LU TP 94-19
October 1994

Estimating Nonlinear Regression Errors without Doing Regression

Hong Pi[1]

Department of Computer Science & Engineering, Oregon Graduate Institute, P.O. Box 91000, Portland, Oregon 97291-1000

Carsten Peterson[2]

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

Abstract:

A method for estimating nonlinear regression errors and their distributions without performing regression is presented. Assuming continuity of the modeling function the variance is given in terms of conditional probabilities extracted from the data. For N data points the computational demand is N^2.The method is successfully illustrated with data generated by the Ikeda and Lorentz maps augmented with noise. As a by-product the embedding dimensions of these maps are extracted. Comparing the predicted residual errors with those from linear models provides a signal for nonlinearity.

PACS numbers: 05.45.+b, 05.40.+j

[1] pihong@cse.ogi.edu
[2] carsten@thep.lu.se

Background

Most measurements of physical processes are noisy. This is often due to the fact that all independent variables have not been measured. Being able to estimate the noise distribution and its variance directly from data with no assumptions about the underlying signal function is most desirable. It would provide a natural step prior to any modeling of a system (e.g. artificial neural network) since one then knows the optimal performance limit of the fit in advance. Furthermore, methods for filtering data often require prior estimate of noise variance.

To be more concrete, given 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 estimating 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 estimating σ_r^2 are model-based. One fits the data to a model, a particular choice of F, and then interprets the deviation of the fit as noise. In the special case of linear regression models [1] where F takes the form

$$\hat{y} = a_0 + \sum_{k=1}^{d} a_k x_k, \tag{2}$$

a sample estimate for σ_r^2 is explicitly given by

$$\sigma_r^2 = \sigma^2 - \sum_{k=1}^{d} a_k \langle y, x_k \rangle \tag{3}$$

where σ denotes the y-variable variance and the angled brackets covariances. In this letter we devise a method for estimating the optimum σ_r when the modeling function F is not restricted to be linear. The estimate does not rely on any conjecture about the form of F. The only assumption is that F is uniformly continuous. Loosely speaking we extract probability densities from the data and by sampling data with decreasing bin sizes such that any noise point will appear as a discontinuity. The concept of using the requirement of continuity for establishing dependencies on x_k was previously explored in the δ-test method [2], where noise levels had to be estimated by making assumptions about the probability distributions of r. The novel and unique feature of the present approach is that no assumption about the form of r is needed – σ_r^2 is computed directly as an integral over data densities, and the distribution of the noise can also be identified.

We illustrate the power of the method with two examples of chaotic time series augmented with noise: the Ikeda [3] and Lorentz [4] maps.

In addition to finding noise levels, the method also provides means for determining embedding dimensions and for establishing nonlinearities.

Method

The goal is to derive a statistical estimate on the variance of r for the optimum model describing eq. (1). By optimum we mean a model F (eq. (1)) such that r and $\hat{y}$ are uncorrelated and that r represents identically and independently distributed (**i.i.d**) noise. Most adaptive algorithms such as neural network models are designed to find such an optimum function.

We use the conditional probabilities defined in the δ-test [2] – for a pair of positive real numbers ϵ and δ, one constructs directly from the data the conditional probability

$$P(\epsilon|\ \delta) \equiv P(|\Delta y| \leq \epsilon \mid |\Delta \mathbf{x}| \leq \delta) \tag{4}$$

where $|\Delta \mathbf{x}| \equiv \max_k |x_k - x'_k|$. In the limit $\delta \to 0$, one obtains

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

where the property of function continuity, $F(\mathbf{x}) - F(\mathbf{x}') \to 0$ for $\mathbf{x} \to \mathbf{x}'$, is exploited. Eq. (5) establishes a connection between the probability distribution of the residuals $\rho(|\Delta r|)$ to the quantity $P(\epsilon)$, which is directly calculable from the data sample.

