Available online at www.sciencedirect.com
ScienceDirect
Physica D 229 (2007) 1–8
PHYSICA D
www.elsevier.com/locate/physd

A note on the Gamma test analysis of noisy input/output data and noisy time series

Antonia J. Jones[a], D. Evans[a], S.E. Kemp[b,*]

[a] School of Computer Science, Cardiff University, PO Box 916, Cardiff CF24 3XF, UK
[b] Department of Computing and Mathematical Sciences, Faculty of Advanced Technology, University of Glamorgan, Pontypridd, Wales, CF37 1DL, UK

Received 9 November 2005; received in revised form 27 September 2006; accepted 18 December 2006
Available online 28 February 2007
Communicated by C.K.R.T. Jones

Abstract

In a smooth input/output process $y = f(\boldsymbol{x})$, if the input data $\boldsymbol{x} \in \mathbb{R}^d$ is noise free and only the output data y is corrupted by noise, then a near optimal smooth model $\hat{g}$ will be a close approximation to f. However, as previously observed, for example in [H. Kantz, T. Schreiber, Nonlinear Time Series Analysis, 2nd ed., Cambridge Univ. Press, 2004], if the input data is also corrupted by noise then this is no longer the case. With noise on the inputs, the best predictive smooth model based on noisy data need not be an approximation to the actual underlying process; rather, the best predictive model depends on both the underlying process *and* the noise. A corollary of this observation is that one cannot readily infer the nature of a process from noisy data. Since almost all data has associated noise this conclusion has some unsettling implications. In this note we show how these effects can be *quantified* using the Gamma test.

In particular we examine the Gamma test analysis of noisy time series data. We show that the noise level on the best predictive smooth model (based on the noisy data) can be much higher than the noise level on individual time series measurements, and we give an upper bound for the first in terms of the second.

Keywords: The Gamma test; Non-linear analysis; Process identification; Noise estimation; Time series

1. Introduction

In the analysis of time series, we often hypothesize that the variable of interest is just one of a number of variables of a complex dynamic system, described by a system of differential equations. Following the work of Takens [10] we seek to predict the next value (output) based on a number d of previous values (input). In this context, the input is called a *delay vector* and d is called the *embedding dimension*.

For a time series (z_t), Takens' theorem [10] and its subsequent extensions ensure, under a broad range of circumstances, that there exists a *smooth* function f with bounded partial derivatives such that

$$z_t = f(z_{t-1}, z_{t-2}, \ldots z_{t-d}) \tag{1}$$

which, provided d is sufficiently large to unfold the dynamics, can be used as the basis for a recursive one-step prediction.

By *smooth function*, we mean throughout that f and its partial derivatives of first (and possibly higher) orders exist, are continuous over a compact region, and are therefore bounded. To be explicit we suppose $|\nabla f|^2 \leq B$ over the region in question.

1.1. Stochastic time series

We draw a distinction between the subject of our paper – what we have called *noisy* time series – and that of *stochastic* time series. For a univariate stochastic time series with additive noise (often assumed to be Gaussian), the process is defined according a recursive rule of the form

$$z_t = f(z_{t-1}, z_{t-2}, \ldots z_{t-d}) + e_t \tag{2}$$

where f is a smooth function and e_t is a realization of some random variable (if f is linear, these are called *linear* stochastic

* Corresponding address: The Hollies, Curtis Lane, Headley, GU35 8PH Bordon, Hampshire, UK. Fax: +44 0 7880740220.
E-mail address: kemp.samuel@googlemail.com (S.E. Kemp).

0167-2789/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.physd.2006.12.013

time series). The significant fact is that the noise e_{t-1} associated with the *previous* value z_{t-1} of the time series feeds through to affect the next value z_t, so this noise plays a role in determining the evolution of the time series.[1]

1.2. Noisy time series

We consider the case where a noise-free time series (z_t) is observed under additive noise, i.e.

$$y_t = z_t + r_t \qquad (t = 1, 2, 3, \ldots, M) \tag{3}$$

where the *true* values z_t are subject to independent and identically distributed random perturbations r_t having expectation zero.

We assume that the noise-free value z_t is determined by a smooth function $f : \mathbb{R}^d \to \mathbb{R}$ of some number d of the previous noise-free values $z_{t-1}, \ldots, z_{t-d}$,

$$z_t = f(z_{t-1}, \ldots, z_{t-d}). \tag{4}$$

Thus we imagine that in reality, the variable z_t (part of a high-dimensional non-linear dynamical process) is evolving according to an unknown but smooth rule such as (4), but that what we *actually* observe, typically as a consequence of measurement error, are the corrupted values $y_t = z_t + r_t$. Importantly, and in contrast with stochastic time series, the noise associated with previous values such as y_{t-1} does *not* feed through to affect the value y_t.

Of course, in many real world situations a time series may be *both* stochastic and noisy. However, here we seek to examine just those features specifically relating to noisy time series.

In the context of non-linear dynamic systems time series state space reconstruction, the noise that we have considered is termed 'observational noise'. The question of optimal prediction for time series under observational noise is also considered as a special case in Casdagli et al. [1], which studies in considerable detail the more general issues surrounding state space reconstruction under noise.

