March 1993

LU TP 93-4

Finding the Embedding Dimension and Variable Dependences in Time Series

Hong Pi[1] and Carsten Peterson[2]

Department of Theoretical Physics, University of Lund
Sölvegatan 14A, S-223 62 Lund, Sweden

Submitted to **Neural Computation**

Abstract:

We present a general method, the *δ-test*, which establishes functional dependencies given a table of measurements. The approach is based on calculating conditional probabilities from vector component distances. Imposing the requirement of continuity of the underlying function the obtained values of the conditional probabilities carry information on the embedding dimension and variable dependencies. The power of the method is illustrated on synthetic time-series with different time-lag dependencies and noise levels and on the sunspot data.

The virtue of the method for preprocessing data in the context of feed-forward neural networks is demonstrated. Also, its applicability for tracking residual errors in output units is stressed.

[1]pihong@thep.lu.se
[2]carsten@thep.lu.se

Introduction

The behaviour of a dynamical system is often modelled by analyzing a time series record of certain system variables. Using artificial neural networks (ANN) to model such systems has recently attracted much attention. The success of such models relies heavily upon identifying the underlying structure in the time series – it is advantageous to know in advance the the embedding dimension, most relevant inputs, noise level etc.. In this paper we devise a simple and easy-to-use method based on continuity requirements on statistical measures for identifying such essential properties in a time series record. Even though the language is that of time series the approach applies to any continuous function mapping problem.

Time series can have a wide range of behaviour ranging from being entirely random and uncorrelated to being completely deterministic. In reality one is often in between these two extremes. Existing approaches to determine dependencies are either based on entropy measures [1, 2], or on elaborate autocorrelation measures [3, 4, 5]. Our approach, which has its roots in the latter philosophy, aims at establishing noise levels, determining the embedding dimension and to pin down sensitivity on the various variables.

In ref. [4] a method inspired by the Grassberger-Procaccia correlation integral [3] is devised that can be used to verify the degree of (independent) randomness. This test is further developed in ref. [5] into a conditional probability approach in which the degree of variable dependence may be quantified. Although the latter method has its merit in brevity, the message it gives is not without ambiguities.

Strongly inspired by ref. [5] we propose a method from a different viewpoint, which exploits the definition of function continuity. This definition, which can be easily connected to the behaviour of the conditional probabilities, gives clear signatures (apart from ambiguities arising from insufficient statistics) with respect to variable dependences and the embedding dimensionality. To our knowledge the proposed method does not exist in the literature despite its conceptual simplicity.

Successful explorations are made on different maps with and without noise and with a variety of time-lag dependencies. Also, the underlying dynamics of a sunspot series is studied.

The relevance of the method for feed-forward network training is illustrated with the sunspot series, where it is shown that feeding the network with the established minimum embedding dimension vectors gives rise to state-of-the-art generalization performance. Also, the method can be used to track residual dependencies of the output errors in a Multilayer Perceptron (MLP).

General Formulation

Consider a discrete-time system, manifested as a time series x_t, $t = 1, 2, 3, ..., N$, where we are interested to know if there exists a continuous map relating future values to the past ones, *i.e.* if it is possible to identify a state equation[3]

$$x_t = f(x_{t-1}, x_{t-2}, ..., x_{t-d}) + r_t \tag{1}$$

The "noise"-term r_t represents an indeterminable part which originates either from insufficient dimension of the measurements or from real noise. In general r_t should decrease with d. If the system is completely deterministic, r_t should vanish entirely as d exceeds the minimum embedding dimension d_{min}.

For a map given as a table of measurements, one wants to know: (1) the minimum embedding dimension d_{min}, (2) the sensitivity of x_t with respect to each of the dependent variables, and (3) an estimate of the size of the noise. With "variable dependence" we mean the primary dependence, not the induced ones. If $x_t = f(x_{t-1}) = f(f(x_{t-2}))$ we say that x_{t-1} is the primary dependent variable, the induced dependence on x_{t-2} is of no interest. Variables with no primary dependence are denoted "irrelevant".

