


---

**Monograph:**
Aguirre, L.A. and Billings, S.A. (1993) Identification Models for Chaotic Systems from a Noisy Data: Implications for Performance and Nonlinear Filtering. Research Report. ACSE Research Report 485 . Department of Automatic Control and Systems Engineering

---



# Identification of Models for Chaotic Systems from Noisy Data: Implications for Performance and Nonlinear Filtering

L A Aguirre and S A Billings

Department of Automatic Control and Systems Engineering
University of Sheffield
P.O. Box 600
Mappin Street
Sheffield S1 4DU
United Kingdom

Research Report No 485

September 1993

# Identification of Models for Chaotic Systems from Noisy Data: Implications for Performance and Nonlinear Filtering

LUIS A. AGUIRRE† and S. A. BILLINGS

Department of Automatic Control and Systems Engineering
University of Sheffield
P.O. Box 600, Mappin Street — Sheffield S1 4DU - UK

**Abstract**

This paper investigates the identification of global models from chaotic data corrupted by purely additive noise. It is verified that noise has a strong influence on the identification of chaotic systems. In particular, there seems to be a critical noise level beyond which the accurate estimation of polynomial models from chaotic data becomes very difficult. Similarities with the estimation of the largest Lyapunov exponent from noisy data suggest that part of the problem might be related to the limited ability of predicting the data records when these are chaotic. A nonlinear filtering scheme is suggested in order to reduce the noise in the data and thereby enable the estimation of good models. This prediction-based filtering incorporates a resetting mechanism which enables filtering chaotic data. Numerical examples which consider the double-scroll attractor and the Duffing-Ueda oscillator are provided to illustrate the main points of the paper.

## 1 Introduction

When the nonlinearities in a data set cannot be neglected it is necessary to use nonlinear representations in the modelling of the underlying dynamics. Pioneering techniques for the identification of nonlinear systems were based on Volterra and Wiener functional expansions. Such models were, in principle, able to reproduce typical nonlinear phenomena such as bifurcations, limit-cycles, quasiperiodic motions and chaos. The great difficulty with such methods however was the prohibitively large number of parameters needed to model even simple nonlinear functions [1].

The use of nonlinear representations which are linear-in-the-parameters was proposed as an alternative solution to the identification of nonlinear systems [2, 3]. However, if every term of these functions were considered, these models would become impractically large for relatively small degrees of nonlinearity and number of degrees of freedom. This demanded techniques which would select the most relevant terms among a large (typically thousands) number of candidate terms and is one of the great challenges in nonlinear identification [4, 5].

†e-mail:aguirre@acse.sheffield.ac.uk

It is now recognised that the structure of a model influences the dynamics and also that certain model structures are more appropriate for modelling specific systems than others [6]. For instance, it is known that polynomial models are not adequate for modelling systems with outputs which vary rapidly. Furthermore, it is widely believed that polynomials present restrictions even when the output is a smooth function because the number of terms composing the model grows exponentially and consequently the resulting model does not extrapolate accurately beyond the domain of validity and may even become unstable [7, 8, 9].

Another difficulty which must be faced in most practical applications is the presence of noise. The effects of noise on estimated models is highly system-dependent in nonlinear systems. Thus model estimation from noisy data is more robust in some cases than in others. In particular, it seems that the effects of noise on chaotic systems is somewhat greater than on non-chaotic systems [10].

This paper is concerned with the estimation of global polynomials from chaotic data corrupted by noise. In order to significantly reduce the difficulties which are usually attributed to polynomial models due to an excessively large number of terms, an effective algorithm is used to select the structure of the model. Thus the estimated polynomial models are composed of a few terms (typically less than 20) and consequently, in several examples such models perform very well.

It has been observed that the presence of noise contaminating chaotic data is a major obstacle to be overcome in the identification of chaotic systems. Curiously, the same noise levels in applications concerned with nonchaotic systems seem not to have such a devastating effect. This paper investigates the effects of noise on the identification of polynomial models for chaotic systems and possible causes are conjectured.

In order to verify the validity of some of the conjectures made, a nonlinear filtering procedure is investigated. Unlike most of the existing filtering techniques, the main objective of filtering the data in this paper is to enable the identification of polynomial models from the filtered data and not necessarily to recover the noise-free orbit.

To illustrate that in some cases filtering the data makes the estimation of good models possible, some examples are presented which use the equations of Chua's circuit operating in the double-scroll attractor regime and the Duffing-Ueda oscillator operating over a wide range of parameter values. Such results seem to confer strength to some of the conjectures made.

Because the main aim of the present study is to demonstrate the importance of determining the model structure and, in this context, the usefulness of new prediction based filtering methods for noise reduction, the results have been implemented based on polynomial model expansions. The chief aim is to demonstrate the value of the new algorithms in the simplest and most transparent way and this is best achieved using the polynomial model. It is important to emphasise however that the results can readily be extended to other more complex model forms.

The paper is organised as follows. In § 2 the main techniques used in the identification of global polynomial models is reviewed. Section 3 describes the two systems used in the numerical examples, namely Chua's circuit and the Duffing-Ueda oscillator. Section 4 describes the use of prediction-based techniques in the identification of the systems described in § 3 when the data are corrupted with purely additive white noise. The effects of the noise in the estimation of the largest Lyapunov exponent are verified and certain similarities with the estimation of polynomial models are highlighted. This comparison motivates some conjec-

or

$$y(t) = G^{yu}[\cdot] \; + \; G^{yu\xi}[\cdot] \; + \; G^{\xi}[\cdot] \; + \; \xi(t) \tag{7}$$

where $\xi(t)$ is the residual at time $t$ and is defined as

$$\xi(t) \doteq y(t) - \hat{y}(t) \tag{8}$$

and

$$\hat{y}(t) = \Psi_{yu}^{T}(t-1)\hat{\Theta}_{yu} \; + \; \Psi_{yu\xi}^{T}(t-1)\hat{\Theta}_{yu\xi} \; + \; \Psi_{\xi}^{T}(t-1)\hat{\Theta}_{\xi} \tag{9}$$

is called the *one-step-ahead* (OSA) predictor and $\hat{y}(t)$ is the OSA prediction of $y(t)$. Finally, equation (6) can be expressed in concise form as

$$\begin{aligned} y(t) \; &= \; \left[\Psi_{yu}^{T}(t-1) \; \Psi_{yu\xi}^{T}(t-1) \; \Psi_{\xi}^{T}(t-1)\right] \begin{pmatrix} \hat{\Theta}_{yu} \\ \hat{\Theta}_{yu\xi} \\ \hat{\Theta}_{\xi} \end{pmatrix} \; + \; \xi(t) \\ y(t) \; &= \; \Psi^{T}(t-1)\hat{\Theta} \; + \; \xi(t) \end{aligned} \tag{10}$$

The following cost function can be defined based on the last equation

$$J_{LS}(\hat{\Theta}) \doteq \| \; y(t) - \Psi^{T}(t-1)\hat{\Theta} \; \| \tag{11}$$

where $\| \cdot \|$ is the Euclidean norm. A typical *least squares* (LS) problem is to find $\hat{\Theta}$ such that $J_{LS}(\hat{\Theta})$ is minimised [14].

A similar cost function can be defined as follows

$$J_{PE}(\hat{\Theta}) \doteq \log_e \det Q(\hat{\Theta}) \tag{12}$$

where $Q(\hat{\Theta})$ is the sample covariance matrix of the residuals and is defined as

$$Q(\hat{\Theta}) \doteq \frac{1}{N} \sum_{t=1}^{N} \xi(t)\xi^{T}(t) \tag{13}$$

One of the major difficulties in solving equation (10) is that such a set of equations is typically ill-conditioned, especially if the number of terms is large. To circumvent this problem orthogonal techniques may be used [4, 15].

