Proceedings of the Fourth International Workshop on Software Engineering and Artificial Intelligence for High Energy and Nuclear Physics, eds. B. Denby and D. Perret-Gallix, International Journal of Modern Physics **C6**, 611 (1995). [*LU TP 95-4*]

Determining Dependency Structures and Estimating Nonlinear Regression Errors Without Doing Regression

Carsten Peterson[1]

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

Invited talk presented at
Fourth International Workshop on Software Engineering and Artificial Intelligence for High Energy and Nuclear Physics,
Pisa, Italy, April 3-8, 1995

Abstract:

A general method is discussed, the δ-test, which establishes functional dependencies given a table of measurements. The approach is based on calculating conditional probabilities from data densities. Imposing the requirement of continuity of the underlying function the obtained values of the conditional probabilities carry information on the variable dependencies. The power of the method is illustrated on synthetic time-series with different time-lag dependencies and noise levels. For N data points the computational demand is N^2. Also, the same method is used for estimating nonlinear regression errors and their distributions without performing regression. Comparing the predicted residual errors with those from linear models provides a signal for nonlinearity. The virtue of the method in the context of feedforward neural networks is stressed with respect to preprocessing data and tracking residual errors.

[1] carsten@thep.lu.se

Motivation

Successful regression of a system given tables of data with no access to first principles models relies heavily upon identifying the underlying structure – embedding dimension, most relevant inputs and noise levels. Finding relevant inputs is a natural step prior to any artificial neural networks (ANN) processing. Furthermore, if the noise variance can be estimated then one knows the optimal performance limit of the fit in advance. Also, methods for filtering data often require prior estimate of noise variance. To be more explicit, consider a table of data, $\{(y^{(i)}, \mathbf{x}^{(i)}),\ i = 1, 2, ..., N\}$, where y is the dependent variable and the d-dimensional vector $\mathbf{x}$ denotes the set of explanatory variables. One aims at determining sensitivities upon the different $\mathbf{x}$-components and the variance of r (σ_r^2) for

$$\hat{y} = F(\mathbf{x}) + r \tag{1}$$

where F represents the optimum model. Conventional procedures for accomplishing this are model-based. One fits the data to a model, a particular choice of F, and then interprets the results. In the special case of linear regression models [1] where F takes the form $\hat{y} = a_0 + \sum a_k x_k$, the dependencies are simply given by the covariances $\langle y, x_k \rangle$ and the sample variance σ_r^2 is explicitly given by

$$\sigma_r^2 = \sigma^2 - \sum_{k=1}^{d} a_k \langle y, x_k \rangle \tag{2}$$

where σ denotes the y-variable variance. With ANN regression methods one first makes a fit to the data. The dependency on the different input variables is then extracted either implicitly by including complexity terms in the error term [2] or explicitly by inspecting e.g. suitable derivatives with frozen weights.

Here we report on a method, the δ-test [3, 4], where the dependency structures and noise variance are extracted with no modeling involved and no assumption made about the noise distribution. The estimates from the δ-test only relies upon the assumption that F is uniformly continuous and that the noise in eq. (1) is additatively added. Prior approaches to determine dependencies are either based on entropy measures [5, 6], or on elaborate autocorrelation measures [7, 8, 9]. Our approach, has its roots in the latter philosophy. The power of the method is illustrated with different examples of chaotic time series augmented with noise: the logistic map, the Henon map and the Ikeda [12]map. Other successful applications [10, 11] also exist.

Method

We approach the problem by constructing conditional probabilities $P_d(\epsilon|\delta)$ for different embedding dimensions d and different choices of positive real numbers ϵ and δ from pairs of data points as follows

$$P_d(\epsilon|\,\delta) \equiv P(|\Delta y| \leq \epsilon \mid |\Delta \mathbf{x}| \leq \delta) \tag{3}$$

where $|\Delta \mathbf{x}| \equiv \max_k |x_k - x'_k|$. The calculational demand for this is $1/2N(N-1)$ for N data points. What does $P_d(\epsilon|\,\delta)$ tell us? The following important observations can be made:

1. For completely random data one has for any choice of ϵ and δ

$$P_0(\epsilon) = P_1(\epsilon|\,\delta) = \ldots = P_d(\epsilon|\,\delta) = \ldots \tag{4}$$

2. In the limit $\delta \to 0$, one obtains

$$\begin{aligned} P_d(\epsilon) &\equiv \lim_{\delta\to 0} P_d(\epsilon|\,\delta) \\ &= P(|F(\mathbf{x}) - F(\mathbf{x}') + r - r'| \leq \epsilon \mid |\mathbf{x} - \mathbf{x}'| \to 0) \\ &= \mathrm{Prob}(|\Delta r| \leq \epsilon), \end{aligned} \tag{5}$$

   where the property of function continuity, $F(\mathbf{x}) - F(\mathbf{x}') \to 0$ for $\mathbf{x} \to \mathbf{x}'$, is exploited. Hence in the noiseless case (Δr=0) one has $P_d(\epsilon)$=1.

