Chapter 9

Nonlinear Methods

9.1 Introduction

This chapter covers entropy, mutual information, correlation sums, source entropy and nonlinear prediction.

To motivate the use of nonlinear methods we give a simple example of where other methods fail. Our example is the logistic map

$$x_{t+1} = Rx_t(1 - x_t) \tag{9.1}$$

which is nonlinear because of the x_t^2 term. Different values of R are known to produce different dynamics; R=3.5 and 3.6 produce periodic dynamics and R=4 produces *chaotic dynamics*. A 'chaotic' system is a low-dimensional nonlinear determnistic system which is sensitive to initial conditions. Because of the 'folding' in the logistic

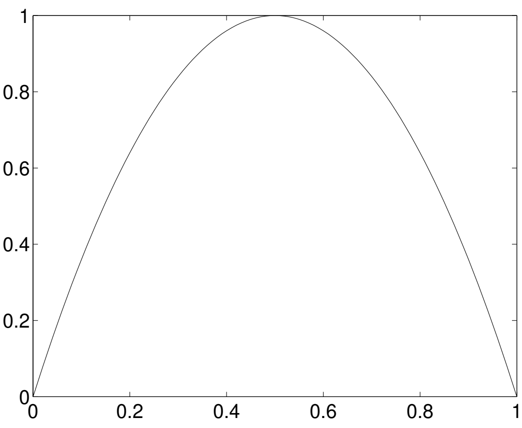


Figure 9.1: *A plot of x_{t+1} versus x_t for logistic map function $x_{t+1} = 4x_t(1 - x_t)$. If $x_{t+1} = 0.7$, then what was x_t ? Was it 0.23 or 0.77 ?*

map, for example, the system quickly forgets where its been before. Also, a slight change in the initial conditions soon leads to a big change in the subsequent state of the system.

For $R = 4$ the Power Spectral Density (PSD) is flat which is reminiscent of white noise (the corresponding autocovariance is only signficiantly non-zero at zero lag). Application of autoregressive models yields prediction errors with the same variance

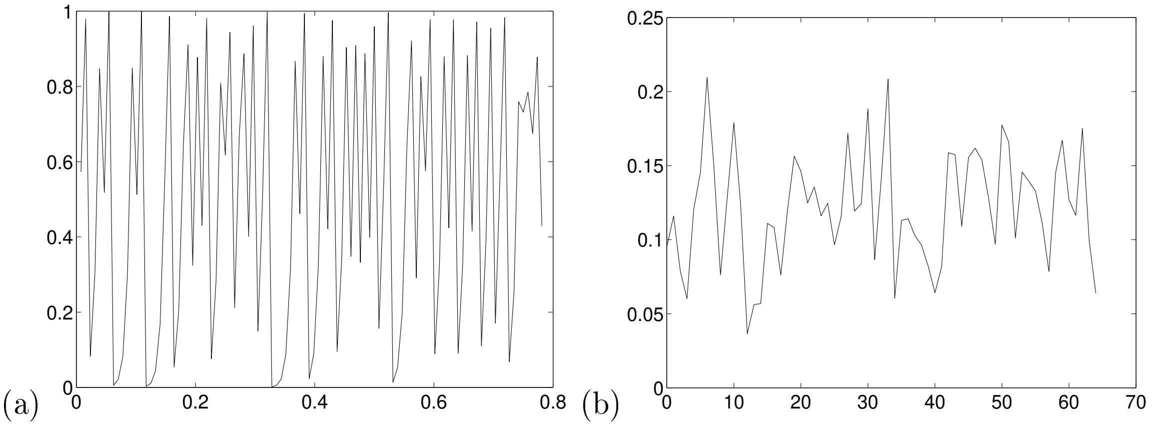


Figure 9.2: *(a) Time series from the logistic map ($R = 4$) and (b) its Power Spectral Density*

as the signal itself; ie. they are unable to detect any deterministic component in the signal. Thus, the application of linear methods would lead us to mistakenly conclude that the signal is purely stochastic when in fact it is purely deterministic.