Effective solutions to handle the problem of determining the structure of nonlinear models are available in the literature [4, 5]. One solution is based on the *error reduction ratio* (ERR) test which provides an indication of which terms to include in the mode. Two advantages of this approach are i) it does not require the estimation of a complete model to determine the significance of a candidate term and the respective statistical contribution to the output, and ii) the ERR test is derived as a by-product of the orthogonal estimation algorithm. For details see [4, 16].

Once a model is estimated it should be submitted to a number of tests which should check if the model is adequate and, hopefully, will also provide a measure of goodness for

where $A_i \in \mathbb{R}$ are the values of the control parameter $A$ for which the system bifurcates. The $a_i$'s are defined likewise and the summation is taken over all the $(N_b)$ bifurcation points of interest.

# 3 The original nonlinear systems

Chua's circuit is certainly one of the most well studied nonlinear circuits and a great number of papers ensure that the dynamics of this circuit are also well documented [23]. The normalised equations of Chua's circuit can be written as [24]

$$\begin{cases} \dot{x} = \alpha(y - h(x)) \\ \dot{y} = x - y + z \\ \dot{z} = -\beta y \end{cases} \tag{17}$$

where

$$h(x) = \begin{cases} m_1 x + (m_0 - m_1) & x \geq 1 \\ m_0 x & |x| \leq 1 \\ m_1 x - (m_0 - m_1) & x \leq -1 \end{cases} \tag{18}$$

In what follows $m_0 = -1/7$ and $m_1 = 2/7$. Varying the parameters $\alpha$ and $\beta$ the circuit displays several regular and chaotic regimes. The famous double scroll attractor, for instance, is obtained for $\alpha = 9$ and $\beta = 100/7$ and has the largest Lyapunov exponent and the Lyapunov dimension equal to $\lambda_1 = 0.23$ and $D_L = 2.13$, respectively [25]. These parameter values will be used henceforth.

Figure 1 shows a trajectory on the double scroll reconstructed using the $z$ component and plotting $z(t) \times z(t - T_p)$ with $T_p = 0.3$. The trajectories of the system were obtained by digital simulation using a fourth-order Runge-Kutta algorithm with an integration interval equal to $1 \times 10^{-3}$.

The Duffing-Ueda equation [26]

$$\ddot{y} + k\dot{y} + y^3 = u(t) \tag{19}$$

was originally proposed as a model for nonlinear oscillators and has become a bench test for the study of nonlinear dynamics. It has also been considered as a simple paradigm for chaotic dynamics in electrical science [27]. One of the main reasons for this is that in spite of being simple this model can produce a variety of dynamical regimes, from period-one motions to chaos [28, 26].

To obtain the bifurcation diagrams and Poincaré sections shown in this paper, the input was chosen to be of the form $u(t) = A\,cos(\omega t)$ where the maximum input amplitude $A$ was used as the control parameter. The bifurcation diagram shown in figure 2a was obtained by taking $k = 0.1$, $\omega = 1$ rad/s and simulating equation (19) digitally using a fourth-order Runge-Kutta algorithm with an integration interval equal to $\pi/3000$. Figure 2b shows the Poincaré section of the attractor at $A = 5.7$.

The systems used in the numerical examples were chosen mainly because of three factors, namely i) such systems are well documented in the literature, ii) the relative simplicity of the equations facilitate the presentation of numerical experiments and permits that the discussions remain focussed on the principal points of the paper, and iii) differential equations were preferred to discrete maps for generating the data records because in practice most systems are better represented as continuous processes.

# 4 Prediction-based estimation of chaotic maps

The chief objective of this section is to illustrate the kind of problems faced when dealing with chaotic data contaminated with additive noise. Two different problems are investigated, namely the identification of NARMAX polynomial models for chaotic systems and the estimation of the largest Lyapunov exponent. A common feature to these problems is that in both cases the algorithms are based on short-term predictions. This observation will be used to suggest a possible explanation for the estimation problems verified.

## 4.1 Estimation from noisy data — examples

Two examples are provided in this subsection. The first example considers the Chua system and the second example uses the Duffing-Ueda oscillator. In both instances white noise is added to data which were obtained by digital simulation of the differential equations governing the systems.

### 4.1.1 Example 1

In this example equations (17) and (18) were used to simulate Chua's circuit operating in the double scroll region. After transients had died out, 1750 data points sampled at $T_s = 0.15$ and corresponding to the $z$ component were recorded and white noise with variance $\sigma_e^2 = 0.021$ was added to the data which resulted in a *signal to noise ratio* (SNR) equal to $20\log(2.62/0.021) \approx 42\,\text{dB}$.

The techniques mentioned in § 2 were then used to identify NARMA polynomial models from the data. The parameters used were $\ell = 4$, $n_y = 5$, $n_u = 0$ and $n_e = 20$. The total number of candidate process terms was 125 and the most significant were chosen based on the ERR test. Several different models were estimated by varying the number of terms allowed in the model. A typical model follows

$$\begin{aligned}
z(k) &= 0.19429\times 10 z(k-1) + 0.13823 z(k-3)z(k-4)^2 - 0.61780 z(k-3) \\
&+ 0.10904 z(k-1)^2 z(k-4) - 0.20556 z(k-1)z(k-5)^2 \\
&- 0.73391\times 10^{-1} z(k-1)z(k-2)z(k-5) + 0.51675 z(k-5) \\
&- 0.12494 z(k-4)^2 z(k-5) - 0.23710 z(k-4) - 0.43345 z(k-2) - 0.050681 z(k-1)^3 \\
&+ 0.10732 z(k-1)z(k-2)z(k-3) - 0.36301 z(k-1)z(k-3)z(k-4) \\
&- 0.60481\times 10^{-1} z(k-5)^3 + 0.29419 z(k-3)z(k-5)^2 - 0.43793 z(k-3)^2 z(k-5) \\
&+ 0.69855\times 10^{-1} z(k-2)z(k-5)^2 + 0.51947 z(k-1)z(k-3)z(k-5) \\
&- 0.14679 z(k-1)^2 z(k-5) + 0.15630 z(k-3)^2 z(k-4) \\
&+ \Psi_\xi^{\mathrm{T}}(k-1)\hat{\Theta}_\xi + \xi(k) \qquad (20)
\end{aligned}$$

with $\sigma_\xi^2 = 0.042$. The attracting set for this model is shown in figure 3. Clearly, the identified model settles to a chaotic attractor which is very different from the original one shown in figure 1.

even for chaotic models since such models should give accurate one-step-ahead predictions for short sampling intervals[3].

The estimation algorithm is illustrated in figure 5. It should be noted that because the OSA prediction at time $t$ is obtained by taking measured data up to time $t-1$ and then predicting just one step ahead into the future, the OSA predictions are normally very close to the data and do not usually drift away even when the data/model are chaotic. In other words, the OSA predictor is reset at each iteration with measured data and this resetting action maintains the OSA predictions close to the original data. Because least squares and prediction error estimation algorithms use the OSA predictions to update parameter estimates this is a possible justification for the success of such parameter estimation techniques for some chaotic systems when the data are clean [6].

The resetting effect obtained when the model is used to compute the OSA predictions has no counterpart when a model is used to simulate the system many steps into the future. In the latter case, at each iteration the model is initialised with the data predicted in previous iterations. Thus if the model is sensitive to small variations in the initial conditions the long-term predictions will eventually drift away from the original data.