1.3. Effective noise

For $d \in \mathbb{N}$, let $\boldsymbol{x}_{d+1}, \ldots, \boldsymbol{x}_M$ denote the noisy delay vectors:

$$\boldsymbol{x}_t = (y_{t-1}, \ldots, y_{t-d}) \in \mathbb{R}^d. \tag{5}$$

Using only the noisy time series data $(\boldsymbol{x}_t, y_t)$, we seek to identify a smooth function $g : \mathbb{R}^d \to \mathbb{R}$ that 'best explains' the observed behaviour of the time series. We first clarify what is meant by 'best explains', i.e. what is an optimal smooth model in this context.

Let $S = \{h : \mathbb{R}^d \to \mathbb{R} \mid h \text{ smooth}, |\nabla h|^2 \leq B\}$, i.e. S is the class of smooth functions in the sense described earlier. For each h consider the mean squared error

$$\mathrm{MSE}(h) = \mathcal{E}((y - h(\boldsymbol{x}))^2) \tag{6}$$

where the expectation is taken over all realizations of the input/output pair $(\boldsymbol{x}, y)$. The set of *optimal* predictive smooth data models is defined to be

$$S_{\mathrm{opt}} = \{g \in S : \mathrm{MSE}(g) \leq \mathrm{MSE}(h) \text{ for all } h \in S\}. \tag{7}$$

Let $g \in S_{\mathrm{opt}}$. We write

$$y = g(\boldsymbol{x}) + R \tag{8}$$

where R is a zero-mean random variable, called the *effective noise* on the output, which accounts for all variation in the output that cannot be accounted for by *any* smooth transformation of the input.

Note that

$$\mathcal{E}(R^2) = \mathcal{E}((y - g(\boldsymbol{x}))^2) = \mathrm{MSE}(g) \tag{9}$$

so the variance of the effective noise coincides with the minimum achievable mean squared error by a smooth data model based on the given selection of inputs.

To best model the time series data, we need to identify a function $\hat{g} \in S$ which is as close as possible to an optimal data model $g \in S_{\mathrm{opt}}$. Such a model will have close to minimal $\mathcal{E}\left((y - \hat{g}(\boldsymbol{x}))^2\right)$ and will not change significantly as more and more data is used in the model construction, i.e. as $M \to \infty$. We describe such a model as 'asymptotically stable'.

By (6) and (8),

$$\mathrm{MSE}(\hat{g}) = \mathcal{E}(R^2) + \mathcal{E}((g(\boldsymbol{x}) - \hat{g}(\boldsymbol{x}))^2). \tag{10}$$

Once the model selection process has been completed, it is tempting to assume that $g = f$, and hence that $\hat{g}$ is an approximation to the original function f that generated the noise-free data z_t. However, as observed in Kantz and Schreiber [6] and as we illustrate here, this is not necessarily the case. The main contribution of this note is to illustrate how these differences can be quantified using the Gamma test.

1.4. Model construction

In practice, given a noisy time series (y_t), we seek to construct an asymptotically stable model $\hat{g}$ for which the *empirical* mean squared error, defined by

$$\mathrm{MSE}_{\mathrm{emp}}(\hat{g}) = \frac{1}{M - d} \sum_{t=d+1}^{M} (y_t - \hat{g}(\boldsymbol{x}_t))^2 \tag{11}$$

is as close as possible to $\mathcal{E}(R^2)$, the variance of the effective noise. By (10), this ensures that $\hat{g}$ is as close as possible to an optimal predictive data model $g \in S_{\mathrm{opt}}$ (in a mean squared sense).

Our tool for estimating $\mathrm{var}(R)$ is the *Gamma test* [5], and we show how the results of a Gamma test analysis should be interpreted first when applied to input/output data with noisy inputs, and second when applied to noisy time series data.

[1] This type of noise is often called 'dynamic noise' in the literature on dynamic systems, see for example Casdagli et al. [1].

1.5. The Gamma test

The Gamma test is a fast, scalable algorithm for estimating the noise variance present in a data set modulo the best smooth model for the data, regardless of the fact that this model is unknown. A useful overview and general introduction to the method and its various applications is given in Jones [5].

In the standard Gamma test analysis we consider vector-input/scalar-output data sets of the form

$$\{(\boldsymbol{x}_i, y_i) \mid 1 \leq i \leq M\} \tag{12}$$

where the input vector $\boldsymbol{x}_i \in \mathbb{R}^d$ is confined to some closed bounded set $C \subset \mathbb{R}^d$. Under additive noise, the relationship between input and output is expressed by

$$y_i = f(\boldsymbol{x}_i) + r_i \tag{13}$$

where f is a smooth function with bounded gradient and r is noise with expectation zero. Despite the fact that f is unknown, the Gamma test computes an estimate for the noise variance $\mathrm{var}(r)$ directly from the data set (12).