If we apply nonlinear methods, however, then the underlying determinism can be discovered. This holds the promise of short-term predictability when, under the hypothesis of linear dynamics the system was considered to be unpredictable.

Also most early claims that physiological systems were chaotic have since been discredited. What is a more plausible working hypothesis, however, is that whilst these systems may not be nonlinear and *deterministic* they may very well be nonlinear and *stochastic*, and there is much evidence for this [23].

We look at methods for detecting nonlinear dependencies such as the *mutual information* and *marginal mutual information* and methods for exploiting these dependencies for purposes of prediction, such as *local-linear methods* and *neural networks*.

9.2 Lyapunov Exponents

A defining characteristic of a chaotic system is sensitivity to initial conditions. Points which are near at time 0 become exponentially far apart at time t. This can be captured in the relation

$$d_t = d_0 e^{\lambda t} \tag{9.2}$$

where d_0 is the initial distance, d_t is the distance at time t and λ is the *Lyapunov exponent*. Re-arranging the above equation gives

$$\lambda = \lim_{t \to \infty} \log \frac{d_t}{d_0} \tag{9.3}$$

λ_1	λ_2	λ_3	Attractor
-	-	-	Fixed Point
0	-	-	Cycle
0	0	-	Torus
+	0	-	Chaotic

Table 9.1: *Relation of sign of Lyaponov exponents to type of attractor.*

Negative λ's indicate convergence (damping) and positive λ's indicate divergence. Exponents equal to zero indicate cycles.

If the points are in a d-dimensional embedding space then neighboring points will initially be contained in a small multidimensional sphere. As time progresses this sphere will be stretched to form an ellipsoid with the length of the ith principal axis at time t given by $d_i(t)$. There is a corresponding *spectrum* of Lyapunov exponents; one for each axis. If we consider a 3-dimensional system, for example, then the relation between the signs of the Lyapunov exponents and the type of attractors is shown in Table 9.2. See [41] for more details.

The exponents can be calculate from a data set using the relation

$$\lambda_i = \lim_{t \to \infty} \log \frac{d_i(t)}{d_0} \tag{9.4}$$

Lyapunov exponents can be calculated from box-counting algorithms or from predictive models. In the last approach, for example, we can fit a neural network to the data, calculate the networks *Jacobian* matrix $\boldsymbol{J}$ (the derivative of the network's output with respect to its inputs - see Bishop [3] for details) and find λ_i from an eigendecomposition of $\boldsymbol{J}$ ([30] page 174). See also [13].

9.3 Measures of Information

See earlier lecture on Information Theory.

9.3.1 Continuous variables

In order to apply information theory to continuous variables we can partition continuous space into a number of discrete bins [1]. If we use M bins and observe n_i occurences in the ith bin then the probability of the value x_i occuring is

$$p(x_i) = \frac{n_i}{N} \tag{9.5}$$

[1] An alternative is to use a parametric model to estimate the probability density $p(\boldsymbol{x})$ from which $H(\boldsymbol{x})$ can be calculated. The entropy of such a continuous variable is known as the differential entropy [12].

where N is the total number of samples.

As we increase the number of bins, so the entropy increases.

If we have two continuous variables x and y and partition the two-dimensional space into bins where the number of levels in each dimension is M then the probability of a vector is given by

$$p(x_i, y_i) = \frac{n_{ij}}{N} \tag{9.6}$$

where there are n_{ij} samples in the i, jth bin and a total of N samples. The total number of bins will be M^2. The entropy of the above distribution is the joint entropy (see equation 4.5) and the mutual information can be calculated from 4.15. In general, these discretization procedures can be applied to d variables. But because the number of bins is M^d we need a lot of data to estimate the probabilities. As an alternative to box-counting algorithms we could use tree search algorithms or correlation sum methods (see later). See Pineda and Sommerer [48] for a review.