However, it should be noted (see figure 5) that only a part of the model is reset, namely the terms involving inputs and outputs. The rest of the model will be initialised from previously simulated data via the residuals. Thus the residuals are fedback into the estimation algorithm in future iterations. This observation motivates a closer inspection of the residuals especially in cases where adequate parameter estimation is precluded.

Figure 6 shows the variance of the residuals as a function of the number of noise iterations which are performed during parameter estimation of the Duffing-Ueda oscillator. For most systems it is expected that such a variance converges to a value which is relatively close to the noise variance. This figure reveals that the residual variance of the models obtained with five or less noise iterations is close to the noise variance, 0.015 (note that the model in equation (21) was obtained performing four noise iterations) but that the variance of the residuals grows monotonically with the number of noise iterations. This seems to indicate difficulties with parameter convergence.

## 4.3 Estimation of the largest Lyapunov exponent

The estimation of Lyapunov exponents is known to be a nontrivial task. The simplest algorithms [30, 27] can only reliably estimate the largest Lyapunov exponent, $\lambda_1$. Such algorithms are based on predictions of small perturbations along a an attacting set. Estimating the entire spectrum is a typically ill-conditioned problem and requires more sophisticated algorithms [30, 31].

Figures 7a and 7b show the estimated values of $\lambda_1$ for varying noise levels. Figure 7a corresponds to the double scroll displayed by Chua's circuit and figure 7b was obtained using the Duffing-Ueda equation equation with $A=11$ (marked with circles) and $A=4.5$ (marked with asterisks). In both figures the Lyapunov interval was made equal to the respective sampling interval used in the identification, that is $\Delta L=0.15$ and $\Delta L=\pi/60$, respectively.

[3]The sampling rate is usually chosen fast enough in order to guarantee that no relevant high frequency information is lost due to poor sampling. Therefore in the context of system identification the sampling interval can usually be considered 'short' when compared to the prediction capability of the model.

A common feature to these graphs is that, in estimating $\lambda_1$, there seems to be a limited tolerance to noise. Thus after a certain critical value it becomes increasingly difficult to estimate this exponent accurately. Moreover, this critical point for $A=4.5$ (which corresponds to a non-chaotic regime) is somewhat higher than its counterpart at $A=11$.

Concerning the curve in figure 7b marked with asterisks, two interesting points are worth mentioning. First, even for high noise levels good estimates were obtained. Second, with only one exception, whenever the estimated value of $\lambda_1$ was not considered accurate such a value was positive.

These figures suggest that, at least for some systems, estimating $\lambda_1$ would be more difficult when this exponent is positive than if it were negative. This would point to the limited capability of making accurate predictions when the data are too noisy as one possible reason for some of the difficulties encountered and this seems to account, at least partially, for the sharp decline in estimation accuracy when the uncertainty (noise) associated with the data exceeds a certain value. Another possible source of errors is that as the noise level is increased the Jacobian will tend to be evaluated (more and more frequently) at points which are more distant from the noise-free trajectory. Because the Jacobian is defined in a small neighbourhood of the noise-free trajectory, increasing the noise level will weaken the validity of the evaluated Jacobian which is actually used in the calculations [12].

## 4.4 Causes of the problem — conjectures

Before conjecturing about possible reasons for the difficulties introduced by noise, two facts should be remarked, namely i) the deleterious effect of the noise seems to be far more dramatic in chaotic than in non-chaotic systems, and ii) both the estimation of $\lambda_1$ and the identification of discrete models from the data were carried out by prediction-based algorithms.

Although the ability of a chaotic model to accurately predict a long time-series is limited, because of the shadowing lemma, it is still possible to make accurate estimates of $\lambda_1$ [30]. Moreover, it is also possible to identify good chaotic models because only short-term predictions are used in such cases [22, 6]. Clearly, some other reason should be found to explain the observed difficulty. A possible reason is suggested in what follows.

There seems to be a critical value of the noise level beyond which the uncertainty in the data is such that even short-term[4] predictions are precluded to the extent that accurate estimates of NARMAX polynomials become very difficult to obtain using prediction error methods.

Thus if the uncertainty in the data is such that predictions of $T_s$ time units into the future are somehow affected this will be reflected in the residuals which, during the estimation procedure, will be fedback into the model in posterior iterations. This iterative procedure will enable the effects of the uncertainty in the data to build up within the estimation algorithm (via the residuals) and ultimately prevent good estimates. It should be noted that even for nonchaotic systems the residuals are usually proportional to the noise, thus the larger the noise variance the larger the residuals for a given model. However, it appears that

[4]In this context 'short' should be understood in relation to the prediction intervals used in the algorithms, that is, the Lyapunov interval, $\Delta L$, and the sampling interval, $T_s$, and no longer in relation to the Lyapunov time $-\log_2 \sigma_e/\lambda_1$.

the estimation algorithm is able to handle large residuals rather comfortably for nonchaotic systems.

Table 1 shows some dynamical invariants for models estimated from data on the double scroll attractor with different noise levels.

Table 1. Statistics for identified models of the double scroll

| SNR (dB) | $\sigma_e^2$ | $\sigma_\epsilon^2$ | $\lambda_1{}^a$ | $D_c$ |
|---|---|---|---|---|
| 50.8 | 0.0075 | 0.016 | 0.221 | 2.05±0.025 |
| 46.7 | 0.0134 | 0.027 | 0.206 | 2.04±0.020 |
| 43.8 | 0.0169 | 0.035 | 0.238 | 2.07±0.015 |
| 42.7 | 0.0193 | 0.039 | 0.226 | 2.03±0.019 |
| 42.0 | 0.0209 | 0.042 | 0.130 | 1.84±0.008 |

[a] Calculated using $\log_e$

As can be seen, up to the model identified from the data with noise variance 0.0193 relatively good models were estimated. For the present purposes a model is considered good if it reproduces the geometry of the double scroll attractor shown in figure 1 and also possesses similar statistics such as $\lambda_1$ and the correlation dimension, $D_c$.

It is interesting to note that a relatively small increase in the noise variance from 0.0193 to 0.0209 ($\approx 8\%$) was sufficient to hamper the estimation of a good model using the same values of $\ell$, $n_y$ and $n_p$. Note that the percentage variation in $\lambda_1$ was around 42%. This suggests that in some sense, NARMAX model estimation from chaotic data may suffer from limitations of which some are similar to those encountered in the estimation of $\lambda_1$ ($\lambda_1 > 0$), which also displays a rather well-defined critical value beyond which accurate estimation becomes sensibly more difficult.

It is not being advocated that the source of errors in the estimation of $\lambda_1$ and NARMAX models is one and the same. However, some of the results presented so far suggest that, at least in some cases, the difficulties in estimating $\lambda_1$ and NARMAX models from noisy data manifest in similar ways and apparently share common features.

Thus high noise variance seems to imply i) high uncertainty in the data which is used to initialize the predictor in the parameter estimation step. This uncertainty appears to limit the predictor accuracy based on which the model parameters are updated, and ii) high residual variance which is fedback into the model after each noise iteration and in the case of chaotic models seems to 'grow' during parameter estimation.

# 5 Chaotic data filtering

If the conjectures made in the last section are correct, it would be expected that the reduction of the noise level in the data would enhance the quality of the estimated models. Therefore in practice it might be helpful to reduce the noise level to acceptable values by means of filtering techniques. After a brief overview of existing approaches, a simple way of reducing the noise level is suggested. Throughout this paper it is assumed that the noise is purely additive, or in other words the noise is entirely observational. This has become a standard

procedure in the literature [32, 9] because "while there are situations where, say, parametric or nonlinear fluctuation coupling are appropriate, experience has shown that the additive form is adequate for most modeling purposes" [11].