The Gamma test estimates $\mathrm{var}(r)$ in $\mathrm{O}(M \log M)$ time by first constructing a kd-tree using the input vectors $\boldsymbol{x_i}$ $(1 \leq i \leq M)$ and then using the kd-tree to construct lists of the kth $(1 \leq k \leq p)$ nearest neighbours $\boldsymbol{x}_{N[i,k]}$ $(1 \leq i \leq M)$ of $\boldsymbol{x}_i$. Here p is fixed and bounded, typically $p \approx 10$. The algorithm next computes

$$\delta_M(k) = \frac{1}{M} \sum_{i=1}^{M} |\boldsymbol{x}_{N[i,k]} - \boldsymbol{x}_i|^2 \tag{14}$$

where $|.|$ denotes Euclidean distance,[2] and

$$\gamma_M(k) = \frac{1}{2M} \sum_{i=1}^{M} (y_{N[i,k]} - y_i)^2 \tag{15}$$

where $y_{N[i,k]}$ is the output value associated with $\boldsymbol{x}_{N[i,k]}$ (note that $y_{N[i,k]}$ is not necessarily the kth nearest neighbour of y_i in output space). Finally the regression line $\gamma_M(k) = \Gamma + A\delta_M(k)$ of the points $(\delta_M(k), \gamma_M(k))$ $(1 \leq k \leq p)$ is computed and the vertical intercept Γ returned as the estimate for $\mathrm{var}(r)$. The slope parameter A is also returned as this normally contains useful information regarding the complexity of the unknown surface $y = f(\boldsymbol{x})$.

The main result of Evans and Jones [2] is that, subject to certain conditions, if C is a compact convex body in $\mathbb{R}^m$ and data samples $\boldsymbol{x} \in C$ are selected with a smooth positive sampling density ϕ, then the number Γ returned by the algorithm converges in probability to Var(r) as $M \to \infty$.[3]

In the first instance the algorithm relies on the assumption that measurement error is confined only to the output, i.e. the inputs are assumed to be *noise-free*. If noise is confined only to the output, an approximately optimal smooth model $\hat{g}$ (in the sense described above) *will* also be a close approximation to f.

However, if the inputs are also corrupted by noise, as is the case for noisy time series (represented as input/output systems using delay vectors), we shall see that the effective noise is increased and that *it is no longer the case* that an approximately optimal $\hat{g}$ need approximate f.

Remark. In Evans and Jones [2] it is shown that the estimate computed by the Gamma test is (weakly) consistent, in the sense that it converges (in probability) to the true noise variance as the number of points $M \to \infty$. In particular, the input points $\boldsymbol{x}_i$ are assumed to be noise free, and confined to a closed and bounded subset C of $\mathbb{R}^d$. In the case of noisy inputs, then depending on the distribution of the input noise, the points $\boldsymbol{x}_i$ may be unbounded. In practice, provided that the tails of the component noise distributions approach zero sufficiently quickly, it appears that large values of $\boldsymbol{x}$ are so improbable as to not affect the conclusions of Evans and Jones [2].

Furthermore, that the Gamma test can be applied to time series embeddings follows because $z_t = f(z_{t-1}, \ldots, z_{t-d})$ where f is a smooth function. However, the theoretical analysis of the method presented in Evans and Jones [2] does not extend to the case where the dynamics are *chaotic*, although in practice, it has been demonstrated that the algorithm works well under these conditions.[4]

In Section 2.1 we first discuss the case of input/output data with noisy inputs, and illustrate this by a salutary example. In Section 2.2 we go on to apply the results to the case of noisy time series, and in Section 3 we give an example using time series data from the Hénon map with normally distributed additive noise.

2. Estimates of effective noise

The notion of effective noise introduced in Section 1.2 is not the same as residual error measured against a particular model. Effective noise essentially arises as the residual error against a best possible smooth model and can occur for a variety of reasons (see Evans and Jones [5]). Residual errors against a particular model can occur simply because the model is poorly chosen. However, even with the best possible model one cannot regularly produce an error variance lower than the effective noise variance imposed by the constraints of the overall situation, e.g. measurement error, incorrectly chosen embedding dimension in Eq. (8) etc. As remarked in Jones [5], even with no measurement error whatsoever, effective noise may still be present, for example as a consequence of choosing non-optimal input variables for the model.[5]

2.1. Input/output data with noisy inputs

For brevity, we suppress the dependence on the index t, and consider a 'typical' data point $(\boldsymbol{x}, y)$. If the input components

[2] Other metrics can be used.

[3] It seems likely that this result might be strengthened so that the convergence is 'almost surely'.

[4] Further information, references, and various software implementations of the Gamma test are available from the Gamma archive at http://users.cs.cf.ac.uk/Antonia.J.Jones/.

[5] In such cases the effective noise distribution is partly determined by the input distribution.

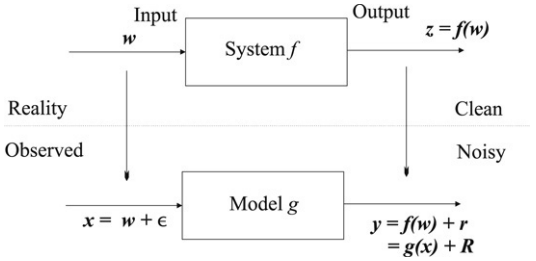

Fig. 1. Reality versus observation.

$x_1, \ldots, x_d$ of $\boldsymbol{x}$ are themselves subject to noise, this will increase the effective noise variance on the output. The Gamma test can still be used to estimate the variance of this effective noise (modulo the predictively optimal smooth model).

As illustrated in Fig. 1, we define

$$\boldsymbol{x} = \boldsymbol{w} + \boldsymbol{\epsilon} \quad \text{and} \quad y = f(\boldsymbol{w}) + r \tag{16}$$

where $\boldsymbol{w}$ and $f(\boldsymbol{w})$ represent the noise-free input and output, while $\boldsymbol{\epsilon}$ and r represent the input noise and output noise, respectively.