9.3.2 Measures of Information for Time Series

If our d continuous variables have come from a d-dimensional embedding of a time series eg.

$$\boldsymbol{x}_i = [x_i, x_{i-1}, ..., x_{i-d+1}] \tag{9.7}$$

and we partition the d-dimensional space into bins where the number of levels in each dimension is M then the probability of a vector is given by

$$p_d(\boldsymbol{x}_i) = \frac{n_i}{N - d + 1} \tag{9.8}$$

where there are n_i samples in the ith bin and a total of $N - d + 1$ samples. The total number of bins will be M^d so we need long time series to get good probability estimates.

Given a signal that has a range V the bin width will be $r = V/M$. The entropy of the above distribution is the joint entropy

$$H_d(\tau, r) = -\sum_{i=1}^{M^d} p_d(\boldsymbol{x}_i) \log p_d(\boldsymbol{x}_i) \tag{9.9}$$

where τ is the lag between samples. The *mutual information*, defined for $d = 2$, is

$$I(\tau, r) = 2H_1(\tau, r) - H_2(\tau, r) \tag{9.10}$$

It tells us about the nonlinear (or linear) correlation between $x_{t-\tau}$ and x_t and by varing τ we can plot an autocorrelation function. Figure 9.3 shows a plot of this for the logistic map time series. The entropies were calculated using a correlation sum method (see later) rather than a box-counting method. The mutual information reduces from about 4 at a lag of zero to nearly zero after 5 time steps. This makes sense as with the logistic map we lose about 1 bit of information per iteration. The

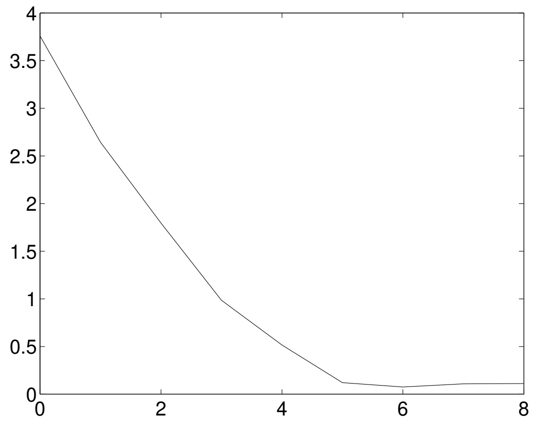


Figure 9.3: *Mutual Information, $I(\tau, r)$ versus lag τ for Logistic Map data. A resolution $r = 0.1\sigma_x$ was used where σ_x is the standard deviation of the data*

folding of the attractor acts like a switch and we lose about 1 bit of information per switch press.

For general d we can define the *joint mutual information* as the difference between the scalar entropies and the joint entropy

$$I_d(\tau, r) = dH_1(\tau, r) - H_d(\tau, r) \tag{9.11}$$

The joint mutual information measures the amount of information about x_t contained *independently* in the previous d samples ie. if we were to build a predictor, each of the previous d samples could be used but no interaction terms would be allowed.

9.3.3 Marginal Mutual Information

The joint mutual information measures the difference between the measured joint entropy of d variables and their joint entropy as if they were independent. For the special case $d = 2$ it therefore measures the amount of information about x_t contained in the previous sample $x_{t-\tau}$. For $d = 3$ and above, however, the corresponding measure is the *marginal mutual information* (or incremental mutual information or redundancy)

$$R_d(\tau, r) = I_d(\tau, r) - I_{d-1}(\tau, r) \tag{9.12}$$

We can re-write this in terms of joint entropies

$$R_d(\tau, r) = H_1(\tau, r) + H_{d-1}(\tau, r) - H_d(\tau, r) \tag{9.13}$$

Here the effect of the $d - 1$ previous variables is considered jointly (in the second term) whereas in the joint mutual information they were considered independently. The marginal mutual information, $R_d(\tau, r)$ measures the amount of information about x_t contained in the previous d samples. For $d = 2$ the marginal mutual information reduces to the mutual information.