3. In the presence of noise, $P_d(\epsilon)$ will no longer saturate to 1 as ϵ drops below Δr. Eq. (5) establishes a relation between the unknown distribution of the residuals $\rho(|\Delta r|)$ and $P_d(\epsilon)$

$$\rho(|\Delta r|) = -\frac{d}{d|\Delta r|}\mathrm{Prob}(|\Delta r'| > |\Delta r|) = \left[\frac{d}{d\epsilon}P_d(\epsilon)\right]_{\epsilon=|\Delta r|}. \tag{6}$$

   Thus $\langle(\Delta r)^2\rangle$ can be computed from $P_d(\epsilon)$ using eqs. (5, 6) and partial integration. For independent and identically distributed random numbers ($\sigma_r^2 = 1/2\langle(\Delta r)^2\rangle$) one obtains

$$\sigma_r^2 = \int_0^\infty d\epsilon\;\epsilon\,[1 - P_d(\epsilon)] \tag{7}$$

How does $P_d(\epsilon|\,\delta)$ vary as a function of δ for fixed ϵ? For $\delta \to \infty$ the conditions have no effect and one has $P_d(\epsilon|\,\delta)|_{\delta\to\infty} = P_0(\epsilon)$. As $\delta \to 0$, $P_d(\epsilon|\,\delta)$ should increase monotonically and saturate to 1 for $d \geq d_0$, where $d_0 + 1$ is the minimum embedding dimension. This behavior is shown schematically in fig. 1a. $P_d(\epsilon)$ measures how well the dynamics can be modeled in terms of the d variables. To quantify the dependence on each of the variables, it is convenient to define a *dependability index*

$$\lambda_d(\epsilon) = \frac{P_d(\epsilon) - P_{d-1}(\epsilon)}{1 - P_0(\epsilon)}, \quad d = 1, 2, \ldots \tag{8}$$

or its average over ϵ, $\bar{\lambda}_d$. For a noise-free deterministic map, $P_d(\epsilon)$ saturates to 1 for $d \geq d_0$ and one has

$$\sum_{d=1}^{d_0} \overline{\lambda}_d = \sum_{d=1}^{d_0} \lambda_d(\epsilon) = 1 \tag{9}$$

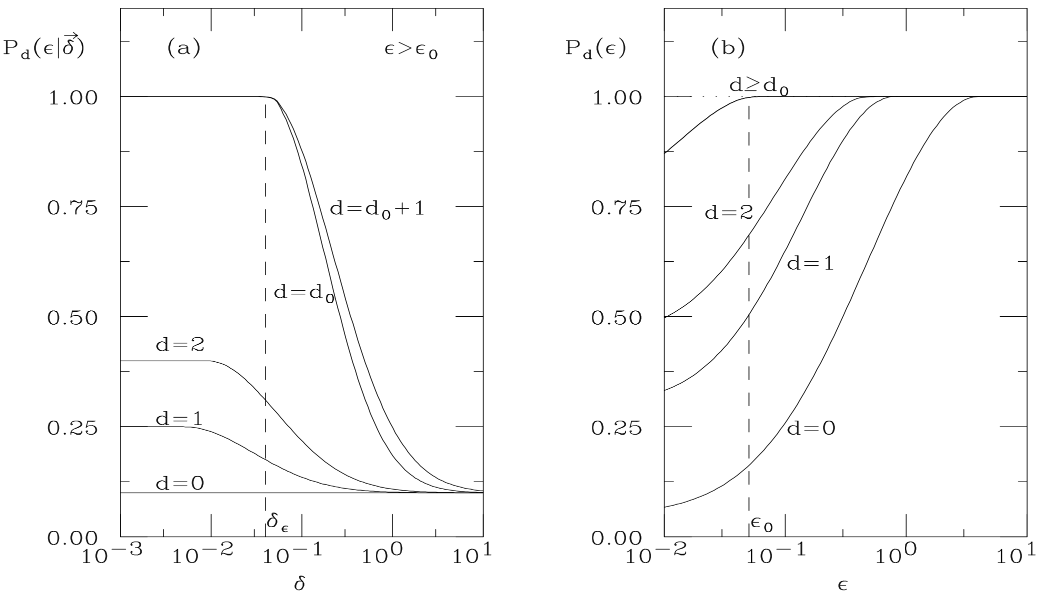

Figure 1: (a). $P_d(\epsilon|\,\delta)$ as a function of δ for fixed ϵ. (b). The maxima $P_d(\epsilon)$ as a function of ϵ. Saturation to 1 would be observed for $d \geq d_0$. In the presence of noise the saturation deviates from 1 around $\epsilon_0 \sim \Delta r_{max}$.