We approach the problem by constructing conditional probabilities in embedding spaces of various dimensions d. The time series in eq. (1) is represented as a series of N points $\mathbf{z}(i)$ in a $(d+1)$-dimensional space $(d = 0, 1, 2, ...)$

$$\mathbf{z}(i) = (z_0(i), z_1(i), .., z_k(i), ., z_d(i)) \tag{2}$$

where $z_k(t) = x_{t-k}$. The distances between the k:th components of two vectors $\mathbf{z}(i)$ and $\mathbf{z}(j)$ are defined as

$$l_k(i,j) = |z_k(i) - z_k(j)|, \quad k = 0, 1, ..., d \tag{3}$$

Given a set of positive numbers, ϵ and $\vec{\delta} = (\delta_1, ..., \delta_d)$, one can construct the following joint probabilities from the data

$$P(l_0 \leq \epsilon, \vec{l} \leq \vec{\delta}) = \frac{1}{N_{pair}} n(l_0 \leq \epsilon, \vec{l} \leq \vec{\delta}) \tag{4}$$

$$P(\vec{l} \leq \vec{\delta}) = \frac{1}{N_{pair}} n(\vec{l} \leq \vec{\delta}) \tag{5}$$

where N_{pair} is total number of vector pairs, and $n(l_0 \leq \epsilon, \vec{l} \leq \vec{\delta})$ and $n(\vec{l} \leq \vec{\delta})$ are the number of the pairs satisfying the corresponding distance constraints. Throughout this paper we freely use the notation $\vec{l} \leq \vec{\delta}$ for $\{(l_1 \leq \delta_1), (l_2 \leq \delta_2), ..., (l_d \leq \delta_d)\}$. Also we set $\delta_i = \delta$ for all i.

[3]The series is assumed to be bounded and stationary and x_t can take either real or complex values.

Next we form the conditional probabilities

$$P_d(\epsilon|\,\vec{\delta}) \equiv P(l_0 \leq \epsilon|\,\vec{l} \leq \vec{\delta}) = \frac{P(l_0 \leq \epsilon, \vec{l} \leq \vec{\delta})}{P(\vec{l} \leq \vec{\delta})}. \tag{6}$$

How is $P(l_0 \leq \epsilon|\,\vec{l} \leq \vec{\delta})$ expected to vary under different conditions? The following important observations can be made:

1. For a completely random time series one has

$$P_0(\epsilon) = P_1(\epsilon|\,\vec{\delta}) = ... = P_d(\epsilon|\,\vec{\delta}) = ... \tag{7}$$

   This identity, which should be understood in a statistical sense, holds for any choice of positive ϵ and δ.

2. If a continuous map exists as in eq. (1) with no intrinsic noise, then for any $\epsilon > 0$ there exists a δ_ϵ such that

$$P_d(\epsilon|\,\vec{\delta}) = 1 \quad \text{for} \quad \delta \leq \delta_\epsilon \quad \text{and} \quad d \geq d_0 \tag{8}$$

   The smallest integer d_0 for which eq. (8) holds is identified with $(d_{min} - 1)$.

3. In the presence of noise r, $P_d(\epsilon|\,\vec{\delta})$ will no longer saturate to 1 as ϵ drops below the width Δr_{max} of the noise.

Eq. (8) is a direct consequence of the definition of function continuity, which states that if $z_0 = f(z_1, ..., z_d)$, then for any $\epsilon > 0$ there exists a $\delta > 0$ such that the conditions $|z_1 - z_1'| < \delta$,...,$|z_d - z_d'| < \delta$ guarantee $|z_0 - z_0'| < \epsilon$. With the presence of noise, however, $|z_0 - z_0'| = |f(z_1, ...) - f(z_1', ...) + r - r'|$, which as $\delta \to 0$ can not be made smaller than $\Delta r = |r - r'|$. This justifies the statement # 3.