9.3.4 Source Entropy

The *Approximate Source Entropy* statistics [47] are defined as

$$ApEn(d,r,N) = H_d(\tau,r) - H_{d-1}(\tau,r) \tag{9.14}$$

and

$$ApEn(d,r) = \lim_{N\to\infty} [H_d(\tau,r) - H_{d-1}(\tau,r)] \tag{9.15}$$

They are approximations to the *source entropy* or *KS-entropy* (from Mr. Kolmogorov and Mr Sinai) which is defined as

$$h_{KS}(\tau) = \lim_{r\to 0} \lim_{d\to\infty} ApEn(d,r) \tag{9.16}$$

Now, because of the limits, the $KS - Entropy$ can never be estimated experimentally (and, besides, it is only really of interest for purely deterministic sytems). But $ApEn$ can, and as long as the embedding dimension is large enough and the resolution fine enough it will provide a good approximation. That is,

$$h_{KS}(\tau) \approx ApEn(d,r) \tag{9.17}$$

Moreover, we can relate it to the marginal mutual information. If we substitute the above relation into equation 9.13 we get

$$R_d(\tau,r) = H_1(\tau,r) - h_{KS}(\tau) \tag{9.18}$$

Given that (see Weigend [63] page 50, or equation 9.42 later on)

$$h_{KS}(\tau) = \tau h_{KS} \tag{9.19}$$

then we have

$$R_d(\tau,r) = H_1(\tau,r) - \tau h_{KS} \tag{9.20}$$

Thus h_{KS} is the gradient of a plot of $R_d(\tau,r)$ versus τ. The d previous samples contain an amount of information $R_d(\tau,r)$ about the present sample which decreases as the time lag τ is increased. The rate of decrease is governed by the source entropy.

So, at a time lag of zero, the second term on the right is zero. The marginal mutual information is equal to the scalar entropy of the signal and the signal is completely predictable.

At each additional time step our predictive accuracy (which is governed by the marginal mutual information) loses h_{KS} bits. After a certain number of time steps, p_t, the marginal mutual information will fall to zero and all prediction accuracy will be lost.

In practice, zero prediction accuracy occurs when the the variance of the prediction error equals the variance of the signal σ_x^2. Given a prediction accuracy at zero lag of e_0 (equal to the resolution of the signal) after p_t time steps the accuracy will be

$$\sigma_x = e_0 2^{p_t h_{KS}} \tag{9.21}$$

Taking logs (to the base 2) gives

$$p_t = \frac{\log(\sigma_x / e_0)}{h_{KS}} \tag{9.22}$$

Therefore we must know the initial conditions exponentially more accurately (exponential decrease in e_0) to get a linear increase of the prediction horizon p_t. By measuring h_{KS} we can estimate the prediction horizon. Conversely, by measuring the prediction horizon, from a predictive model (see later), we can estimate h_{KS}.

9.3.5 Correlation Sums

As an alternative to box-counting algorithms we can use *correlation sums* to estimate the joint entropy (and therefore the mutual information and the source entropy). If we embed a time series in d-dimensional lag space such that

$$\boldsymbol{x}_i = [x_i, x_{i-1}, ..., x_{i-d+1}] \tag{9.23}$$

then we can measure the maximum distance between two points as

$$|\boldsymbol{x}_i - \boldsymbol{x}_j| = \max_k \{x_{i-k+1} - x_{j-k+1}\} \tag{9.24}$$

ie. look along the k out of d dimensions and pick the biggest distance. If we define the step function (or *Heaviside* function) as $h(x) = 1$ for $x \geq 0$ and $h(x) = 0$ for $x < 0$ then the indicator function

$$I_r(\boldsymbol{x}_i, \boldsymbol{x}_j) = h(r - |\boldsymbol{x}_i - \boldsymbol{x}_j|) \tag{9.25}$$

is 1 if the maximum distance between two points is less than r, and zero otherwise. We can now define the *pointwise correlation sum* as

$$C_i^d(r) = \frac{1}{N-d+1} \sum_{j=1}^{N-d+1} I_r(\boldsymbol{x}_i, \boldsymbol{x}_j) \tag{9.26}$$

which is the proportion of points within distance r of the point $\boldsymbol{x}_i$. As such this provides a good estimate for the probability density at point i

$$p_d(\boldsymbol{x}_i) = C_i^d(r) \tag{9.27}$$

The joint entropy can be approximated as the average log of this inverse probability [16]

$$H_d(r) = \frac{-1}{N-d+1} \sum_{i=1}^{N-d+1} \log p_d(\boldsymbol{x}_i) \tag{9.28}$$