We assume that each input noise component ϵ_j has mean zero and bounded variance $\text{var}(\epsilon_j)$, and is independent of every other noise component ϵ_k $(j \neq k)$ and of r. In particular, because the ϵ_j are zero mean, this implies that $\mathcal{E}(\epsilon_j \epsilon_k) = 0$ whenever $j \neq k$.

Proposition 1. *Provided the input noise $\boldsymbol{\epsilon}$ is 'small' in some sense, the effective noise variance* $\text{var}(R)$ *satisfies*

$$\text{var}(r) \leq \text{var}(R) \leq \text{var}(r) + \mathcal{E}(|\nabla f|^2) \max_j \{\text{var}(\epsilon_j)\} \tag{17}$$

where $\mathcal{E}(|\nabla f|^2)$ is taken with respect to the $\boldsymbol{w}$ input sampling distribution, $\text{var}(\epsilon_j)$ *is the variance of the noise distribution on the jth input coordinate, and* $\text{var}(r)$ *is the variance of the output noise distribution.*

Remark. Here we seek to explain the practical consequences of our observations, rather than supply a formal mathematical analysis, so we have intentionally been somewhat cavalier in the statement. What is interesting is that these conclusions are largely independent of the exact nature of the distributions involved.

Proof. As we have seen, the variance of the effective noise is the minimum mean squared error achievable by any asymptotically stable smooth data model. In particular, because the underlying function f is smooth,

$$\mathcal{E}(R^2) \leq \text{MSE}(f) = \mathcal{E}((y - f(\boldsymbol{x}))^2). \tag{18}$$

Furthermore, $y = f(\boldsymbol{w}) + r$ where $\boldsymbol{w}$ is the clean input and r is the output noise. Hence by Taylor's theorem, for small $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_d)$,

$$\begin{aligned} y - f(\boldsymbol{x}) &= r + f(\boldsymbol{w}) - f(\boldsymbol{x}) = r + f(\boldsymbol{w}) - f(\boldsymbol{w} + \boldsymbol{\epsilon}) \\ &\approx r - (\nabla f(\boldsymbol{w})).\boldsymbol{\epsilon}. \end{aligned}$$

Squaring and taking expected values, because the output noise r is independent of both the noise-free input $\boldsymbol{w}$ and the input noise $\boldsymbol{\epsilon}$, and because $\mathcal{E}(r) = 0$,

$$\text{MSE}(f) \approx \mathcal{E}(r^2) + \mathcal{E}((\nabla f(\boldsymbol{w}).\boldsymbol{\epsilon})^2). \tag{19}$$

If we assume that the input noise $\boldsymbol{\epsilon}$ is independent of $\nabla f(\boldsymbol{w})$, for the $\boldsymbol{w}$ on which it is measured, and writing

$$\nabla f = (f_{x_1}, \ldots, f_{x_d}) = \left(\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_d} \right)$$

then

$$\begin{aligned} \mathcal{E}((\nabla f.\boldsymbol{\epsilon})^2) &= \mathcal{E}\left(\left(\sum_{j=1}^{d} f_{x_j} \epsilon_j \right)^2 \right) \\ &= \sum_{j=1}^{d} \mathcal{E}(f_{x_j}^2 \epsilon_j^2) + \sum_{j \neq k} \mathcal{E}(f_{x_j} \epsilon_j f_{x_k} \epsilon_k) \\ &= \sum_{j=1}^{d} \mathcal{E}(f_{x_j}^2) \mathcal{E}(\epsilon_j^2) + \sum_{j \neq k} \mathcal{E}(f_{x_j} f_{x_k}) \mathcal{E}(\epsilon_j \epsilon_k) \end{aligned}$$

which, since $\mathcal{E}(\epsilon_j \epsilon_k) = 0$ for all $j \neq k$, gives

$$\mathcal{E}((\nabla f.\boldsymbol{\epsilon})^2) = \sum_{j=1}^{d} \mathcal{E}(f_{x_j}^2) \mathcal{E}(\epsilon_j^2) \leq \mathcal{E}(|\nabla f|^2) \max_j \{\mathcal{E}(\epsilon_j^2)\} \tag{20}$$

i.e.

$$\mathcal{E}((\nabla f.\boldsymbol{\epsilon})^2) \leq \mathcal{E}(|\nabla f|^2) \max_j \{\text{var}(\epsilon_j)\}. \tag{21}$$

Finally, since $\text{var}(R) \leq \text{MSE}(f)$, we conclude from (18), (19) and (21) that

$$\text{var}(r) \leq \text{var}(R) \leq \text{var}(r) + \mathcal{E}(|\nabla f|^2) \max_j \{\text{var}(\epsilon_j)\}. \quad \square$$

As an upper bound for $\text{var}(R)$, this is a worst case analysis in that f itself certainly provides one *particular* smooth data model, although it may not be the function that best predictively models the data. The effective noise variance (corresponding to the best data model g) is therefore bounded above by $\text{MSE}(f)$, which in turn is bounded above by $\text{var}(r) + \mathcal{E}(|\nabla f|^2) \max_j \{\text{var}(\epsilon_j)\}$ for small ϵ_j. In particular, note that $\text{var}(R) \to \text{var}(r)$ as $(\max_j \{\text{var}(\epsilon_j)\}) \to 0$, i.e. as the noise on the inputs vanishes the effective noise variance approaches the variance of the noise on the output.