## 5.1 Filtering and noise reduction — existing techniques

One way of eliminating unwanted frequencies in a signal is by filtering. When the main objective is to 'clean' the data from additive noise a simple alternative is the use of low-pass filters. An obvious deficiency of this approach is that the filter will also attenuate frequency components of the signal which are above the filter cut-off frequency. This could be particularly detrimental if the signal to be filtered is chaotic with a broad-band spectrum such as those produced by the logistic and Hénon maps. Moreover, moving average (MA), global linear fitted maps and linear low-pass filters may distort the signal badly unless the data are highly oversampled [9]. Autoregressive (AR) filters can increase the dimension of the attractor [33] especially if the damping is too weak [9].

Ways of overcoming some of the aforementioned problems have been suggested in the literature and include the use of acausal filters [34] and reverse filtering [35].

A particular aspect of filtering that has attracted some attention among chaoticists is the *noise reduction problem.* Given a chaotic time series $x(t)$ contaminated by additive noise $e(t)$, it is desired to filter the measured data $y(t)=x(t)+e(t)$ in order to recover $x(t)$. This is useful in 'cleaning' Poincaré sections and embedded attractors which have been blurred by noise.

Another aspect of this problem is to find a 'noise-reduced' orbit $\bar{y}(t)$ from which statistics such as $\lambda_1$, $D_c$ and the attractor geometry can be more accurately estimated than if the noisy data $y(t)$ were used. This is sometimes referred to as *statistical noise reduction* as opposed to recovering $x(t)$ from $y(t)$ which has been called *detailed noise reduction* [36].

Filtering based on model predicted outputs, whilst reducing the noise content in the data, will not guarantee that $\bar{y}(t)$ remains close to $y(t)$ (and ultimately close to $x(t)$) if the latter is chaotic. Note that 'closeness' between $\bar{y}(t)$ and $y(t)$ is not necessarily required in *statistical filtering* but it is usually a requirement in other applications.

Ways of ensuring that $\bar{y}(t)$ remains close to $y(t)$ have been suggested which demand that $\bar{y}(t)$ be found by minimization of a cost function of the form

$$J_{NR} = \sum_{k=1}^{N} \{J_1 [\bar{y}(k) - g_k(\bar{y}(k-1))] \; + \; J_2 [\bar{y}(k) - y(k)]\} \tag{22}$$

where $J_1[\cdot]$ and $J_2[\cdot]$ indicate functions which are usually metric norms and $g_k(\cdot)$ are linear maps which describe the dynamics in a neighbourhood of a point on the true orbit. Clearly $J_1[\cdot]$ penalizes deviations from the true deterministic dynamics described by $g_k(\cdot)$ while $J_2[\cdot]$ guarantees that the cleaned orbit remains close to the measured orbit.

In particular, Kostelich and Yorke (1988, 1990) have used

$$J_1[\cdot] = \| \bar{y}(k) - g_k(\bar{y}(k-1)) \|^2 + \| \bar{y}(k+1) - g_k(\bar{y}(k)) \|^2 \tag{23}$$

where $\| \cdot \|$ is the Euclidean norm, and

$$J_2[\cdot] = \| \bar{y}(k) - y(k) \|^2 \tag{24}$$

while Farmer and Sidorowich (1991) suggested that $J_2[\cdot]$ be chosen as above and

$$J_1[\cdot] = 2 \parallel g_k(\bar{y}(k)) - \bar{y}(k+1) \parallel^{\mathrm{T}} \mu_k \tag{25}$$

where $\mu_k$ are Lagrange multipliers. The minimization of $J_{NR}$ is a typically ill-conditioned problem, especially for chaotic time series [36]. Some improvement in the numerical conditioning however can be attained at the expense of performance [37].

A clear limitation in any real noise reduction problem is that the underlying dynamics are not usually known *a priori*. If the underlying dynamics were perfectly known then maps describing the dynamics could be used to separate the predictable part of the orbit from the unpredictable portion, which is the noise. But, as often happens in practice, if the map has to be estimated (learned) from the noisy data, the noise will pose limitations on the accuracy with which the map can actually be estimated and, of course, an inaccurate map will not be able to exactly separate the noise from the true orbit.

It is therefore not surprising that the method proposed by Hammel (1990), which was derived from the proof of the *shadowing lemma*, outperforms the aforementioned methods because it assumes that the maps describing the underlying dynamics are known in advance. A method suggested by Marteau and Abarbanel (1991) does not assume that the dynamics are known but requires that some noise-free data be available to estimate a set of conditional probabilities which are subsequently used to reduce the noise in a different noisy time series. Although the importance and contribution of the two latter methods are not being questioned here, it seems, however, that the assumptions made are somewhat restrictive in most practical situations.

In the field of system identification, improving the *signal to noise ratio* (SNR) is also of interest because this facilitates both the unbiased estimation of the parameter vector and the correct determination of the model structure. The chief idea is to estimate the noise-free data and then use this estimate to perform the parameter estimation. A way of doing this is to use the following predictor which can be derived from equation (10)

$$\hat{y}(t) = \Psi_{\hat{y}u}^{\mathrm{T}}(t-1)\hat{\Theta}_{yu} \tag{26}$$

It should be realised that in the last equation the parameter vector $\hat{\Theta}_{yu}$ was estimated from the original noisy data as is indicated by the absence of the hat on the subscript $y$. On the other hand, the matrix $\Psi_{\hat{y}u}^{\mathrm{T}}(t-1)$ was formed using predicted values of the data, that is $\hat{y}(t)$ up to and including time $t-1$. Because $\hat{y}(t)$ is an estimate of $x(t)$, equation (26) can be used in suboptimal parameter estimation schemes [38].

However, if the data were chaotic after a few iterations $\hat{y}(t)$ would not be an accurate estimate of $x(t)$ because of the sensitive dependence on initial conditions. Therefore the use of $\hat{y}(t)$ in suboptimal schemes seems somewhat restricted for chaotic systems. It should be noted however that even if $\hat{y}(t)$ is not close to $x(t)$ the former might convey consistent information about the underlying dynamics, but for filtering purposes it seems appropriate to require that the filtered data resembles the original records.

## 5.2 Separating noise from data — preliminaries

Equation (5) reveals that the noisy data, $y(t)$, is composed by predictable and unpredictable components[5] and that the unpredictability stems from the inability to predict the noise at time $t$ based upon measurements up to time $t-1$. In principle, if the data were noise-free $F_u[\cdot]=0$ and $F_p[y(s),u(s)]$ would be completely deterministic.

For the time being it will be assumed that the noise $e(t)$ is white, uncorrelated with the input and purely additive, therefore $e(t)$ is totally unpredictable. Consequently a predicted time series would not include $e(t)$. Thus in what follows a predictor is sought to separate the noise from the data.

In order to use equation (5) as a predictor the following assumption should be made

**Assumption 5.1** The map $F_p[\cdot]$ and the past values of the noise, that is $e(s)$ $s \leq t-1$, are known.

The data can then be predicted by neglecting $F_u[\cdot]$, which cannot be used because $e(t)$ is unknown. Thus

$$\hat{y}(t) = F_p[y(s), u(s), e(s)] \qquad s \leq t-1 \tag{27}$$

Billings and Voon have argued that even if the noise is purely additive it will induce cross-product terms in the model, represented by $G^{yu\xi}[\cdot]$ in equation (7) [38]. If such terms are significant and are not included in the model, parameter estimates will become biased. For the sake of simplicity the following assumption is made

**Assumption 5.2** The cross-product terms, induced by the noise and the nonlinearities, are negligible[6].

In this case the cross-product terms of $F_p[\cdot]$ consist of output and input terms only. Consequently equation (27) can be rewritten as

$$\hat{y}(t) = F_{pyu}[y((s), u(s)] + F_{pe}[e(s)] \qquad s \leq t-1 \tag{28}$$

Under the assumption that $e(t)$ is white, or in other words structureless, $F_{pe}[\cdot]$, which can be viewed as the noise model, will be zero[7]. Thus

$$\hat{y}(t) = F_{pyu}[y((s), u(s)] \qquad s \leq t-1 \tag{29}$$

which indicates that in ideal conditions $\hat{y}(t)$ is the purely deterministic noise-free component of the data.