Note that the sum is now over i whereas before it was over j. This method was used to calculate the mutual information in the earlier example. Now the probability $p_d(\boldsymbol{x}_i)$ can be decomposed as

$$\begin{aligned} p_d(\boldsymbol{x}_i) &= p(x_i^1, x_i^2, .., x_i^d) \\ &= p(x_i^d | x_i^1, x_i^2, .., x_i^{d-1}) p(x_i^1, x_i^2, .., x_i^{d-1}) \\ &= p(x_i^d | x_i^1, x_i^2, .., x_i^{d-1}) p_{d-1}(\boldsymbol{x}_i) \end{aligned} \tag{9.29}$$

Substituting this into the definitions for the joint entropies gives an expression for the approximate source entropy

$$ApEn(d, r, N) = \frac{-1}{N-d+1} \sum_{i=1}^{N-d+1} \log p(x_i^d | x_i^1, x_i^2, .., x_i^{d-1}) \tag{9.30}$$

Therefore, the approximate source entropy can be interpreted as the average log of a conditional probability; the probability that points are within distance r in embedding dimension d *given* that they were within this distance in embedding dimension $d-1$. Application of $ApEn$ to the logistic map shows that it is able to detect the difference between the 'simpler' periodic regime and the more complex 'chaotic' regime. Application of $ApEn$ to physiological signals is discussed in [23, 52, 47]. See Pincus

R	$ApEn$
3.5	0.0
3.6	0.229
3.8	0.425

Table 9.2: *Approximate entropy of the logistic map time series with* $d = 3$, $N = 300$, $r = 0.1\sigma_x$. *Increasing* R *increases the complexity of the time series which is reflected in higher values of* $ApEn$.

[47] for a discussion on how to select r.

9.4 Nonlinear Prediction

Given a time series x_n where $n = 1..N$ we wish to predict future values of the series ie x_{N+1}, x_{N+2} etc. If we view the time series up to time N as a fixed data set D then this can be achieved by inferring a statistical model from the data and using this model to predict future values of the signal.

This could, for example, be achieved by an autoregressive model which predicts the next value in the time series eg x_{N+1} as a linear combination of the p previous values

$$\hat{x}_{N+1} = w_1 x_N + w_2 x_{N-1} + ... + w_k x_{N-k+1} \tag{9.31}$$

where w_k are the autoregressive coefficients (see earlier lecture). These can be 'learnt' by tuning the model to the data set D.

This same process can be repeated but with a more powerful class of predictive models; nonlinear predictors. These replace the linear function in the above equation with a nonlinear function

$$\hat{x}_{N+1} = f(\boldsymbol{w}, x_N, x_{N-1}, .., x_{N-k+1}) \tag{9.32}$$

having parameters $\boldsymbol{w}$. Nonlinear predictors may be categorized into two broad classes (i) Local methods and (ii) Global methods.

9.4.1 Local methods

Given a data set of N embedded points $D = \{\boldsymbol{x}_n\}$ we can make a nonlinear prediction of a future time series value x_{p+T} from the embedded data point $\boldsymbol{x}_p$ as follows. Firstly, we find the k-nearest neighbours amongst D. That is, the k points in D which minimise the distance

$$||\boldsymbol{x}_n - \boldsymbol{x}_p|| \tag{9.33}$$

Put these points, $\tilde{\boldsymbol{x}}_n$, in rows of a matrix $\boldsymbol{X}$ and put the corresponding 'future' values $\tilde{x}_{n+T}$ into the vector Y. We now fit a linear model

$$\boldsymbol{Y} = \boldsymbol{w}\boldsymbol{X} \tag{9.34}$$

in the usual manner

$$\boldsymbol{w} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} \tag{9.35}$$