How does $P_d(\epsilon|\,\vec{\delta})$ vary as a function of δ for fixed ϵ? For $\delta \to \infty$ the conditions have no affect. Hence one has $P_d(\epsilon|\,\vec{\delta})|_{\delta\to\infty} = P_0(\epsilon)$. As $\delta \to 0$, $P_d(\epsilon|\,\delta)$ should increase monotonically and saturate to 1 for $d \geq d_0$. This behaviour is shown schematically in fig. 1a. If d is not a primary variable, it is tempting to assume that $P_d(\epsilon|\,\vec{\delta})$ will exhibit no change as compared to $P_{d-1}(\epsilon|\,\vec{\delta})$. However this is not always true for arbitrary δ. Effects from nonuniform curvatures [5] in the data can induce sizable changes in $P_d - P_{d-1}$ for a certain range of δ-values even if d is an irrelevant variable. For this reason we examine the maxima

$$P_d(\epsilon) = \max_{\delta > 0} P_d(\epsilon|\,\vec{\delta}) = P_d(\epsilon|\,\vec{\delta})|_{\delta \leq \delta_\epsilon}. \tag{9}$$

Saturation of the maxima as d increases singles out the irrelevant variables. How the maxima $P_d(\epsilon)$ change with d and ϵ provides basically all the information we need (see fig. 1b).

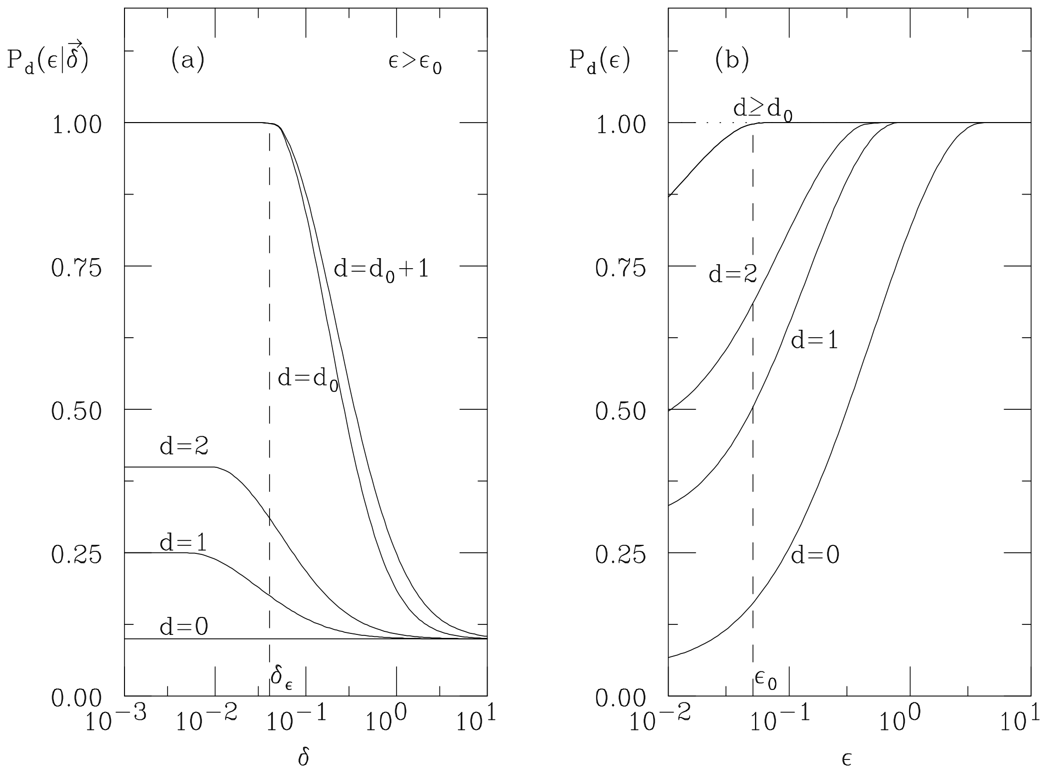

Figure 1: (a). $P_d(\epsilon|\,\vec{\delta})$ as a function of δ for fixed ϵ. (b). $P_d(\epsilon)$ as a function of ϵ. Saturation to 1 would be observed for $d \geq d_0$. In the presence of noise the saturation deviates from 1 around $\epsilon_0 \sim \Delta r_{max}$.