---

[5]It is realised that a purely deterministic chaotic system is long-term unpredictable along the unstable manifold although the dynamics can, in principle, be predicted to a certain extent over short periods of time along such a manifold. In what follows predictable and unpredictable should be understood in a statistical sense, see footnote 2.

[6]This assumption is made for the sake of a clearer argument but it is not needed in practice. Should this assumption be untrue in a certain application, this would be revealed by the correlation functions described in § 2.

[7]In practice, even when $e(t)$ is white $F_{pe}[\cdot]$ is allowed to have some terms in order to guarantee that the residuals will be white as well. Moreover, this also accounts for inaccuracies which are not necessarily related to the data such as uncertainties in the model structure, roundoff and numerical errors.

The basic idea behind this filtering scheme is to discriminate between the unpredictable part of the signal, that is the noise, and the purely deterministic component which is the noise-free data.

Summarising, two things should be noted. First, the use of a predictor, in principle, enables the separation of the predictable and unpredictable components. Second, the predictable component is still stochastic because $F_p[\cdot]$ in equation (27) depends on $e(s)$ $s \leq t-1$. Hence two things should be assumed in order that $F_p[\cdot]$ be purely deterministic, namely, i) that the cross-product terms involving the noise be negligible (see assumption (5.2)), and ii) that the noise be completely structureless such that $F_{pe}[\cdot] = 0$. However, two practical difficulties can be pointed out i) $F_p[\cdot]$ is not known *a priori*, and ii) because the noise cannot usually be measured separately from the data, $e(t)$ will not be known either. Consequently assumption (5.1) will not hold in most real applications. In the next subsection practical solutions to these problems are suggested.

## 5.3 The resetting filter

It has been argued that, because the noise is totally unpredictable, the true deterministic signal can be distinguished from the noise by means of prediction techniques. In other words, if a predictor is found such that the deterministic part of the data is *perfectly* predicted, the predicted trajectory will be the original noise-free data. Therefore this procedure eliminates the noise by not predicting it (note that this is most natural since the noise has been assumed to be white and consequently unpredictable in nature). This approach contrasts with some classical filtering techniques which eliminate the noise by frequency attenuation.

Hence it is crucial in this approach that most of the dynamics in the data be predicted and to achieve this the model structure must be 'flexible' enough. A common practice in linear filtering is the use of overparametrized *moving average* (MA) and *autoregressive* (AR) models [15]. However, it seems that even overparametrized linear models will not in general predict the nonlinearities in the data, unless highly oversampled data are used which, on the other hand, will induce numerical difficulties.

Because $F_p[\cdot]$ is not known, in practice this map has to be estimated from the data and even if the noise is assumed to be white, noise terms should be included to guarantee that the residuals will also be white. In other words noise terms are included to ensure that all the dynamics in the data have been learned and will be predicted.

From what has been discussed, a practical filter would be of the form

$$\hat{y}(t) = \Psi_{\hat{y}u}^{\mathrm{T}}(t-1)\,\hat{\Theta}_{yu} + \Psi_{\hat{\xi}}^{\mathrm{T}}(t-1)\,\hat{\Theta}_{\xi} \tag{30}$$

where the hat indicates estimated values and the residuals have been used as estimates of the noise.

Two difficulties however still need to be settled. First, in § 5.1 it was argued that prediction-based filters would not in general perform satisfactorily if the data were chaotic because the filtered signal would not be constrained to remain close to the original data. Second, from equation (30) it is clear that $\hat{y}(t)$ depends on $\xi(s)$ $s \leq t-1$, and because $\hat{y}(t)$ is used to determine $\hat{y}(t+1)$ the residuals would be reintroduced into the predictor. Even if the residuals are zero-mean and white, because of the nonlinearities in $\Psi_{\hat{y}u}^{\mathrm{T}}(t-1)$, the effect of the 're-used' residuals cannot be assumed to converge to zero.

The second difficulty can be overcome by using suboptimal least squares techniques which neglect $\Psi_{\xi}^{T}(t-1)\,\hat{\Theta}_{\xi}$ [38]. This can only be done safely if the residuals are white or nearly white and this presupposes that the structure of $\Psi_{\hat{y}u}^{T}(t-1)$ has been adequately chosen and that $\hat{\Theta}_{yu}$ is unbiased. However, these assumptions are not realistic in many situations because the noise will preclude the correct choice of model structure and the unbiased estimation of the parameter vector. Moreover, this approach leaves the first difficulty unresolved.

The following predictor overcomes the two aforementioned difficulties

$$\hat{y}(t) = \Psi_{yu}^{T}(t-1)\,\hat{\Theta}_{yu} + \Psi_{\xi}^{T}(t-1)\,\hat{\Theta}_{\xi} \tag{31}$$

It should be noted that in this case $\hat{y}(t)$ is predicted based on previous values of the measured data $y(s)$ $s \leq t-1$ and not based on previously predicted values such as in equation (30). Moreover, since this predictor is used to predict only one step into the future, the predicted value $\hat{y}(t)$ is, in most cases, guaranteed to remain close to the data $y(t)$. This can be interpreted as being a consequence of the *resetting* effect achieved by using measured data to initialise the predictor at each step. The predictor in equation (31) will be referred to as the *resetting filter* (RF) and it is adequate for filtering chaotic signals.

**Remark 5.1** The similarity of equations (31) and (9) is evident. The resetting filter in equation (31) does not include the terms $\Psi_{yu\xi}^{T}(t-1)\,\hat{\Theta}_{yu\xi}$ because of assumption 5.2. However, in practice this is not necessary and if such terms are *not* negligible (this would become evident from the correlation tests) the resetting filter would be identical to the OSA predictor shown in equation (9) and the filtered data $\hat{y}(t)$ would be the OSA prediction of $y(t)$.

**Remark 5.2** The qualitative effect attained by the resetting filter is, in some respects, analogous to other methods. This can be verified by considering equation (22). It is worth noticing that the resetting effect of the RF guarantees that $J_2$ is kept small. Moreover, the parameter vector of the RF is obtained by minimising $J_{LS}$ in equation (11), which is clearly analogous to $J_1$ in equation (22). The main difference is that whilst $g_k(\cdot)$ usually represents local linear maps, $\Psi^{T}(t-1)\,\hat{\Theta}$ is a global nonlinear map which may include inputs and residuals in addition to output terms.

**Remark 5.3** Predictor based filtering for chaotic systems will not work in general because of the inability of making long-term acurate predictions along the unstable manifold. Therefore in such directions, the filter would actually amplify the noise [39]. The same is valid for the RF, but to a much lesser extent because of the resetting effect which will guarantee that any noise amplification along the unstable manifold is kept to a minimum.

**Remark 5.4** In this work the final objective is to obtain filtered data from which good NARMAX polynomial models can be estimated. Consequently a relatively small increase in the SNR may well be considered satisfactory provided that such an increase is sufficient to reduce the uncertainty in the data below the critical value conjectured in § 4.5.

