00	0.563×10^{-4}
	17	0.209×10^{-3}	183	1.00	0.269×10^{-3}
Second partial derivative	200	0.001×10^{-4}			
	190	0.258×10^{-4}	10	1.00	0.355×10^{-4}
	180	0.203×10^{-4}	20	1.00	0.267×10^{-4}
	150	0.126×10^{-4}	50	1.00	0.118×10^{-4}
	17	0.337×10^{-4}	183	1.00	0.345×10^{-4}

is measurement noise. The series with measurement noise is calculated by $z_t = x_t + \sigma \epsilon_t$ where ϵ_t is uniformly distributed $U(-1, 1)$ and σ is the percentage times the signal to noise ratio, $\sigma = \kappa \sigma_x / \sigma_\epsilon$, where σ_x is the standard deviation of the data and σ_ϵ is the standard deviation of the noise component. The noise filtering capability of feedforward networks is analysed in five different levels of noise by setting $\kappa = 5, 10, 15, 25$ and 35% . The results are presented in Tables 3 and 4.

For all levels of noise, the fit and the first partial derivative of the Hénon map are approximated quite accurately with a data set of one hundred observations. The one period out-of-sample forecasts of the

fit and the first partial derivative are accordingly quite accurate and 100% sign predictions are obtained in all cases. At 25% and 35% levels of noise, the estimates of the second derivative and accordingly its one-period ahead predictions start to deteriorate. For noise levels larger than 35% the deterioration gets worse and quality of the fit gets poorer. This deterioration shows itself in derivative estimates.

With the noise filtering experiments shown above, one can wonder whether the dynamical invariants of a map can be calculated accurately from the filtered data. For all noise levels, the Lyapunov exponents of the in-sample and out-of-sample data are calculated by the algorithm of Gencay and Dechert [17]. The

Table 3
Conditional mean

Noise	SIC	In-sample fit	Prediction interval	Predicted sign	Prediction
5%	-6.193	0.159×10^{-3}	100	1.00	0.141×10^{-3}
10%	-4.811	0.628×10^{-3}	100	1.00	0.570×10^{-3}
15%	-4.005	0.143×10^{-2}	100	1.00	0.132×10^{-2}
25%	-2.995	0.418×10^{-2}	100	0.99	0.401×10^{-2}
35%	-2.337	0.874×10^{-2}	100	0.99	0.900×10^{-2}

Table 4
First and second partial derivatives

	Noise	In-sample fit	Prediction interval	Predicted sign	Prediction
First partial derivative	5%	0.263×10^{-3}	100	1.00	0.338×10^{-3}
	10%	0.138×10^{-2}	100	1.00	0.192×10^{-2}
	15%	0.348×10^{-2}	100	1.00	0.506×10^{-2}
	25%	0.169×10^{-1}	100	1.00	0.186×10^{-1}
	35%	0.349×10^{-1}	100	1.00	0.558×10^{-1}
Second partial derivative	5%	0.756×10^{-4}	100	1.00	0.700×10^{-4}
	10%	0.292×10^{-3}	100	1.00	0.287×10^{-3}
	15%	0.806×10^{-3}	100	1.00	0.822×10^{-3}
	25%	0.323×10^{-2}	100	1.00	0.354×10^{-2}
	35%	0.834×10^{-2}	100	1.00	0.928×10^{-2}

percentage error of the estimated Lyapunov exponents is calculated by

$$pe = \left| \frac{\sum_i^2 |\hat{\lambda}_i| - \sum_i^2 |\lambda_i|}{\sum_i^2 |\lambda_i|} \right|. \quad (9)$$

The results presented in Table 5, which show that even with 25% and 35% measurement noise, the Lyapunov exponents of the Hénon map are captured accurately with a percentage error not exceeding 5%.

5. Bifurcation prediction

We follow the procedure adopted in Ref. [2] to sweep through the parameters of the Hénon map by

$$x_{t+1} = 1 - \epsilon_t x_t^2 + y_t, \quad y_{t+1} = 0.3x_t, \quad (10)$$

and $\epsilon_{t+1} = \epsilon_t + \mu$, where μ is set to 0.0005 and $\epsilon_1 = 0.1$. 2500 observations are generated and used in the estimation. A feedforward network with 16 hidden units

is used and it gives a SIC value of -6.234 . The IMSE obtained from this network is 0.926×10^{-3} . The actual bifurcation diagram is presented in Fig. 1 and the estimated one is presented in Fig. 2. It is clear that a substantial portion of the true structure is recovered. Clearly, the period-doubling phases are recovered. One implication of this exercise is that if the model under study exhibits some type of parameter instability over a certain time period, the feedforward networks are able to capture the change in the qualitative dynamics.