Example 1. Consider the function $f : \mathbb{R} \to \mathbb{R}$ defined by $f(w) = w^2$. We generate $M = 1000$ clean samples (w, z), where w is uniformly distributed in $[1, 2]$ and $z = f(w) = w^2$. Then we add normally distributed[6] noise ϵ, having mean zero and variance 0.075, to the clean inputs w, yielding the noisy inputs $x = w + \epsilon$. In this example, we set the output noise to zero, so that $y = w^2$. Fig. 2 shows the spread of data and the $f(w) = w^2$ curve.

Next we run the Gamma test on the resulting input/output pairs (x, y) and obtain $\Gamma = 0.336593$ as an estimate for

[6] We have added normally distributed noise as this is typical of measurement error, but the results herein are (subject to some reasonable conditions) largely independent of the precise nature of the noise distribution.

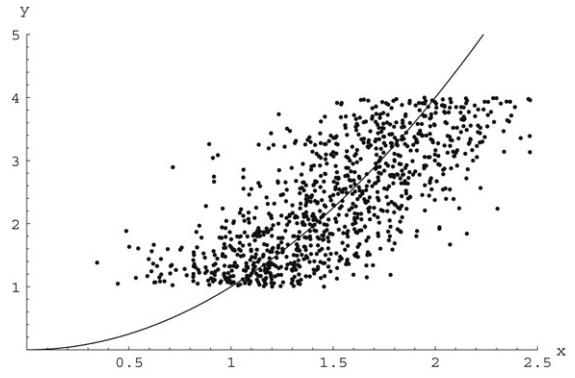

Fig. 2. The data set having noisy inputs (x, y) and the function $y = w^2$.

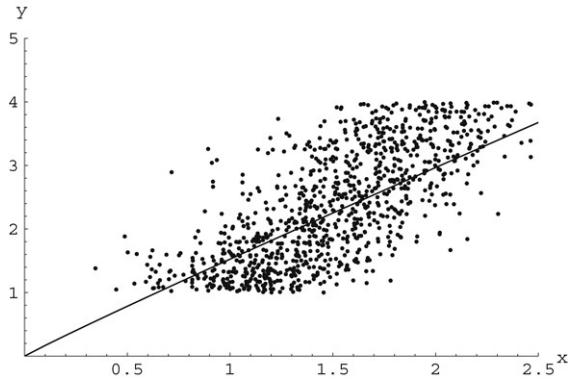

Fig. 3. The data set of noisy inputs and the 'best-fit' function.

the effective noise variance $\mathrm{var}(R)$, which indicates the best possible modelling error achievable by a smooth data model. For comparison, we also compute

$$\mathrm{MSE}(f) \approx \frac{1}{M}\sum_{i=1}^{M}(y_i - f(x_i))^2 = 0.713487. \tag{22}$$

This is the residual mean squared deviation of the data with respect to the original function $f(w) = w^2$ used to generate the data, and is more than twice the Gamma statistic. By comparison, direct numerical computation[7] of $\mathcal{E}(|\nabla f|^2)\mathrm{var}(\epsilon)$, which (since $\mathrm{var}(r) = 0$) is the upper bound given by Proposition 1, yields 0.7. From (22) this is in close agreement with $\mathrm{MSE}(f)$. This demonstrates that Γ may be significantly less than $\mathrm{var}(r) + \mathcal{E}(|\nabla f|^2)\mathrm{var}(\epsilon)$.

The difference between the estimate $\Gamma = 0.336593$ and $\mathrm{MSE}(f) = 0.713487$ prompts us to ask whether we can find a function g that better fits the data?

To this end, we transform the data to $(\log(x), \log(y))$ and perform a least-squares linear fit on the resulting data set, from which we obtain the relation $\log y = 0.959745 \log x + 0.422462$. Exponentiating the data in the form $y = \alpha x^{\beta}$ then leads to the model $\hat{g}(x) = 1.52501x^{0.959745}$. Fig. 3 shows the spread of data and the $\hat{g}(x)$ curve.

[7] In practice, we do not know this value because the underlying function f is unknown. In fact, because $f(w) = w^2$ and the (clean) inputs w are uniformly distributed in $[1, 2]$, it is easily shown that $\mathcal{E}(|\nabla f|^2) = 28/3 \approx 9.33$ and hence, because $\mathrm{var}(r) = 0$ and $\mathrm{var}(\epsilon) = 0.075$, that the upper bound of Proposition 1 is equal to 0.7.

Table 1
Summary of Example 1 results ($M = 1000$)

Function	g	$\hat{g}$	f
Noise	$\mathrm{var}(R)$	$\mathrm{MSE}(\hat{g})$	$\mathcal{E}(\lvert\nabla f\rvert^2)\mathrm{var}(\epsilon) = 0.7$
Experiment	$\Gamma = 0.33659$	0.35216	$\mathrm{MSE}\,(f) = 0.71348$

Adding more data does not essentially change this best-fit model, so the model is asymptotically stable. For example, repeating the exercise for 10 000 data points, we obtain $y = 1.51946x^{0.958342}$, so the values 1.52 and 0.959 were essentially determined by the input noise (and of course the original function f).