The formalism above assumes an 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. One then estimates errors using the standard methods [3]. Note that the integrand in eq. (7) suppresses the small ϵ region, which is desirable in limited statistics situations.

Applications

We demonstrate the power of the method for three synthetic time series problems where $y = x_t$ and $x_1 = x_{t-1}, \ldots, x_d = x_{t-d}$.

Logistic map data is generated with 4000 time steps according to

$$x_t = \eta x_{t-1}(1 - x_{t-1}) + r_t \tag{10}$$

with $\eta = 4$. The iterative noise r_t is 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 $x_t \in (0, 1)$. One obtains $\overline{\lambda}_1 = 1$ for **A**, $\overline{\lambda}_1 = 0.97$ for **B** and $\overline{\lambda}_{d\geq 2} \approx 0$. Note that the method yields signals for primary dependencies, not induced ones. Hence it is satisfying that one gets $\overline{\lambda}_{d\geq 2} \approx 0$ in this case. This is in contrast to covariance and mutual information [6] methods. Using eq. (7) to extract the noise variance gives $r = 0$ and $r = 0.27$ consistent with the generated data.

Hénon map data with dependences on larger lags than usual is generated with 4000 time-steps according to

$$x_t = 1 - a(x_{t-2} - r_{t-2})^2 + b(x_{t-4} - r_{t-4}) + r_t \tag{11}$$

with $a = 1.4$, and $b = 0.3$. Again, two data sets are generated with $r = 0$ (**A**) and $r = 0.14\sigma$ (**B**). One obtains $\overline{\lambda}_{1-4}$ = 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 $\overline{\lambda}_4$. The two noise levels are reproduced approximately by eq. (7).

The **Ikeda map** [12] describes the evolution of a laser in a ring cavity with a lossy active medium. In terms of the complex variable $z(t) = x(t) + i\,y(t)$, the map is defined by

$$z(t+1) = p + B\,z(t)\exp[i\kappa - \frac{i\alpha}{1 + |z(t)|^2}]. \tag{12}$$

Sets of $N = 2000$ data points are generated using eq. (12) with the parameters $p = 1.0$, $B = 0.9$, $\kappa = 0.4$ and $\alpha = 6.0$, and with Gaussian noise added to the x component at the each iteration as $x(t) = x(t) + r$ with standard deviations σ_r=0.0, 0.01 and 0.03 respectively. The results from applying the method on the Ikeda map are shown in table

σ_r	0.00		0.01		0.03	
σ_r/σ	0.0000		0.0208		0.5621	
Variables	$(\hat{\sigma}_r)_{LR}$	$(\hat{\sigma}_r)_{NL}$	$(\hat{\sigma}_r)_{LR}$	$(\hat{\sigma}_r)_{NL}$	$(\hat{\sigma}_r)_{LR}$	$(\hat{\sigma}_r)_{NL}$
$\{\text{none}\}$	1.000	1.003	1.000	1.003	1.000	1.001
$\{x_{t-1}\}$	0.997	0.819	0.998	0.792	0.642	0.644
$\{x_{t-1},\, y_{t-1}\}$	0.887	0.0055	0.889	0.021	0.557	0.563

Table 1: Regression errors on $x(t)$ expressed as fractional errors $\hat{\sigma}_r$ for various sets of variables. The subscripts **LR** and **NL** stand for linear regression (eq. (2)) and the method of eq. (7)) respectively. Due to the effect of the noise, the noise fraction σ_r/σ varies considerably for differing noise levels.

1. One concludes that the method quite accurately estimates the noise fraction σ_r/σ using the variable set $\{x_{t-1},\, y_{t-1}\}$. In the case of $\sigma_r = 0.0$, the linear regression model gives a noise level 0.887, while the current method identifies a negligible noise level (0.005). This indicates that the dependency of x_t on x_{t-1} and y_{t-1} is predominantly nonlinear. Such a signature of nonlinearity exists as long as the noise level is modest – below $\sigma_r = 0.02$ in this case. This is consistent with what can be seen in fig. 2, where the nonlinear structure clearly disappears in (d) when the noise reaches $\sigma_r = 0.03$.