# 6 Requirements on the resetting filter

The aim of this section is to investigate some of the requirements on the RF in order that a *statistically* sound model be estimated from the filtered data.

## 6.1 White noise

In order to keep the analysis focussed on the main ideas a simple example is first presented which assumes that the data are linear.

### 6.1.1 A linear model

In this example it is assumed that the underlying dynamics are described by

$$x(t) = a_1 x(t-1) + b_1 u(t-1) \tag{32}$$

and that the noise $e(t)$ is white. Thus the measured data can be represented as follows

$$\begin{aligned} y(t) &= x(t) + e(t) \\ y(t) &= a_1 x(t-1) + b_1 u(t-1) + e(t) \\ y(t) &= \Psi_{xu}^{\mathrm{T}}(t-1)\Theta_{yu} + e(t) \end{aligned} \tag{33}$$

Consider the following parametrization

$$\begin{aligned} y(t) &= a_1 y(t-1) + b_1 u(t-1) + c_1 \xi(t-1) + \xi(t) \\ y(t) &= \Psi_{yu}^{\mathrm{T}}(t-1)\Theta_{yu} + \Psi_{\xi}^{\mathrm{T}}(t-1)\Theta_{\xi} + \xi(t) \end{aligned} \tag{34}$$

where $\Theta_{yu} = [a_1 \ b_1]^{\mathrm{T}}$ and one noise term has been included in the model. The following resetting filter can be obtained from the last equation

$$\hat{y}(t) = \hat{a}_1 y(t-1) + \hat{b}_1 u(t-1) + \hat{c}_1 \xi(t-1) \tag{35}$$

Expressing the filtered signal in terms of the noise-free data $x(t)$ yields

$$\begin{aligned} \hat{y}(t) &= \hat{a}_1[x(t-1) + e(t-1)] + \hat{b}_1 u(t-1) + \hat{c}_1 \xi(t-1) \\ \hat{y}(t) &= \Psi_{xu}^{\mathrm{T}}(t-1)\hat{\Theta}_{yu} + \hat{c}_1 \xi(t-1) + \hat{a}_1 e(t-1) \end{aligned} \tag{36}$$

The bias[8] of the estimated parameter vector $\hat{\Theta}_{yu}$ is defined as

$$B_{yu} \doteq \mathrm{E}\{\Theta_{\epsilon}\} \tag{37}$$

where $\mathrm{E}\{\cdot\}$ denotes mathematical expectation and

$$\Theta_{\epsilon} = \hat{\Theta}_{yu} - \Theta_{yu} \tag{38}$$

then equation (36) can be written as

$$\hat{y}(t) = \Psi_{xu}^{\mathrm{T}}(t-1)[\Theta_{yu} + \Theta_{\epsilon}] + \hat{c}_1 \xi(t-1) + \hat{a}_1 e(t-1) \tag{39}$$

[8]This definition requires that both the estimated and true parameter vectors have the same dimensions. In the next section a more intuitive concept of bias will be used which will not be restricted to models with the same structure.

Further analysis and numerical calculations show that in this particular case $\hat{a}_1 \approx -\hat{c}_1$. Therefore equation (39) can be rewritten as

$$\hat{y}(t) = \Psi_{xu}^{\mathrm{T}}(t-1)\Theta_{yu} + \Psi_{yu}^{\mathrm{T}}(t-1)\Theta_{\epsilon} + \hat{a}_1[e(t-1) - \xi(t-1)] \tag{40}$$

Comparing the latter equation to (33) it becomes clear that while the measured data, $y(t)$, have a certain degree of uncertainty due to the noise $e(t)$, the filtered data at time $t$, $\hat{y}(t)$, do not depend on the unknown noise at time $t$ which was eliminated during prediction. On the other hand, the filtered data have two other sources of uncertainty, namely i) a term due to the difference, $\Theta_{\epsilon}$, between the true and the estimated parameter vectors, and ii) a term due to the error between the noise and the residuals.

If the map which describes the underlying dynamics were known, $\Theta_{\epsilon}=0$ and consequently a perfect separation of noise and deterministic data would be possible at least in principle. This would therefore imply $\xi(t)=e(t)$ and, in the light of equations (33) and (40) it can be seen that ideally

$$\begin{aligned} \hat{y}(t) &= \Psi_{xu}^{\mathrm{T}}(t-1)\Theta_{yu} = y(t) - e(t) \\ \hat{y}(t) &= x(t) \end{aligned} \tag{41}$$

In order to investigate the bias of a parameter vector estimated from the filtered data, the following parametrization is used[9]

$$\begin{aligned} \hat{y}(t) &= \tilde{a}_1\hat{y}(t-1) + \tilde{b}_1 u(t-1) + \tilde{\xi}(t) \\ \hat{y}(t) &= \Psi_{\hat{y}u}^{\mathrm{T}}(t-1)\Theta_{\hat{y}u} + \tilde{\xi}(t) \end{aligned} \tag{42}$$

Noting that $\xi(t)=y(t)-\hat{y}(t)$, the last equation can be expressed as

$$\begin{aligned} y(t) &= \tilde{a}_1 y(t-1) + \tilde{b}_1 u(t-1) + \tilde{\xi}(t) + \xi(t) - \tilde{a}_1\xi(t-1) \\ y(t) &= \Psi_{yu}^{\mathrm{T}}(t-1)\Theta_{\hat{y}u} + \tilde{\xi}(t) + \xi(t) - \tilde{a}_1\xi(t-1) \end{aligned} \tag{43}$$

The least squares estimate of $\Theta_{\hat{y}u}$ obtained from the last equation is given by the well known expression

$$\hat{\Theta}_{\hat{y}u} = \mathbf{A}_{yu}y(t) \qquad t = 1, 2, \ldots, N \tag{44}$$

where

$$\mathbf{A}_{yu} = [\Psi_{yu}(t-1)\Psi_{yu}^{\mathrm{T}}(t-1)]^{-1}\Psi_{yu}(t-1) \tag{45}$$

Therefore, the bias in the parameter vector estimated from the filtered data is

$$\mathrm{E}\{\hat{\Theta}_{\hat{y}u}\} - \Theta_{yu} = \mathrm{E}\{\mathbf{A}_{yu}y(t)\} - \Theta_{yu} \tag{46}$$

and substituting equation (43) into (46) gives

[9] In a real application this parametrization would correspond to the final model.

$$
\begin{aligned}
\mathrm{E}\{\hat{\Theta}_{\hat{y}u}\} - \Theta_{yu} &= \mathrm{E}\{\mathbf{A}_{yu}\Psi_{yu}^{\mathrm{T}}(t-1)\Theta_{\hat{y}u} - \Theta_{yu}\} + \mathrm{E}\{\mathbf{A}_{yu}\xi(t)\} \\
&\quad -\tilde{a}_1\mathrm{E}\{\mathbf{A}_{yu}\xi(t-1)\} + \mathrm{E}\{\mathbf{A}_{yu}\tilde{\xi}(t)\} \\
&= \mathrm{E}\{\Theta_\epsilon\} + \mathrm{E}\{\mathbf{A}_{yu}\xi(t)\} - \tilde{a}_1\mathrm{E}\{\mathbf{A}_{yu}\xi(t-1)\} + \mathrm{E}\{\mathbf{A}_{yu}\tilde{\xi}(t)\} \qquad (47)
\end{aligned}
$$

Two conditions must be satisfied in order that the bias be zero, namely i) that the resetting filter itself be unbiased, this will imply $\mathrm{E}\{\Theta_\epsilon\} = 0$, by definition, and both the second and third terms of the right hand side of the last equation will also be null because if the filter is unbiased the residual sequence $\xi(t)$ is uncorrelated with the data in $\mathbf{A}_{yu}$, and ii) that the residuals of the model in equation (42) also be uncorrelated with the raw data. It is noted that if the filter is unbiased all the predictable part of the raw data is preserved in $\hat{y}(t)$. Consequently, if $\tilde{\xi}(t)$ is not to be correlated with the raw data it is necessary that all the dynamics, that is the predictable part, of the filtered data be adequately modelled by equation (42). In other words, the fourth term in the right hand side of the last equation will be null if the model in equation (42) is unbiased with respect to the filtered data.