6. Concluding remarks

The noise filtering capabilities of feedforward networks are investigated with measurement noise. Large amounts of measurement noise are added to a chaotic time series. The in-sample fit, in-sample partial derivatives and one-period out-of-sample forecasts for

Table 5
Lyapunov exponent estimates

Noise	In-sample		Out-of-sample		In-sample % error	Out-of-sample % error
	λ_1	λ_2	λ_1	λ_2		
0%	0.408	-1.620				
5%	0.395	-1.598	0.396	1.601	0.017	0.015
10%	0.394	-1.601	0.397	1.604	0.016	0.013
15%	0.398	-1.681	0.397	1.684	0.025	0.026
25%	0.397	-1.705	0.399	1.704	0.036	0.037
35%	0.396	-1.712	0.395	1.718	0.039	0.042

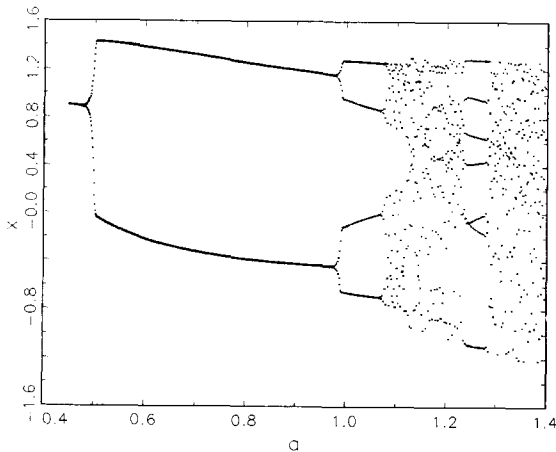


Fig. 1. The actual bifurcation diagram of the Hénon map.

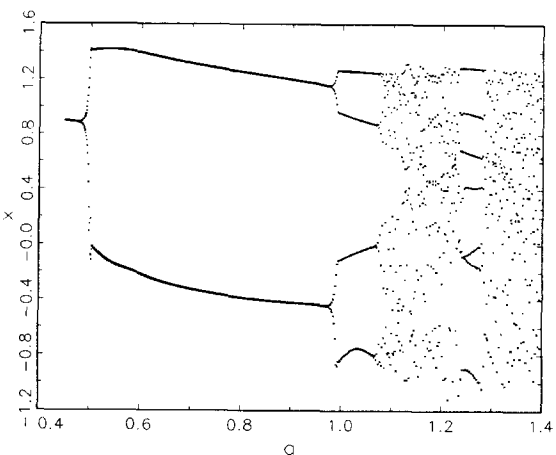


Fig. 2. The estimated bifurcation diagram of the Hénon map.

the fitted map and its partial derivatives are calculated as many as one hundred periods ahead. The results indicate that feedforward networks can filter a realistic amount of noise quite satisfactorily. The Lyapunov exponents of the filtered series are in turn calculated. The calculated Lyapunov exponents of the filtered series are in turn calculated. The calculated Lyapunov exponents are quite accurate as the percentage error of the true Lyapunov exponents from their estimates is less than 5%. The results from the bifurcation analysis indicate that feedforward networks can successfully approximate the qualitative changes in the dynamics of the time series data due to changes in the parameter values of the exogeneous variables.

Acknowledgement

I would like to thank the Natural Sciences and Engineering Council of Canada and the Social Sciences and Humanities Council of Canada for financial support.

References

- [1] F. Takens, in: *Dynamical systems and turbulence*, Warwick, 1980, eds. D. Rand and I. Young (Springer, Berlin, 1981) pp. 366–381.
- [2] M. Casdagli, *Physica D* 35 (1989) 335.
- [3] S.M. Carroll and B.W. Dickinson, in: *Proceedings of the International Joint Conference on Neural networks*, Washington, DC (IEEE Press, New York, 1989) pp. I: 607–611.
- [4] G. Cybenko, *Math. Control Signals Syst.* 2 (1989) 303.
- [5] K. Funahashi, *Neural Netw.* 2 (1989) 183.
- [6] K. Hornik, M. Stichcombe and H. White, *Neural Netw.* 3 (1990) 551.
- [7] C.-M. Kuan and H. White, *Artificial neural networks: an econometric perspective*, *Econometric Rev.* (1991), forthcoming.
- [8] A. Lapedes and R. Farber, *Nonlinear signal processing using neural networks: prediction and signal processing*, Los Alamos National Laboratory Technical Report (1987).
- [9] J.D. Farmer and J.J. Sidorowich, *Phys. Rev. Lett.* 59 (1987) 845.
- [10] J.D. Farmer and J.J. Sidorowich, *Exploiting chaos to predict the future and reduce noise*, in: *Evolution, learning and cognition*, ed. Y.C. Lee (World Scientific, Singapore, 1988).
- [11] A.R. Gallant, *Econometrics* 15 (1981) 211.
- [12] A.R. Gallant and H. White, in: *Proc. 2nd Annual IEEE Conference on Neural networks*, San Diego (IEEE Press, New York, 1988) pp. I: 657–664.
- [13] A. Barron, *Universal approximation bounds for superpositions of a sigmoidal function*, University of Illinois at Urbana-Champaign, Department of Statistics Technical Report 58 (1991).
- [14] E.K. Blum and L.K. Li, *Neural Netw.* 4 (1991) 511.
- [15] A.R. Gallant and H. White, *Neural Netw.* 5 (1992) 129.
- [16] A. Barron, *Approximation and estimation bounds for artificial neural networks*, University of Illinois at Urbana-Champaign, Department of Statistics Technical Report 59 (1991).
- [17] R. Gencay and W.D. Dechert, *Physica D* 59 (1992) 142.
- [18] W.H. Press, B.P. Flannery, S.A. Teukolsky and W.T. Vetterling, *Numerical recipes, the art of scientific computing* (Cambridge Univ. Press, Cambridge, 1986).
- [19] A. Barron, *Complexity regularization with application to artificial neural networks*, University of Illinois at Urbana-Champaign, Department of Statics Technical Report 57 (1990).