The probability density of the residual separation $|\Delta r|$ is given by

$$\begin{aligned} \rho(|\Delta r|) &= -\frac{d}{d|\Delta r|}\text{Prob}(|\Delta r'| > |\Delta r|) \\ &= \left[\frac{d}{d\epsilon}P(\epsilon)\right]_{\epsilon = |\Delta r|} . \end{aligned} \tag{6}$$

Thus $\langle |\Delta r| \rangle$ and $\langle (\Delta r)^2 \rangle$ can be related to $P(\epsilon)$ using eqs. (5, 6). Using partial integration one obtains

$$\langle |\Delta r| \rangle = \int_0^\infty d\epsilon\ [1 - P(\epsilon)] \tag{7}$$

and

$$\langle (\Delta r)^2 \rangle = 2 \int_0^\infty d\epsilon\ \epsilon\, [1 - P(\epsilon)] \tag{8}$$

respectively. If r is **i.i.d** one has $\langle(\Delta r)^2\rangle = 2\sigma_r^2$. Our estimate for the residual variance of the optimum model is then given by

$$\sigma_r^2 = \int_0^\infty d\epsilon\, \epsilon\, [1 - P(\epsilon)]. \tag{9}$$

We note that the integrand in eq. (9) suppresses the small ϵ region. This feature is desirable in limited statistics situations with few high resolution (small ϵ) data points.

Eq. (6) shows that $P(\epsilon)$ measures the cumulative distribution of the noise separations. It is then possible to discern the noise distribution through $P(\epsilon)$. In particular if the noise is Gaussian with standard deviation σ_r, $P(\epsilon)$ should be

$$P(\epsilon) = \int_0^\epsilon \rho(|\Delta r|)\, d|\Delta r| = erf(\frac{\epsilon}{2\sigma_r}), \tag{10}$$

where $erf(.)$ is the error function.

Eq. (8) is equivalent to calculating the expectation value [5],

$$\sigma_r^2 = E[(\Delta y)^2|\, |\Delta\mathbf{x}| \leq \delta]_{\delta\to 0}. \tag{11}$$

which may be easier to implement numerically, if one only wants to estimate σ_r^2.

So far we have assumed an infinite amount of data. Some implementation issues are pertinent with limited statistics. In what follows each variable is transformed to zero mean and unit standard deviation.

Consider a subset of the data $S(\delta)$ defined as

$$S(\delta) = \left\{(y^{(j)}, \mathbf{x}^{(j)})|\ \exists k \neq j,\ \ |\mathbf{x}^{(j)} - \mathbf{x}^{(k)}| \leq \delta\right\}. \tag{12}$$

where each data point has at least one "neighbor". Clearly $S(\delta)$ is the set of points relevant for the probabilities defined in eqs. (4, 5) and for the expectation value in eq. (11). Denote the number of points in $S(\delta)$ by $n(\delta)$ and $N(\delta)$ as the number of *pairs* of points satisfying $|\mathbf{x}^{(i)} - \mathbf{x}^{(i')}| \leq \delta$. When both $n(\delta)$ and $N(\delta)$ are lower than N_{min}, δ should not be decreased further.

$S(\delta)$ varies in the $\delta \to 0$ limit. In extreme cases one may end up sampling a very small cluster of data points. The constructed probability densities are then far off from the true densities. To reduce the effect of this sampling bias, we monitor $\sigma(\delta)$, the standard deviation of y for the points in $S(\delta)$. Ideally the point distribution of $S(\delta)$ should resemble that of the entire data set. When $\sigma(\delta)$ falls below σ_{cut}, we conclude that $S(\delta)$ is sufficiently biased and the statistics in e.g. eq. (4) are deemed unacceptable. Throughout this paper we use $N_{min} = 30$ and $\sigma_{cut} = 0.98$.