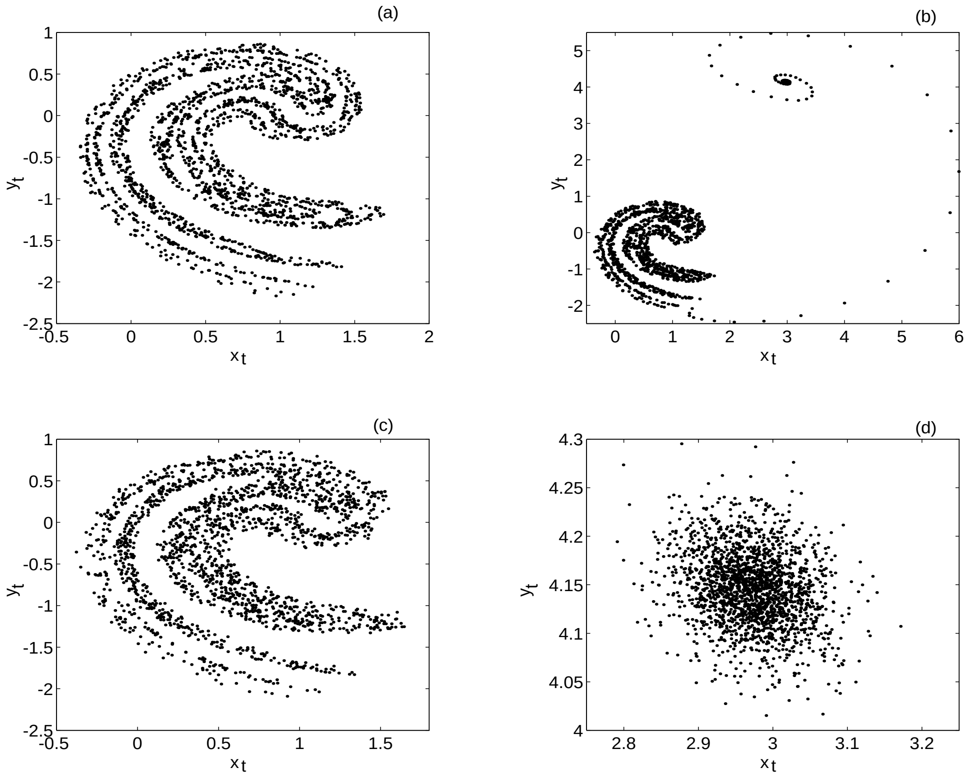

Figure 2: The Ikeda map shown in its x-y phase space. A Gaussian noise term with standard deviation $\sigma_r = 0.0$ (a), 0.01 (b), 0.02 (c) and 0.03 (d) is added iteratively to the $x(t)$-component.

Summary

The δ-test is efficient 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. Automated procedures for setting bin sizes, cutoffs etc. and error analysis are feasible and public domain software exists [13, 14]. It has been profitably exploited in selecting relevant time lags for ANN processing in connection with sunspot data [3], in predicting utility consumptions [10] and in analysis of ECG data [11]. The method has been successfully illustrated here with time series examples. The δ-test of course works equally well in cases with "horizontal dependencies" – variables measured at equal times. The general method for extracting noise variances assumes nothing about the noise distributions – but these can also be extracted [4]. By comparing the obtained noise variances with those derived from assumed linear dependencies, signals of nonlinearities are obtained.

References

[1] See e.g. J. D. Hamilton, "Time Series Analysis", Princeton University Press (Princeton 1994).

[2] S.J. Nowlan and G. Hinton, *Neural Computation* **4**, 473 (1992).

[3] H. Pi and C. Peterson, *Neural Computation* **6**, 509 (1994).

[4] H. Pi and C. Peterson, "Estimating Nonlinear Regression Errors without Doing Regression" *LU TP 94-19*.

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

[6] A. M. Fraser, *IEEE Trans. Info. Theory* **IT-35**, 245 (1989).

[7] P. Grassberger and I. Procaccia, *Physica* **D 9**, 189 (1983).

[8] W.A. Brock et al., "A Test for Independence Based on the Correlation Dimension", University of Wisconsin Technical Report (1988).

[9] R. Savit and M. Green, *Physica* **D 50**, 95 (1991).

[10] M. Ohlsson, C. Peterson, H. Pi, T. Rögnvaldsson and B. Söderberg, "Predicting Utility Loads with Artificial Neural Networks – Methods and Results from the Great Energy Predictor Shootout", *1994 Annual Proceedings of the American Society of Heating, Refrigerating and Air-Conditioning Engineers, Inc.*, 1063 (1994).

[11] B. Heden, M. Ohlsson, L. Edenbrandt, R. Rittner, O. Pahlm and C. Peterson, "Artificial Neural Networks for Recognition of Electrocardiographic Electrode Misplacement", *American Journal of Cardiology* **75**, 929 (1995).

[12] K. Ikeda, *Opt. Commun.* **30**, 257 (1979).

[13] H. Pi and C. Peterson, "Delta 2.0 – A Program for Finding Dependencies in Tables of Data", *Computer Physics Communications*, **83**, 293 (1994).

[14] H. Pi and C. Peterson, `Delta 3.0`, available with anonymous ftp at `thep.lu.se` in directory `pub/LundPrograms` – `delta.tar.Z`.