$P_d(\epsilon)$ measures how well the dynamics can be modelled in terms of the d variables. To quantify the dependence on each of the variables, it is convenient to define a *dependability index*[4]

$$\lambda_d(\epsilon) = \frac{P_d(\epsilon) - P_{d-1}(\epsilon)}{1 - P_0(\epsilon)}, \quad d = 1, 2, \ldots \tag{10}$$

and its average over ϵ

$$\overline{\lambda}_d = \frac{\int_0^\infty d\epsilon\, \lambda_d(\epsilon)(1 - P_0(\epsilon))}{\int_0^\infty d\epsilon\, (1 - P_0(\epsilon))} = \frac{\int_0^\infty d\epsilon\, (P_d(\epsilon) - P_{d-1}(\epsilon))}{\int_0^\infty d\epsilon\, (1 - P_0(\epsilon))}. \tag{11}$$

For a noise-free deterministic map, $P_d(\epsilon)$ saturates to 1 for $d \geq d_0$ and one has

$$\sum_{d=1}^{d_0} \overline{\lambda}_d = \sum_{d=1}^{d_0} \lambda_d(\epsilon) = 1 \tag{12}$$

[4]This is similar to the index defined in ref. [5] apart from a different definition of $P_d(\epsilon)$ and the normalization.

Irrelevant variables are identified by having $\overline{\lambda}_d \approx 0$.

So far the formalism has assumed infinite amount of data. With limited statistics very low δ-values or large d's may give rise to a picture not as crisp as the one in fig. 1. A rough estimate of the errors in such a case is

$$\Delta P_d(\epsilon|\,\vec{\delta}) = \sqrt{\frac{P_d(\epsilon|\,\vec{\delta})}{n(\vec{l} \leq \vec{\delta})}}. \tag{13}$$

Theoretically $P_d(\epsilon|\,\vec{\delta})$ is expected to be a smooth and monotonically decreasing function of δ. It is generally flat near $\delta = 0$ and has a plateau that extends to a finite δ_ϵ. Finding its approximate maximum is not difficult, provided that there is a reasonable amount of statistics to probe the plateau region. In cases with very limited statistics it is advantageous to set an irrelevant variable k *inactive*, which means that the condition $l_k \leq \delta$ is omitted when computing $P_d(\epsilon|\,\vec{\delta})$ for $d > k$. By removing the unnecessary restrictive conditions in this way improved statistics is obtained without affecting the evaluation of P_d's in higher dimensions.

Implementation Issues

Next we give a systematic procedure following the above on how to read off the key properties of $P_d(\epsilon|\,\vec{\delta})$ given a data set. We propose the following scheme:

0. Compute $P_0(\epsilon)$.
1. Starting with d=1 compute $P_d(\epsilon|\,\vec{\delta})$ and find the maxima $P_d(\epsilon)$.
2. Evaluate the dependability index $\overline{\lambda}_d$.
3. (Optional) Variable elimination: if the d-th variable is irrelevant ($\overline{\lambda}_d \sim O(0)$), set it inactive.
4. Increment d by 1 and repeat steps 1-3 until $P_d(\epsilon)$ saturates.
5. Identify d_0 for which $P_d(\epsilon)$ begins to saturate, and also the point ϵ_0 for which $P_{d_0}(\epsilon)$ begins to deviate from 1. The minimum embedding dimension is then $d_0 + 1$, and the noise width is estimated as ϵ_0.

If option # 3 is used the dependability index in eq. (10) is modified according to: $\lambda_d(\epsilon) = (P_d(\epsilon) - P_{d'}(\epsilon))/(1 - P_0(\epsilon))$, where $d' \leq d-1$ is the closest *active* variable and the summations in eq. (12) are restricted to active variables only.

In what follows ϵ and δ are expressed in units of σ, the standard deviation of the data set. In implementations of step # 1 the $\ln \epsilon$-$\ln \delta$ plane is discretized by 20×30 bins with $\ln \epsilon_{min} \leq \ln \epsilon \leq \ln 4$ and $\ln \delta_{min} \leq \ln \delta \leq \ln 4$. A lower cut $\epsilon_{min} = 0.1$ is used for the ϵ-integrations in eq. (11).

Applications

We apply the method to two synthetic time series and the sunspot data problems.