For the expectation value of eq. (11), instead of cutting off the statistics at $\sigma(\delta) \leq \sigma_{cut}$, the effect of the sampling bias can be approximately corrected for by the modification

$$\sigma_r^2 = \left\{ \frac{1}{\sigma^2(\delta)} E[(\Delta y)^2 | \, |\Delta \mathbf{x}| \leq \delta] \right\}_{\delta \to 0} \tag{13}$$

A suitable binning of the $\log \epsilon$-$\log \delta$ plane for evaluating the probabilities and estimating their statistical errors is given in [2]. In fig. 1 a typical example of $P(\epsilon| \, \delta)$ is shown. For a fixed ϵ, $P(\epsilon| \, \delta)$ rises and reaches a plateau as δ decreases. $P(\epsilon)$ is determined by the largest point with acceptable statistics in the plateau region.

$$P(\epsilon) \equiv \max_{\delta > 0} P(\epsilon| \, \vec{\delta}) \tag{14}$$

The integral of eq. (9) is easily computed with e.g. the Simpson method. A package in C for performing what is reported in this letter is available [6].

Explorations

The Ikeda Map

The Ikeda map [3] 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

$$z(t+1) = p + B\, z(t) \exp[i\kappa - \frac{i\alpha}{1 + |z(t)|^2}]. \tag{15}$$

Sets of $N = 2000$ data points are generated using eq. (15) with the parameters $p = 1.0$, $B = 0.9$, $\kappa = 0.4$ and $\alpha = 6.0$ [11], 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, 0.02 and 0.03 respectively.

We now apply our method to estimate what the error would have been if we had regressed $x(t)$ on various sets of explanatory variables. The results are shown in table 1, from which we conclude that our method gives quite accurate estimates on the noise fraction σ_r/σ when the explanatory variable set $\{x_{t-1}, \, y_{t-1}\}$ is assumed. In the case of $\sigma_r = 0.0$, the linear regression model gives a noise level 0.887, while our 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 $\sigma_r = 0.03$.