### 6.1.2 Nonlinear models

The preceding analysis was carried out for a simple linear system. Nonetheless this included all the main points which are also relevant to nonlinear systems and which will be discussed further in what follows.

The main conclusions of § 6.1.1 were that in order to avoid introducing bias in the filtered data it is necessary that the filter be unbiased with respect to the raw data and that the final model be unbiased with respect to the filtered data.

These conclusions remain true for nonlinear systems. The difference, however, resides in the concept of bias. A linear model is unbiased if $\Phi_{\xi\xi}(\tau)$ and $\Phi_{\xi u}(\tau)$ satisfy the first two conditions in equation (14). If the data are nonlinear, it is known that these correlation tests are not sufficient and will not, in general, indicate the presence of unmodelled nonlinear dynamics in the residuals [18]. Thus a nonlinear model is unbiased if *all* the tests of equation (14) are satisfied, in the case of nonautonomous systems, or if *all* the conditions in equation (15) are met in the case of autonomous systems. Consequently the resetting filter must satisfy the aforementioned correlations tests in order not to introduce bias in the filtered data.

A useful way of interpreting these results is to see the act of filtering as the prediction of the underlying dynamics. Thus predicting the data has the desirable side effect of not predicting whatever cannot be predicted and in so doing reducing the noise. If the filter is unbiased it will not leave any relevant dynamics unmodelled in the residuals. In other words, if the filter is unbiased all the dynamics have been learned by the filter and will be used to reduce the noise. Hence the resetting filter can be seen as *predicting the dynamics through the noise.*

If the data are nonlinear, in order that the filter be able to learn the underlying dynamics, it is necessary that the nonlinearities in the data be well represented in the structure of the filter. If this is achieved the underlying dynamics will be adequately learned during parameter estimation and the filter will be unbiased.

Therefore the concept of bias can be associated to the mechanism of finding mathematical representations for the dynamics in the data and including such representations in the

model. Thus if the dynamics are *well represented* by the model structure, they will also be learned during parameter estimation and the final model will be unbiased. Conversely, if the structure of a model is not adequate the dynamics will not be accurately learned and consequently unmodelled dynamics will appear in the residuals and will be detected by the correlation tests indicating bias.

Concerning the choice of the parametrizations in equations (34) and (42), it should be realised that the choice of the filter structure is not as critical as it is for the final model. The only requirement on the filter structure is that it should be complex enough to adequately capture the underlying dynamics in a local sense. Because the resetting filter is used to predict just one step ahead, the harmful effects of overparametrization as well as the inaccuracies induced by the presence of noise in the raw data apparently do not jeopardise the performance as much as if the filter were used in a global way, that is with no resetting. The primary concern of the filter is to predict whatever is predictable in the data on a one-step-ahead basis.

By contrast, the structure of the final model should be carefully chosen since it is desired that this model be a faithful approximation of the underlying system and not just a prediction of one data set. It is well known that global polynomials do not always extrapolate well beyond their domain of validity [7], that they tend to oscillate wildly [8] and that for systems with many degrees of freedom global polynomials become impractically large [9]. Nevertheless it has been suggested that these effects and others such as the appearance of spurious dynamical regimes, ghost effects and instabilities are usually a consequence of overparametrization and can sometimes be considerably alleviated if appropriate structure detection methods are used [6].

Summarising, in practice there is no need to choose the same structure for both the filter and the final model. This was done in § 6.1.1 for the sake of simplicity only. However, the filter structure must be complex enough to enable adequate learning of the underlying dynamics. This implies that the models must be unbiased and in the case of nonlinear systems this can be easily verified using correlation tests. Other aspects related to the choice of the filter and model structures will be considered in § 6.3 and § 8.

## 6.2 Correlated noise

In this section it is assumed that the noise corrupting the data is expressed by

$$\eta(t) = c\,e(t-1) + e(t) \tag{48}$$

where $e(t)$ is white. Clearly, the noise $\eta(t)$ is *correlated* or, in other words, *coloured*. In this case the measured data are $y(t) = x(t) + \eta(t)$.

The analysis for this case is very similar to the one described in § 6.1 and the main results of that section hold for correlated noise. However, there is a subtle difference which will be pointed out adopting a rather more heuristic approach.

It should be appreciated that if the same procedure described in § 6.1.1 were followed in the case of correlated noise, the RF would still reduce the unpredictable part of the noise which is $e(t)$. Because the noise is correlated, there is a 'predictable' part of the noise which is represented by the term $c\,e(t-1)$. This portion of the noise is not removed from the data by the RF in equation (35). In order to see this in a different way, consider the following

$$\begin{aligned} \hat{y}(t) &= y(t) - \xi(t) + f(\Theta_\epsilon) \\ \hat{y}(t) &= x(t) + \eta(t) - \xi(t) + f(\Theta_\epsilon) \end{aligned} \tag{49}$$

where $f(\Theta_\epsilon)$ indicates a term which is due to the difference between the true and the filter (estimated) parameter vectors. If the map describing the underlying dynamics is known then $f(\Theta_\epsilon)=0$ and $\xi(t)=e(t)$. If the noise is white $\eta(t)=e(t)$ and consequently $\hat{y}(t)=x(t)$. However if the noise is correlated such as in equation (48) then

$$\hat{y}(t) = x(t) + c\,e(t-1) \tag{50}$$

As argued above, $\hat{y}(t)$ contains the predictable part of the measured data. In the case where the noise is coloured, the predictable part of the noise is correlated with previous measurements and consequently it appears in the filtered data together with the purely deterministic orbit.

Equation (50) also suggests that an estimate of $x(t)$ can be obtained by removing from $\hat{y}(t)$ the correlated part of the noise, that is $\hat{y}(t)-c\,e(t-1)=x(t)$ where in practice an estimate of $e(t-1)$ would be used. This suggests that the resetting filter in equation (35) could be used if the last term in the equation were omitted, thus

$$\begin{aligned} \hat{y}(t) &- \hat{c}_1\xi(t-1) = \hat{a}_1 y(t-1) + \hat{b}_1 u(t-1) \\ \hat{y}_c(t) &= \hat{a}_1 y(t-1) + \hat{b}_1 u(t-1) \approx x(t) \end{aligned} \tag{51}$$

Hence, a more general resetting predictor can be defined[10]

$$\hat{y}(t) = \Psi_{yu}^{\mathrm{T}}(t-1)\,\hat{\Theta}_{yu} + \Psi_{\xi}^{\mathrm{T}}(t-1)\,\hat{\Theta}_{\xi} - G^{\xi}[\cdot] \tag{52}$$

where $G^{\xi}[\cdot]$ is the noise model and it is suggested that this model can be taken as follows

$$\begin{cases} G^{\xi}[\cdot] = 0 & \text{if the noise is white} \\ G^{\xi}[\cdot] = \Psi_{\xi}^{\mathrm{T}}(t-1)\,\hat{\Theta}_{\xi} & \text{if the noise is correlated} \end{cases} \tag{53}$$

The last equalities were based on the fact that if the noise is white it is also structureless and therefore $G^{\xi}[\cdot]=0$. In the case of correlated noise it is suggested that the 'noise dynamics' are modelled by $\Psi_{\xi}^{\mathrm{T}}(t-1)\,\hat{\Theta}_{\xi}$ and, since $G^{\xi}[\cdot]$ can be viewed as an estimate of the noise model, the second equality above follows.