For the **Logistic map** we generate 4000 time-steps according to

$$x_t = \eta x_{t-1}(1 - x_{t-1}) + r_t \tag{14}$$

with $\eta = 4$ giving $\sigma = 0.35$. The iterative noise r_t consists of random numbers uniformly distributed in $(-r, r)$. Two data sets are generated with $r = 0$ (a) and $r = 0.28\sigma$ (b) respectively. In order to keep the series bounded r_t is constrained such that $x_t \in (0, 1)$.

In fig. 2a we show $P_d(\epsilon|\vec{\delta})$ versus δ at $\epsilon = 0.085$ for various d. We observe a saturation of $P_d(\epsilon|\vec{\delta})$ to 1 for $d \geq 1$ in fig. 2a, as expected for a one dimensional noise-free map. The conditional probabilities still saturate in (b) but only to ~ 0.55, which indicates the presence of noise. In fig. 2b the maximised conditional probability $P_d(\epsilon)$ is shown as a function of ϵ for various dimensions d. For the noise-free data $P_d(\epsilon)$ saturates to 1 for $d \geq 1$ in the entire ϵ range. For the noisy data $P_d(\epsilon)$ still saturates approximately for $d \geq 1$ but the saturation value starts to fall off from 1 somewhere between 0.3 and 0.7, implying a noise level r between 0.15σ and 0.35σ, which is consistent with the generated noise level. Calculating the dependability indices gives $\overline{\lambda}_1 = 1$ for the $r = 0$ data, $\overline{\lambda}_1 = 0.97$ for the noisy data and $\overline{\lambda}_{d \geq 2} \approx 0$.

For the **Hénon map** we generate 4000 time-steps according to

$$x_t = 1 - a(x_{t-2} - r_{t-2})^2 + b(x_{t-4} - r_{t-4}) + r_t \tag{15}$$

with $a = 1.4$, and $b = 0.3$ giving $\sigma = 0.723$. This is the usual Hénon map with the dependences stretched to larger lags and with the noise additively applied. Again we will use two data sets with $r = 0$ (a) and $r = 0.14\sigma$ (b). The variable elimination option is used here. We obtain $\overline{\lambda}_{1-4} = 0.002$, 0.886, -0.023, 0.114 for (a) and 0.052, 0.728, 0.004, 0.128 for (b). The dependences on x_{t-2} and x_{t-4} emerge as large values of $\overline{\lambda}_2$ and $\overline{\lambda}_4$. For (a) $\overline{\lambda}_2 + \overline{\lambda}_4 = 1$, indicating a noise-free map, whereas for (b) the value is 0.86 signalling the presence of noise.

The **sunspot data** contains the annual averaged sunspot activities from 1700 to 1979 (280 points). Since this is a very limited statistics data set we use the variable elimination