Finally we compute the residual mean squared deviation from the model $\hat{g}$:

$$\mathrm{MSE}(\hat{g}) = \frac{1}{M}\sum_{i=1}^{M}(y_i - 1.52501x_i^{0.959745})^2 \approx 0.352164.$$

This is close to the 'best possible' modelling error estimated by $\Gamma = 0.336593$, which indicates that $\hat{g}$ is a close approximation to the optimal data model g. Note that the precise method used to construct $\hat{g}$ is not critical. We summarize the numerical results in Table 1.

One can use any suitable construction that produces a smooth model having a residual error variance close to the Γ value, although Occam's razor suggests that, given several models with similar error variance, one should strive to choose the *simplest*.

2.2. Noisy time series

The clean (noise-free) time series is represented by the sequence $z_1, \ldots, z_M$. As in (1), we assume that there is some smooth function f and a number d such that

$$z_t = f(z_{t-1}, \ldots, z_{t-d}). \tag{23}$$

The underlying clean time series is corrupted by independent and identically distributed additive noise $r_1, \ldots, r_M$, so we observe the noisy time series $y_1, \ldots, y_M$,

$$y_t = z_t + r_t. \tag{24}$$

For $t = d + 1, \ldots, M$ we construct a set of (noisy) input points $\boldsymbol{x}_t \in \mathbb{R}^d$ using *delay vectors*:

$$\boldsymbol{x}_t = \begin{pmatrix} y_{t-1} \\ \vdots \\ y_{t-d} \end{pmatrix} = \begin{pmatrix} z_{t-1} \\ \vdots \\ z_{t-d} \end{pmatrix} + \begin{pmatrix} r_{t-1} \\ \vdots \\ r_{t-d} \end{pmatrix}. \tag{25}$$

We write $\boldsymbol{x}_t = \boldsymbol{w}_t + \boldsymbol{\epsilon}_t$ where

$$\boldsymbol{w}_t = \begin{pmatrix} z_{t-1} \\ \vdots \\ z_{t-d} \end{pmatrix} \quad \text{and} \quad \boldsymbol{\epsilon}_t = \begin{pmatrix} r_{t-1} \\ \vdots \\ r_{t-d} \end{pmatrix} \tag{26}$$

represent the clean input and the input noise respectively.

Applying Proposition 1, because $\mathrm{var}(\epsilon_j) = \mathrm{var}(r)$ for all j we obtain the following:

Corollary 1. *For noisy time series, provided the noise r is 'small' in some sense,*

$$\mathrm{var}(r) \leq \mathrm{var}(R) \leq (1 + \mathcal{E}(|\nabla f|^2))\mathrm{var}(r) \tag{27}$$

where the expectation $\mathcal{E}(|\nabla f|^2)$ is taken with respect to the distribution of the noise-free delay vectors $\boldsymbol{w}$.

2.3. Noise amplification

In the context of noisy non-linear time series what we have called 'effective noise' arises from a combination of observational noise and noise amplification. Noise amplification arises where measurement errors in one dynamic variable feed through the dynamics to produce larger errors in predictive models for some other variable.

In our case we are not considering state space reconstruction *per se*, but merely an 'autoregressive' model of a single noisy time series (y_t) for sole purpose of prediction (although the spirit of our observations extents to the case of using time delay vectors from several variables of the system as inputs to the predictive model). However, it is interesting to note in Corollary 1 that, even in this case, the nature of the particular zero-noise (i.e. ideal) smooth embedding function $z_t = f(z_{t-1}, z_{t-2}, \ldots, z_{t-d})$ of a given dimension d, determines the noise amplification of r_t on z_t. It appears that it does this via a factor proportional to $\mathcal{E}(|\nabla f|^2)$, at least in the low noise case. Of course $\mathcal{E}(|\nabla f|^2)$ is itself determined by the underlying dynamical system. The upper bound for the effective noise variance $\mathrm{var}(R)$, (estimated by the Gamma test) given by Corollary 1 suggests that the combined effect of noise amplification and observational noise can be neatly described in terms of $\mathcal{E}(|\nabla f|^2)$.

3. The Hénon time series

The Hénon time series is generated iteratively using the equation

$$z_t = f(z_{t-1}, z_{t-2}) = -z_{t-1}^2 + bz_{t-2} + a \tag{28}$$

where $z_0 = 0$, $z_1 = 0$, $a = 1.4$ and $b = 0.3$. The points (z_{t-1}, z_{t-2}) of the map ergodically sample the attractor of the system, which is a set of zero measure but positive Hausdorff dimension. This can be extracted from the time series data and visualized by simply plotting the inputs (z_{t-1}, z_{t-2}) against the corresponding output z_t, as shown in Fig. 4.

Example 2. We generate $M = 1000$ points z_t of the Hénon map. The first 100 points are plotted in Fig. 5. Next we add normally distributed noise r_t, having mean zero and variance $\mathrm{var}(r) = 0.075$, to each z_t, which yields a noisy time series $y_t = z_t + r_t$. Fig. 6 illustrates the noise component.