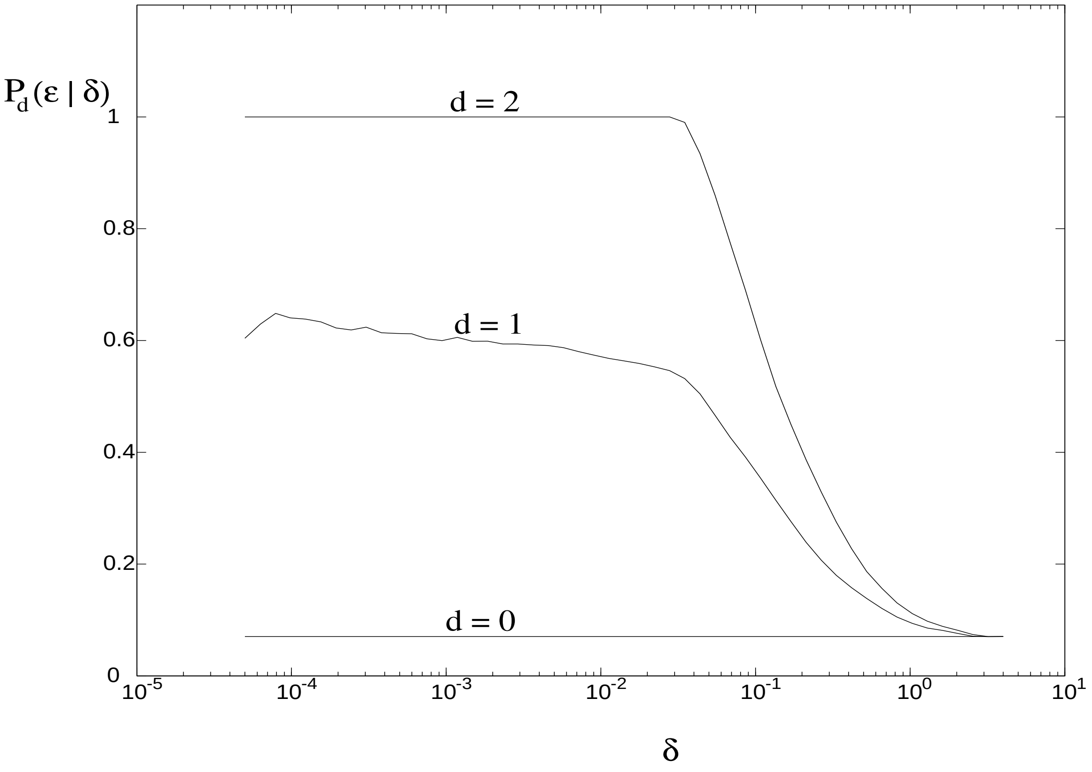

Figure 1: $P_d(\epsilon|\vec{\delta})$ as a function of δ at a fixed $\epsilon = 0.108$ for the Hénon map $[x(t) = 1 - 1.4x_{t-1}^2 + 0.3x_{t-2}]$ with N=3000 data points

In fig. 3 we compare $P(\epsilon)$ from the explanatory variables $\{x_{t-1}, y_{t-1}\}$ with the expression (eq. (10)) for Gaussian noise. Significant deviation between the two would have indicated the noise distribution as non-Gaussian.

Let us next turn to the challenge of determining the embedding dimension. If we assume that the only data we have are the observations of the x component, our variance estimate can be used to identify the minimum embedding dimension, in a procedure similar to the δ-test [2]. What we need to do is to find the (smallest) set of variables that minimizes the residual error. We list the results in table 2, from which we see that the residual error ceases to decrease at $k = 4$. Therefore we identify the embedding dimension $d_E = 5$ for the Ikeda map, as long as the noise level is not too high ($\sigma_r < 0.03$). Using the *False Nearest Neighbors* method, one finds $d_E = 4$ [11]. As shown in the table 2, $d_E = 4$ would result in quite small a residual error of 0.077 and therefore provides a fairly good embedding. However, $d_E = 5$ is a better choice.