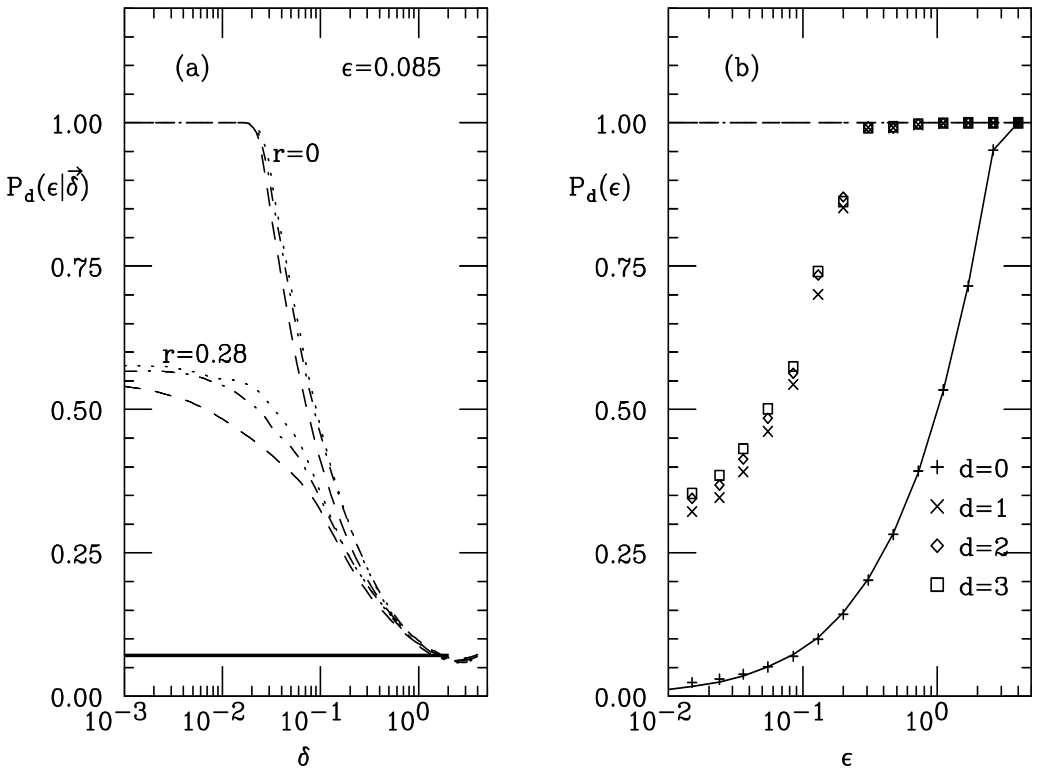

Figure 2: Logistic map: (a) $P_d(\epsilon|\vec{\delta})$ as a function of δ for noise-free ($r = 0$) and noisy data ($r = 0.28$). The curves represent $d = 0$ (solid), $d = 1$ (dashed),$d = 2$ (dash-dotted) and $d = 3$ (dotted) respectively. (b) $P_d(\epsilon)$ as a function of ϵ. The curves corresponds to noise-free data (same notation as in (a) and the points to the noisy data.

option. In fig. 3a we show the dependability indices, from which we conclude that x_{t-1}, x_{t-2}, x_{t-10} and x_{t-9} are the important variables for modelling the series[5]. Summing over the indices on the four variables gives $\sum \overline{\lambda}_d = 0.84$, signalling the presence of noise.

$P_d(\epsilon|\delta)$ and $P_d(\epsilon)$ contain rather large statistical errors. In fig. 4 we see that $P_d(\epsilon|\delta)$ approximately saturates to 1 for $\epsilon \sim 0.94$, which gives 0.47σ as an upper-limit estimation for the noise level. The resulting mean square error in any model fitted the series based on the four variables should be on the order of $(0.47\sigma)^2$.

[5]Due to insufficient statistics in the sunspot series, we can not completely exclude the possibility of minor dependences on the other variables.