Next we form a set of 2-input/1-output data points $(\boldsymbol{x}_t, y_t)$ where $\boldsymbol{x}_t = (y_{t-1}, y_{t-2})$ and $2 \leq t \leq M$, from which we compute an estimate $\Gamma = 0.27713$ for the effective noise variance $\mathrm{var}(R)$. This is an estimate for the best mean squared error we are likely to achieve for a smooth predictive

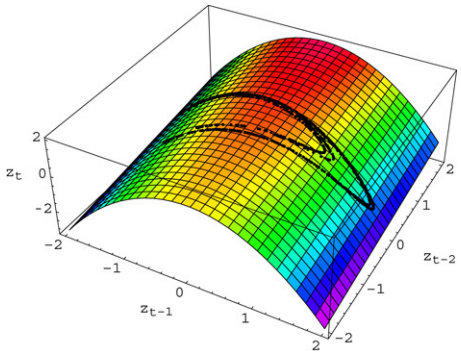

Fig. 4. The Hénon map attractor (black dots) draped over a topographic rendering of the surface $f(u, v) = -u^2 + bv + a$, where u corresponds to z_{t-1} and v to z_{t-2}.

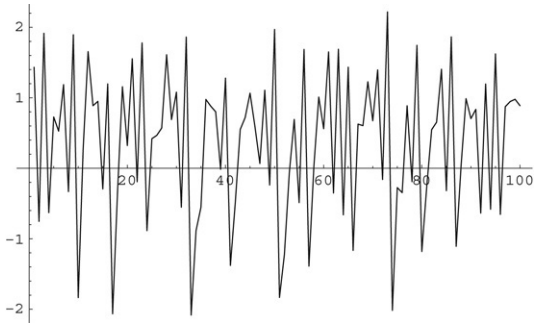

Fig. 5. First 100 points of the clean Hénon time series z_t.

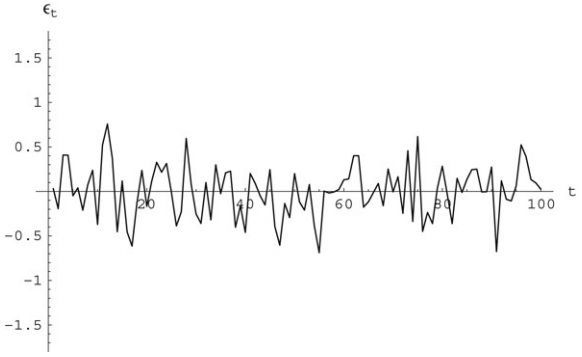

Fig. 6. First 100 points of the noise $r_t = \epsilon_t$.

model using the noisy data, based on the *assumed*[8] embedding dimension $d = 2$. For 10 000 data points, the Gamma statistic becomes 0.26702, so does not change significantly.

Estimating the upper bound given by Corollary 1, we obtain $(1+4.68073) \times 0.075 = 0.426055$, i.e. about twice the Gamma

[8] Given the recursive definition of the time series in (28), at first glance an embedding dimension of $d = 2$ seems reasonable. However, there are many recurrence relations which derive from (28) and can be used to generate the time series (for example in the RHS of (28) substitute $z_{t-2}^2 + bz_{t-3} + a$ for z_{t-1} etc.). Moreover, the effect of introducing noise, and the fact that the best modelling function g actually *depends* on the noise, means we have almost no *a priori* knowledge of the optimal embedding dimension for modelling purposes.

Table 2
Summary of Example 2 results ($M = 1000$)

Function	g	$\hat{g}$	f
Noise	var(R)	MSE($\hat{g}$)	$(1 + \mathcal{E}(\lvert\nabla f\rvert^2))$var($r$)
Experiment	$\Gamma = 0.27713$	0.30662	MSE(f) ≈ 0.42605

Here MSE($\hat{g}$) for the neural network is the weighted average of training and test data.

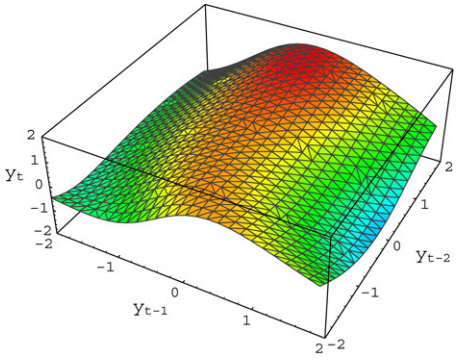

Fig. 7. The topographic neural network surface $\hat{g}$, produced by training on 600 points of the noisy data $(\boldsymbol{x}_t, y_t)$, where $\boldsymbol{x}_t = (y_{t-1}, y_{t-2})$.

statistic. We summarize the numerical results in Table 2. Thus we are in a situation analogous to that of Example 1: the Gamma statistic is suggesting that there is a function g, different from the surface f, which, using the noisy data as inputs, should produce a better prediction model for the noisy output.

Finally we constructed an approximation $\hat{g}$ to the surface g using a $2 \to 5 \to 5 \to 1$ feed-forward neural network, trained on the first 600 data points, using the BFGS algorithm [3]. The network was trained to a mean squared error of 0.264986, which was the Gamma statistic computed using these training points. The mean squared error on the remaining 400 points was 0.32407, somewhat worse than the Gamma statistic.

The trained network surface $\hat{g}$ is shown in Fig. 7. At first sight we might suppose this to be a crude approximation to the surface f of Fig. 4, but using considerably more data (e.g. 5000 points) to train the network does not significantly change the surface, i.e. the model is asymptotically stable. The essential fact is that the g approximated by $\hat{g}$ is a *different* surface from the f of (28).