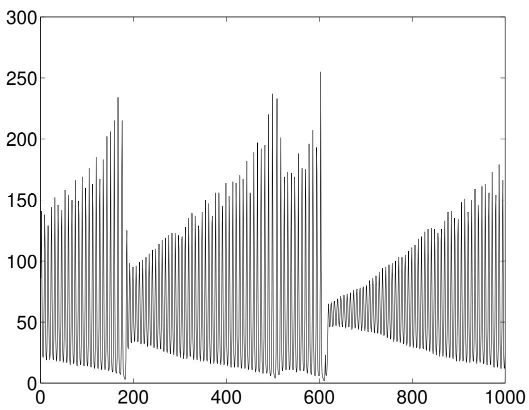
and we can then use it to make the prediction

$$\hat{x}_{p+T} = \boldsymbol{w}\boldsymbol{x}_p \tag{9.36}$$

This constitutes a *local autoregressive* model since only points in the neighbourhood of the predicting region have been used. As $k \rightarrow N$ we get the usual (global) autoregresive model.

A plot of prediction error versus k shows whether a local linear model (which is globally nonlinear) or a global linear model is appropriate. These plots are known as Deterministic versus Stochastic (DVS) plots [9]. For stochastic linear dynamics $k \rightarrow N$ gives the smallest error and for deterministic nonlinear dynamics $k \rightarrow 2d+1$, where d is the dimension of the attractor, gives the smallest error. Physiological data, such as heart rate or EEG, is in-between; it varies from nonlinear-stochastic to linear stochastic.

A cautionary note in the interpretation of these plots is due to the issue of *stationarity*. This is because a nonstationary linear system may be viewed as a stationary nonlinear system. The two viewpoints are both valid descriptions of the same dynamics.

(a) 

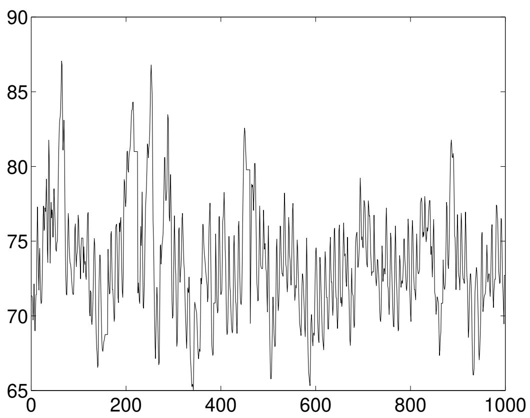
(b) 