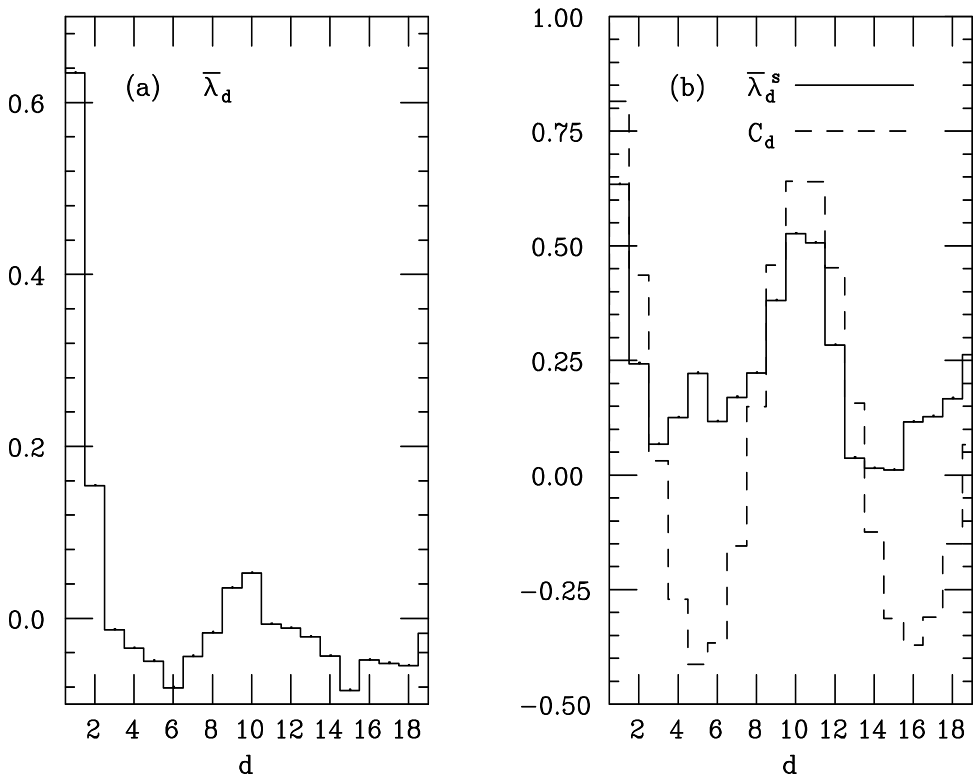

Figure 3: Sunspot data: (a) Dependability index $\overline{\lambda}_d$ versus time lags d. (b) The single variable dependability index $\overline{\lambda}^s_d$ (full line) and the autocorrelation C_d (broken line) plotted as functions of d.

An interesting variant of the δ-test is to perform a "single variable analysis", by checking how the probability distributions are affected by the conditional variables one at a time. The conditional probabilities are now computed as $P_d(\epsilon|\delta) = P_d(l_0 \leq \epsilon | l_d \leq \delta)$, which is equivalent to evaluating $P_d(\epsilon|\delta)$ with all but the d-th variables set inactive. The dependability index is now defined by [6] $\lambda^s_d(\epsilon) = (P_d(\epsilon) - P_0(\epsilon))/(1 - P_0(\epsilon))$. These computations have the minimum statistics requirement. $\lambda^s_d(\epsilon)$ measures the correlations between x_t and x_{t-d}, similar to the usual autocorrelation function C_d except that it is nonlinear. In fig. 3b we show the integrated single variable index $\overline{\lambda}^s_d$ together with C_d. While C_d gives correlations between the values of the variable x_t and x_{t-d}, $\overline{\lambda}^s_d$ establishes or excludes the possibility of a continuous functional relationship between the two variables.

[6]Eq. (12) does not apply to $\lambda^s_d(\epsilon)$.