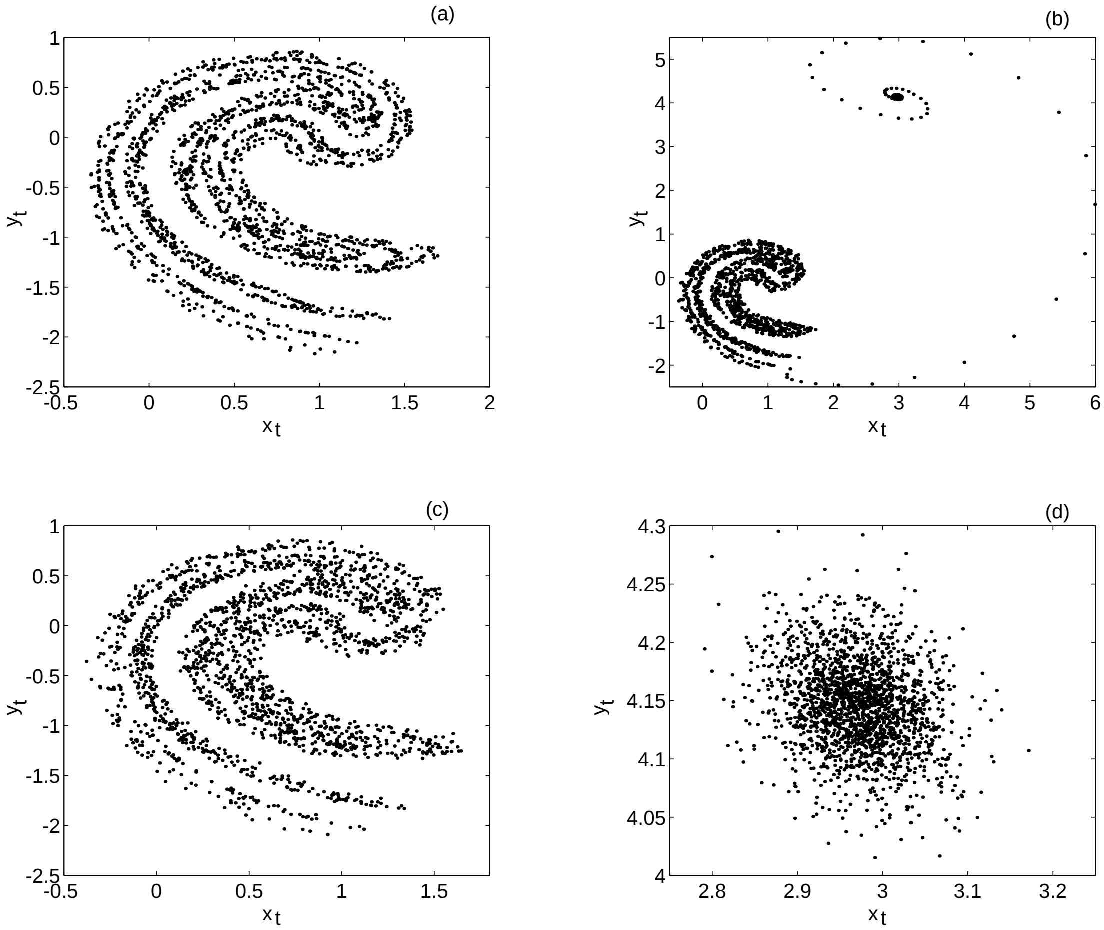

Figure 2: The Ikeda map shown in its x-y phase space. A Gaussian noise term with standard deviation $\sigma_r = 0.0$ (a), 0.01 (b), 0.02 (c) and 0.03 (d) is added iteratively to the $x(t)$-component.

We observe that in the results presented above, the error estimate σ_r is very close to 1.0 when the explanatory variable set is $\{none\}$, as it should be. This provides a consistency check on the applicability of the method on the particular data set. When a significant deviation from 1 is observed, it indicates that some of the assumptions of the method, e.g. stationarity, is violated in the data set. Also note that the noise estimates $\hat{\sigma}_r$ in table 2 does not match the applied relative noise σ_r/σ exactly, since the regression equation is different from the actual generating process.

σ_r	0.00		0.01		0.02		0.03	
σ_r/σ	0.0000		0.0208		0.0424		0.5621	
Variables	$(\hat{\sigma}_r)_{LR}$	$(\hat{\sigma}_r)_{NL}$	$(\hat{\sigma}_r)_{LR}$	$(\hat{\sigma}_r)_{NL}$	$(\hat{\sigma}_r)_{LR}$	$(\hat{\sigma}_r)_{NL}$	$(\hat{\sigma}_r)_{LR}$	$(\hat{\sigma}_r)_{NL}$
{none}	1.000	1.003	1.000	1.003	1.000	1.003	1.000	1.001
$\{x_{t-1}\}$	0.997	0.819	0.998	0.792	0.994	0.813	0.642	0.644
$\{x_{t-1},\ y_{t-1}\}$	0.887	0.0055	0.889	0.021	0.884	0.044	0.557	0.563

Table 1: Regression errors on $x(t)$ expressed as fractional errors $\hat{\sigma}_r$ for various sets of explanatory variables. The subscripts **LR** and **NL** stand for linear regression (eq. (3)) and our method allowing for nonlinear dependencies (eq. (9)), respectively. Due to the effect of the noise, the noise fraction σ_r/σ varies considerably for differing noise levels.