Figure 9.4: *(a) Intensity pulsations of a laser and (b) heart rate.*

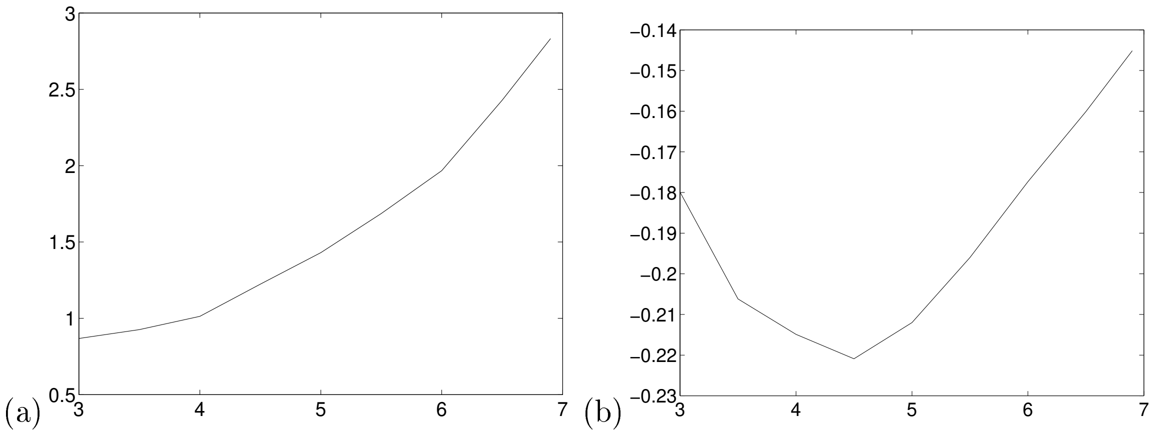


Figure 9.5: *Plots of (log) prediction error, E, versus (log) neighbourhood size, k, for (a) laser data and (b) heart-rate data. The minimum error points are at (a) $logk = 3$, $k = 21$ and (b) $logk = 4.5$, $k = 91$. These indicate that (a) the laser data is nonlinear and deterministic and (b) the heart-rate data is nonlinear and stochastic.*

Denoising

Not only can local methods be used for nonlinear prediction but also for nonlinear denoising. If, for example, the above linear prediction step is replaced by an SVD step we have a local-SVD denoising algorithm. This can also be used in combination with local prediction methods - see Sauer et. al in [63].

9.4.2 Global methods

Probably the most powerful nonlinear predictor is a *Neural Network* and the most commonly used network is the *Multi-Layer Perceptron* (MLP). This consists of a number of layers of processing elements (usually only two). The first layer consists of a number of linear transforms which are then operated on by a nonlinearity. There are $j = 1..p$ such functions each called a *hidden unit*

$$h_j = f(\sum_{i=1}^{d} w_{ij} x_{n-i}) \tag{9.37}$$

where i sums over the embedding and f is usually a sigmoidal nonlinearity

$$f(a) = \frac{1}{1 + e^{-a}} \tag{9.38}$$

The output of the second layer gives the networks prediction which is a linear combination of hidden unit responses

$$\hat{x}_{n+T} = \sum_{j=p}^{d} v_j h_j \tag{9.39}$$

Given a data set of of embedded vectors $\boldsymbol{x}_n$ and corresponding future values x_{n+T} (often $T = 1$) the parameters of the model can be set so as to minimise the prediction error

$$E = \sum_{n=1}^{N} (x_{n+T} - \hat{x}_{n+T})^2 \tag{9.40}$$

This can be achieved by various non-linear optimisation algorithms. The number of hidden units can be chosen according to various model order selection criterion. See Bishop [3] for details.

Application of neural nets to some time series, eg. the laser data, shows them to be better predictors than linear methods by several orders of magnitude [63].

Other global nonlinear methods involve the use of polynomial functions or *Volterra series*. Predictions are formed from linear combinations of quadratic and higher order terms eg.

$$\hat{x}_{n+T} = w_1 x_n + w_2 x_n^2 + w_3 x_n x_{n-1} + w_4 x_{n-1} + ... \tag{9.41}$$

The number and order of such functions can be found empirically or from prior knowledge of the possible interactions.

9.5 Discusion

A nonlinear dynamical system, with or without added stochastic noise, can thus be characterised by a number of measures: (i) source entropy, (ii) prediction error and (iii) Lyapunov exponents and there are relations between them. There are also many more measures that we have'nt discussed. Most of these are relevant to nonlinear *deterministic* systems rather than nonlinear *stochastic* ones. (the most prominent being *correlation dimension* [24]).

To use them to, say, differentiate between different physiological states or experimental conditions requires not just estimating the measures themselves but also providing error bars so we can apply significance tests.

For these 'nonlinear' statistics, these most often take the form of Monte-Carlo estimates. Given a particular time series we compute our measure of interest, say *ApEn*. We then shuffle the data and recompute the statistic. If we do this for a number of shuffles then where on the resulting PDF our original value falls is the significance value.

The sum of the positive Lyapunov exponents is equal to the source entropy

$$h_{KS} = \sum_{\lambda_i > 0} \lambda_i \tag{9.42}$$

This is known as *Pesin's Identity* [2]. This completes the circle: Source Entropy $\rightarrow$ Nonlinear Prediction $\rightarrow$ Lyapunov Exponents $\rightarrow$ Source Entropy etc.

[2] In fact, it is an upper bound on the source entropy [30]