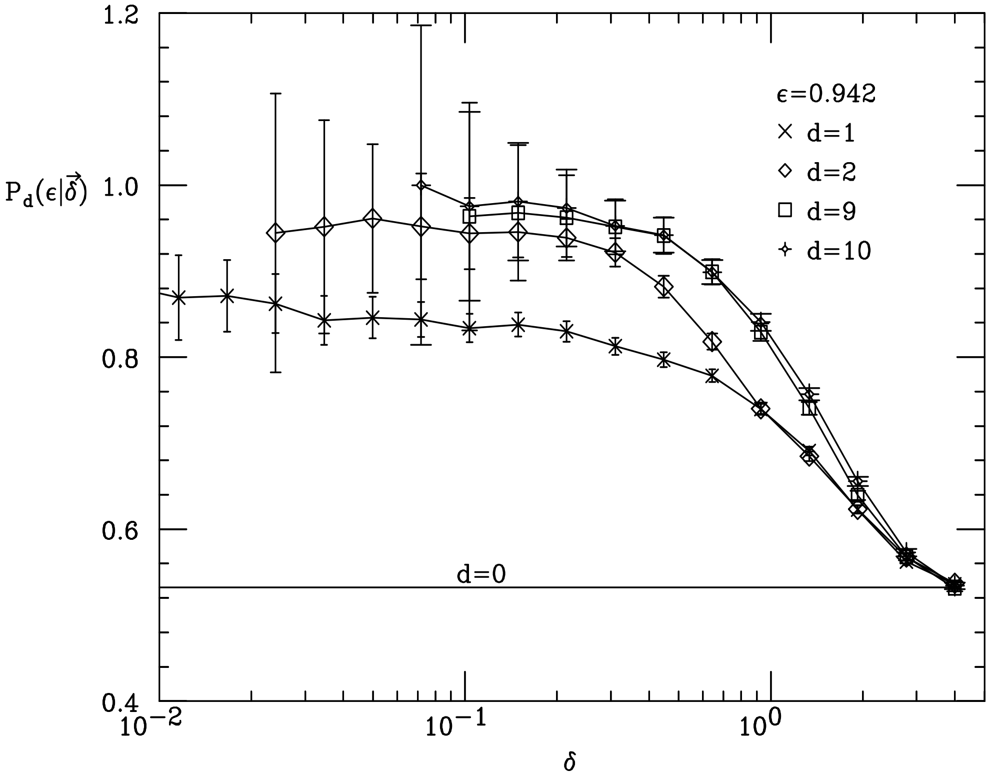

Figure 4: $P_d(\epsilon|\,\vec{\delta})$ as a function of δ for $\epsilon = 0.942$ and various dimensions d. The estimated errors are indicated.

Impact on Neural Network Learning

Feeding an MLP with redundant variables should be avoided since fitting to noise gives rise to poor generalization. In ref. [6] an MLP with one layer of 8 hidden nodes is trained with the sunspot series data. The authors experiment with different number of time-lags as inputs, 6, 12 and 24 and concludes on the basis of generalization performance that 12 is optimal and hence that this number reflects the embedding dimension.

Following the procedures (without weight elimination) of ref. [6] we have trained two MLP with 8 hidden nodes using 12 lag variables as inputs or just the 4 most relevant variables according to the δ-test. With 12 inputs the network learns the training set well, but has the tendency to overfit resulting deteriorated generalization performances. Limiting the input to the 4 most relevant variables gives rise to a similar generalization performance but the problem with overfitting is largely reduced. It is also interesting to notice that after training the network with weight elimination, the authors in ref. [6] find large connections to the hidden nodes from the inputs x_{t-1}, x_{t-2} and x_{t-9}, in agreement with the δ-test findings.

The δ-test is also very powerful when it comes to analyzing the residual error of output units in MLP learning. With perfect training the δ-test should identifies the residual series as an independent random series. If this is not the case the test singles out the relevant input units from which information has not been fully extracted by the network.

Summary

We have devised a general method, the δ-test, for identifying dependencies in continuous functions. It is not limited to linear correlations and it determines the embedding dimensions, dependencies and noise levels fairly accurately even in cases of low statistics. The approach is inspired by ref. [5], which can be regarded as the $\delta \to \epsilon$ limit of ours. The problems with the latter approach are two fold [7]: (1) induced dependences can show up in the index if nonuniform curvature is present in the function map; (2) there is not a saturation measure to indicate the critical embedding dimension. We are able to extract information unambiguously from the region where $P_d(\epsilon|\vec{\delta})$ is maximized. Our approach is similar to the Kolmogorov entropy method [1] in that both methods examine the behaviour of certain set of conditional probability distributions. However evaluating entropy for a high dimensional joint probability distribution requires huge amount of statistics in contrast to the present method [7]

Acknowledgements:

We are indebted to Richard Blankenbecler for bringing ref. [5] to our attention, which inspired this work.

References

[1] A.N. Kolmogorov *Dokl. Akad. Nauk. USSR* **98**, 527 (1959).

[2] J.D. Farmer, "Information Dimension and the Probabilistic Structure of Chaos", *Z. Naturforschung* **37A**, 1304 (1982).

[3] P. Grassberger and I. Procaccia, "Measuring the Strangeness of Strange Attractors", *Physica* **D9**, 189 (1983).

[4] W.A. Brock, W.D. Dechert, J.A. Scheinkman and B. LeBaron, "A Test for Independence Based on the Correlation Dimension", University of Wisconsin Preprint (1988).

[5] R. Savit and M. Green, "Time Series and Dependent Variables", *Physica* **D50**, 95 (1991).

[6] A. Weigend, B.A. Huberman and D. Rumelhart, "Predicting the Future: A Connectionist Approach", *International Journal of Neural systems* **3**, 193 (1990).

[7] H. Pi and C. Peterson, to be published elsewhere.