σ_r	0.00		0.02		0.03	
σ_r/σ	0.0000		0.0424		0.5621	
Variables	$(\hat{\sigma}_r)_{LR}$	$(\hat{\sigma}_r)_{NL}$	$(\hat{\sigma}_r)_{LR}$	$(\hat{\sigma}_r)_{NL}$	$(\hat{\sigma}_r)_{LR}$	$(\hat{\sigma}_r)_{NL}$
{none}	1.000	1.003	1.000	1.003	1.000	1.001
$\{x_{t-1}\}$	0.997	0.819	0.994	0.813	0.642	0.644
$\{x_{t-1},\ x_{t-2}\}$	0.954	0.505	0.950	0.623	0.622	0.636
$\{x_{t-k}\ k=1,..3\}$	0.946	0.077	0.946	0.195	0.611	0.635
$\{x_{t-k}\ k=1,..4\}$	0.936	0.025	0.931	0.074	0.603	0.636
$\{x_{t-k}\ k=1,..5\}$	0.934	0.026	0.931	0.077	0.598	0.645

Table 2: Regression error on $x(t)$ for various sets of explanatory variables. Same notation as in table 1.

The Lorentz Map

A system described by the Lorentz equations [4], which describe meteorological physics,

$$\begin{aligned} \frac{dx(t)}{dt} &= \sigma[-x(t)+y(t)] \\ \frac{dy(t)}{dt} &= rx(t)-y(t)-x(t)z(t) \\ \frac{dz(t)}{dt} &= x(t)y(t)-bz(t) \end{aligned} \tag{16}$$

can display low-dimensional chaotic behavior when the parameters are chosen from the chaotic regime [11]. We adopt the parameters $r = 45.92$, $b = 4.0$ and $\sigma = 16.0$ [11], and solve the equations with $\Delta t = 0.1$ using the fourth and fifth order Runge-Kutta methods. The method is applied iteratively such that the solutions at t are used as the initial values