4. Prediction error and reconstruction

In Casdagli [1] the prediction error is described as arising from a combination of 'noise amplification' and 'estimation error'. Here the estimation error arises from the fact that for a finite data set our reconstruction will always only be approximate. Noise amplification depends on the reconstruction and the variable being measured (often called the 'measurement function'). If for the moment we ignore estimation error, then prediction error is determined by noise amplification of the measurement function and on the reconstruction. The equality in (20) leads to the bound for effective noise variance on the RHS in (27), and this appears to be precisely a special case of Eq. (32) in Casdagli et al. [1], where in our case the measurement function is $y_t = z_t + r_t$. However, what we seek to emphasize here is that, in face of small noise on z_t, the RHS of (27) is only an *upper bound* on the prediction error, i.e. the effective noise variance might well be lower, and we can then do better in predicting y_t than the upper bound suggests.

For an embedding dimension m and a lag, or delay, time τ (the time between successive coordinates of the delay vector) we define the time window $t_w = (m - 1)\tau$. Considerable experimental evidence suggests that for optimal reconstruction it may be more appropriate to optimize t_w rather than m or τ alone. For example, Martinerie et al. [7] showed that the correlation integral is sensitive to t_w, but not to m and τ individually.

In practice it is very useful to have a method for simultaneously optimizing both the embedding dimension m and the delay time τ, and a variety of techniques for doing this have been suggested in numerous papers, e.g. see [9]. One approach, of interest in the context of the present paper, is to be found in Hong-Guang and Chong-Zhao [4]. Here, using an improvement of an idea originally suggested in Otani and Jones [8], a combination of the Gamma test and multiple autocorrelation are employed to produce an efficient, and apparently very effective, algorithm for estimating near optimal parameters for embedding dimension and delay time.

5. Conclusions

We have illustrated that a Gamma test analysis on embeddings of noisy time series data

$$y_t = z_t + r_t$$

does not return an estimate for var(r), but rather returns an estimate for the *effective* noise variance var(R), where R is defined by (8) and g is an unknown but optimal smooth data model. Moreover, g may or may not be an approximation to the original smooth function f which generated the underlying clean time series.

Thus our first example highlights the danger of inferring a *process law* using a *model* constructed from noisy data. The process was actually quadratic, but one might easily be tempted to infer that it was linear from a near optimal predictive model.

It is an unfortunate fact of life that an optimal modelling function g may not bear a close relationship to f [6]. However, this does not reduce the efficiency of a $\hat{g} \approx g$ in producing the best possible predictions using unseen noisy data drawn from the same process.

For time series we have seen that the estimate for the effective noise variance var(R) may be much larger than the variance of the noise on individual measurements var(r). It may seem frustrating that we have no direct access to var(r). However, for the purposes of model building, knowing var(r) is not so critical. Rather, we need to know var(R), e.g. so as to

know when to stop training a neural network, and this is exactly what the Gamma test estimates.

What is of concern is that even if the measurement error variance on the original time series is quite modest, this can lead to a much *larger* error variance var(R) for the predictions, because the fluctuations in the input can be magnified by the transformation f. Thus another moral of the story would seem to be that, when using such predictive techniques, it is essential to reduce the measurement error on time series data as much as possible.

In practice, var(R) is likely to increase with $\mathcal{E}(|\nabla f|^2)$, which is itself *a priori* unknown, and will certainly be different for different embeddings. Nevertheless we remark that it is perfectly valid to use the Gamma statistic Γ to estimate var(R) for different embeddings, and to select, based on a 'minimum Γ' criterion, an appropriate embedding dimension and a suitable irregular embedding for a noisy time series model.

Acknowledgements

Wayne Haythorn, of Granite Software, first raised this issue, and we are indebted to Dr. Tom Westerdale of Birkbeck College, University of London, for his insightful comments.

References

[1] M. Casdagli, S. Eubank, J.D. Farmer, J. Gibson, State space reconstruction in the presence of noise, Physica D 51 (1991) 52–98.
[2] D. Evans, A.J. Jones, A proof of the gamma test, Proc. Roy. Soc. Ser. A (ISSN: 1364-5021) 458 (2027) (2002) 2759–2799.
[3] R. Fletcher, Practical Methods of Optimization, 2nd ed., John Wiley & Sons, 1987.
[4] M. Hong-Guang, H. Chong-Zhao, Selection of embedding dimension and delay time in phase space reconstruction, Electr. Electron. Eng. China 1 (2006) 111–114.
[5] A.J. Jones, New tools in non-linear modelling and prediction, Comput. Manag. Sci. 1 (2004) 109–149.
[6] H. Kantz, T. Schreiber, Nonlinear Time Series Analysis, 2nd ed., Cambridge Univ. Press, 2004.
[7] J.M. Martinerie, A.M. Albano, A.I. Mees, P.E. Rapp, Mutual information, strange attractors, and the optimal estimation of dimension, Phys. Rev. A 45 (1992) 7058–7064.
[8] M. Otani, A.J. Jones, Automated embedding and creep phenomenon in chaotic time series. *EB/OL*, http://users.cs.cf.ac.uk/Antonia.J.Jones/UnpublishedPapers/Creep.pdf, 2000.
[9] M. Rosenstein, J. Collins, C.D. Luca, Reconstruction expansion as a geometry-based framework for choosing proper delay times, Physica D 73 (1994) 82–98.
[10] F. Takens, Detecting strange attractors in turbulence, in: D. Rand, L. Young (Eds.), Dynamical Systems and Turbulence, in: Lecture Notes in Mathematics, vol. 898, Springer-Verlag, 1981, pp. 366–381.