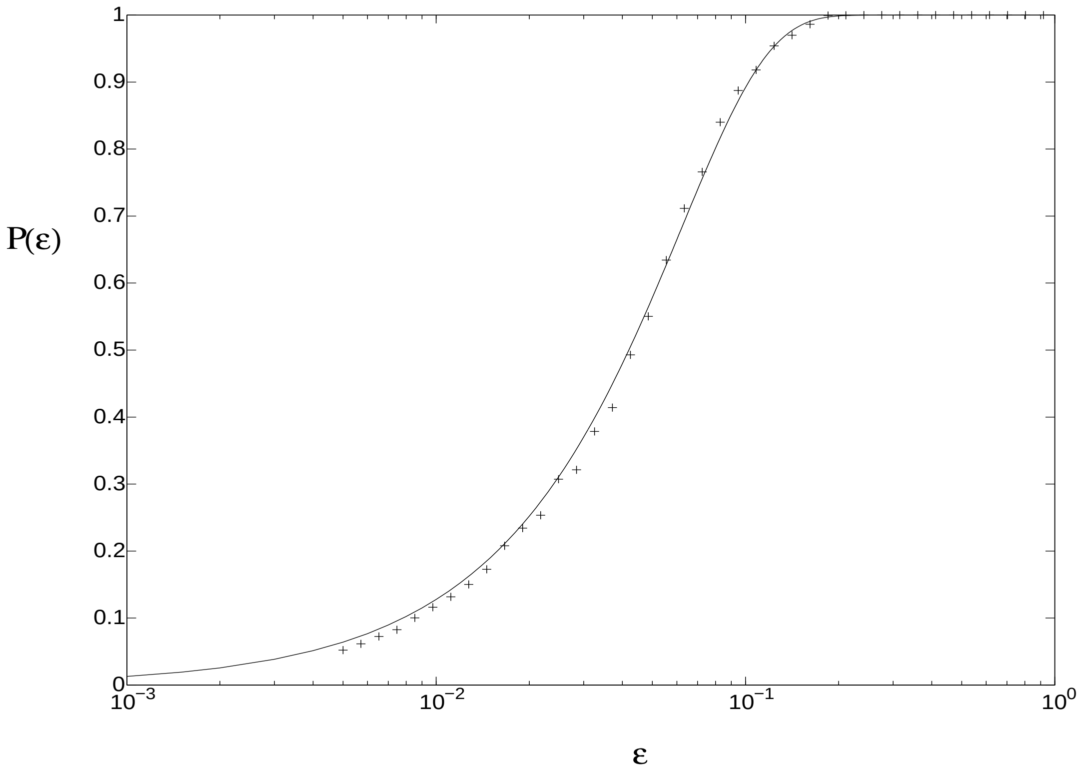

Figure 3: $P(\epsilon)$ versus ϵ for the Ikeda map with Gaussian fractional noise $\sigma_r/\sigma = 0.0424$. The symbols are the values calculated from the data. The line is the analytic expression in eq. (10).

to the differential equations to obtain the values at $t + 0.1$. A portion of the data set is shown in fig. 4.

Gaussian noise of various variances are superimposed onto the original clean data set. The results of the variance estimates are shown in table 3. Nonlinearities are evident from the significant differences between the linear regression errors and the nonlinear estimates. Based on the values of $(\hat{\sigma}_r)_{NL}$ we conclude that three time lag variables are needed to map the variable $x(t)$. Hence one has $d_E = 4$ for the Lorentz map.

The Gaussian noise we imposed have variances of 0.0, 0.5, and 1.0, corresponding to the fractional variances of 0.0, 0.0016, and 0.0062 respectively, which are significantly lower than the estimated fractional residual variances. This is not an inconsistency since the superimposed noise should not be confused with the residual noise of the model as defined

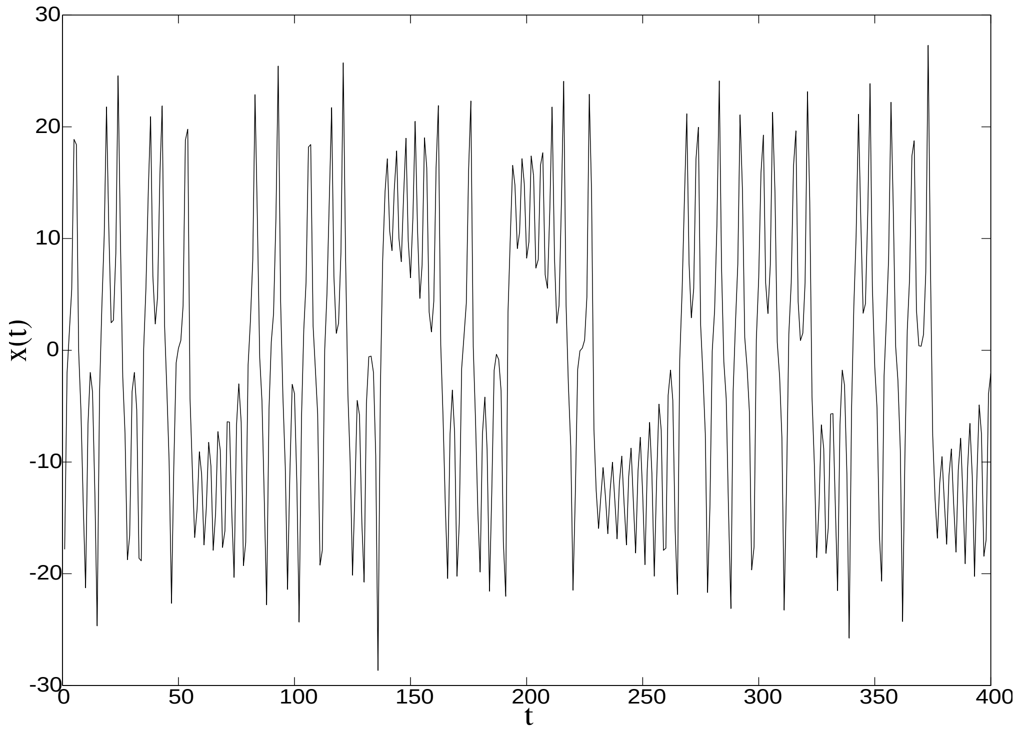

Figure 4: The $x(t)$ component of a Lorentz system as a function of t.

in eq. (1). In this case we are dealing with noisy inputs. The equation becomes

$$\begin{aligned} x(t+1) &= F[x(t)-r(t), x(t-1)-r(t-1), ..] + r(t+1) \\ &= F'[x(t), x(t-1), ..] + r'(t+1). \end{aligned} \tag{17}$$

What the method yields is the variance of the **effective noise** r', which can be quite different from the variance of the superimposed noise r.

Summary

We have developed a general method that efficiently extracts noise variances from raw data with no assumptions about the noise distributions. The method handles nonlinear dependencies provided the underlying function is uniformly continuous.

By comparing the extracted noise variances with those derived from assumed linear de-

σ_r	0.00		0.5		1.0	
σ_r/σ	0.0000		0.0016		0.0062	
	$(\hat{\sigma}_r)_{LR}$	$(\hat{\sigma}_r)_{NL}$	$(\hat{\sigma}_r)_{LR}$	$(\hat{\sigma}_r)_{NL}$	$(\hat{\sigma}_r)_{LR}$	$(\hat{\sigma}_r)_{NL}$
$k=0$	1.000	1.000	1.000	1.000	1.000	1.000
$k=1$	0.670	0.520	0.691	0.517	0.693	0.544
$k=2$	0.653	0.084	0.653	0.110	0.657	0.197
$k=3$	0.640	0.01	0.642	0.084	0.646	0.152
$k=4$	0.640	0.008	0.641	0.084	0.646	0.158
$k=5$	0.634	0.008	0.635	0.084	0.640	0.141

Table 3: Regression errors on $x(t)$ expressed as fractional errors $\hat{\sigma}_r$ for various numbers of time lag variables for the Lorentz map. σ_{LR} gives the linear regression residual error. σ_{NL} is the nonlinear estimate from eq. (13).

pendencies, signals of nonlinearities are obtained.

Estimating the variance is very useful for model selection. As a by-product the embedding dimensions are obtained in a way slightly different from that of ref. [2].

We have illustrated the method with two time series examples. The method of course also works in cases with "horizontal dependencies" – variables measured at equal times.

Existing approaches to determine dependencies aimed beyond the linear regime are either based on entropy measures [12, 13] or on elaborate autocorrelation measures [8, 9, 10]. The *Mutual Information* approach [13] has the shortcoming that it does not disentangle primary dependencies from induced ones. Furthermore, noise levels are not directly extractable. On the other hand its computational effort scales favorable with the number of data points as $N \log N$ as compared to N^2 with the proposed method. Our approach has its roots in autocorrelation methods [8, 9, 10], but is conceptionally very distinct from these since it is based on continuity. For that reason, in contrast to refs. [8, 9, 10], it extracts the noise levels and ignores induced dependencies.

Acknowledgements

This work was supported in part by the the Swedish Board for Industrial and Technical Development (NUTEK) and the Göran Gustafsson Foundation for Research in Natural Science and Medicine. One of the authors (Pi) gratefully acknowledges the funding provided by the Advanced Research Projects Agency and the Office of Naval Research under grant ONR N000-14-92-J-4062. Pi would also like to thank Brian D. Ripley for helpful discussions.

References

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

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

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

[4] E.N. Lorentz map , *J. Atmos. Sci.* **20**, 130 (1963).

[5] This was suggested to us by Dr. B. D. Ripley.

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

[7] M. Hénon, *Commun. Math. Phys.* **50**, 69 (1976).

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

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

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

[11] H. Abarbanel et al., *Rev. Mod. Phys.* **65**, 1331 (1993).

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

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