It should be noted that in many practical situations it will not be known if the noise is white or correlated. Thus the following procedure is suggested. A model is estimated from the raw data. If such a model does not represent the underlying dynamics then, if it is unbiased, take $G^{\xi}[\cdot]=0$ in equation (52) and filter the data. A model is then estimated from these filtered data. If such a model is accurate then the procedure may be stopped, otherwise the data should be filtered again. In this second stage the filtering is based on the second estimated model and the two options of equation (53) are possible thus yielding two different filters. If the noise in the raw data was correlated it would be expected that the

[10]Note that the definition in equation (31) was based on the assumption that the noise was white.

choice $G^{\xi}[\cdot] = \Psi_{\xi}^{T}(t-1)\,\hat{\Theta}_{\xi}$ would yield a filtered data sequence from which a better final model is estimated. This procedure will be illustrated in § 7.

The filtered data is ideally the predictable part of the raw data. All the dynamics in $y(t)$ are modelled by the first two terms in the right hand side of equation (52). If the noise is correlated it will also have some dynamics which will be incorporated in $\Psi_{\xi}^{T}(t-1)\,\hat{\Theta}_{\xi}$. The presence of such dynamics in $\hat{y}(t)$ can be avoided by *not* predicting them. This, of course, can only be done if the portion responsible for predicting the noise dynamics is omitted from the filter. This is indicated by the last term in equation (52).

The discussion above is also valid for nonlinear systems. If the data are nonlinear the resetting filter in equation (52) will also be nonlinear. Because the filtering achieved by the RF is a one-step-ahead-prediction-based procedure it is vital that the underlying dynamics be well represented by $\Psi_{yu}^{T}(t-1)\,\hat{\Theta}_{yu}$ and that, if the noise is correlated, the noise dynamics be well modelled by $\Psi_{\xi}^{T}(t-1)\,\hat{\Theta}_{\xi}$. This requires that the RF be unbiased as discussed in § 6.1.2.

## 6.3 Limitations on the RF performance

The major impediment for achieving complete noise reduction is that the dynamics are not known and should be learned from a finite record of noisy data. Since the RF is estimated in the same way as the final model and in view of the examples provided in § 4.1, it is clear that the identification of a filter suffers from the same limitations as the estimation of the final model itself.

The main difference is that the latter is required to learn the underlying dynamics in a global sense. This means that the model should be able to reproduce dynamical invariants of the system such as attractor geometries, Lyapunov exponents, fractal dimensions, bifurcation patterns etc. On the other hand, the filter, which is also a global polynomial estimated only once based on the entire data records, is only required to perform *locally* on the data. This can be seen as another consequence of the resetting effect peculiar to the RF. As the data is predicted and the noise reduced, the filter is reset at each step. The number of data points required to reset the filter equals the maximum lag of the model, $n_y$, which is usually small (typically less than ten) and thus corresponds to a very narrow window of the data records. Therefore a certain polynomial that would not perform satisfactorily as a global model could be successfully used as a resetting filter. This does not mean, however, that the filter is immune to the noise but that, as a consequence of the way the filter is used, a model may not reproduce dynamical invariants but may give good results as an RF.

This distinction between the performance of a resetting predictor and a polynomial model has been pointed out in a different context: "using a fitted polynomial model to simulate the underlying process may often result in divergent behaviour. It is, however, quite another matter to compute $h$-step-ahead predictions using a polynomial model. This is because in the latter case observed past time-series values are used in the calculation. For a stationary nonlinear time series generated from an underlying process within a stable region, provided that $h$ is not too large, the prediction calculated based on observed time-series values should remain stable" [40].

The deleterious effects of too high a noise level can impair the quality of an estimated model to the extent that even if such a model were used as a resetting filter the final noise reduction attained would be insufficient to enable the estimation of good models from the

filtered data.

Another situation in which the performance of the RF is diminished occurs when the data are, in nature, rather unpredictable such as the time sequences generated by the logistic and Hénon maps. The autocorrelation function of the time series of these maps resemble that of white noise, that is the correlation time is extremely short. In such cases the RF would have difficulties in separating the unpredictable noise from the nearly unpredictable data. Thus it seems that the RF is better suited in applications where the clean data are smooth with relatively long correlation times.

Two ways of improving the performance of the RF are suggested in what follows. The lower bound for the variance of the residuals, $\sigma_\xi^2$, is the variance of the white noise in the data, $\sigma_e^2$, thus $\sigma_\xi^2 \geq \sigma_e^2$. Increasing the number of terms in the RF will, in general, reduce $\sigma_\xi^2$. Therefore terms can be added to the filter until $\sigma_\xi^2$ is reasonably close to $\sigma_e^2$. It should be noted, however, that overparametrizing nonlinear models may induce spurious dynamics and consequently injudicious overparametrization should be avoided.

The performance of the RF can also be improved by performing the filtering on an oversampled set of measured data. After filtering the data should then be decimated in order to provide an adequate sampling. It has been shown that the sampling rate influences the dynamics of the final model and that, for most 'smooth' continuous systems, small variations in the value of the sampling rate are automatically compensated by the identification algorithm by means of selecting a slightly different model structure [6].

Highly oversampled data should always be avoided because they could induce numerical problems due to the data being nearly singular.

Noise reduction techniques based on local maps can achieve high levels of noise reduction [36]. This is a consequence of a higher accuracy achieved in learning the dynamics *locally* over a large (typically fourty) number of neighbourhoods in the reconstructed state space. Such techniques can also be used to filter the data which will be used for identification purposes but would require a time-demanding algorithm which is very different from the one used in the identification.

# 7 Numerical results

In this section some of the ideas discussed in § 4, § 5 and § 6 are illustrated by means of numerical simulations.

## 7.1 The double-scroll attractor

The objective in this example is to identify a NARMAX polynomial model for Chua's circuit from noisy data on the double-scroll attractor. In order to illustrate the procedure, the same SNR (42 dB) and the same model structure ($\ell = 4$, $n_y = 5$ and $n_u = 0$) used in example in § 4.1 were considered. As before, the noise is zero-mean, white and gaussian.

The noisy data were sampled at $T_s = 0.015$ in order to improve the performance of the filter. The first 2000 points of this data set were used to estimate the following model

$$\begin{aligned} z(k) &= 0.80773z(k-1) + 0.50704 \times 10^{-1} z(k-2) + 0.31996z(k-3) \\ &- 0.13370z(k-10) + 0.25039 \times 10^{-1} z(k-4) - 0.50831 \times 10^{-1} z(k-9) \end{aligned}$$

$$
\begin{aligned}
&- \; 0.15806 \times 10^{-1} z(k-5) - 0.14443 \times 10^{-2} z(k-1)^2 z(k-7) \\
&- \; 0.55160 \xi(k-1) - 0.91378 \times 10^{-1} \xi(k-18) + 0.14944 z(k-10) \\
&+ \; 0.57925 \times 10^{-1} \xi(k-15) + 0.72448 \times 10^{-1} \xi(k-17) - 0.24648 \xi(k-3) \\
&- \; 0.52401 \times 10^{-1} \xi(k-6) - 0.40081 \times 10^{-1} \xi(k-13) + \xi(k)
\end{aligned} \tag{54}
$$

where $\xi(k)$ is zero-mean with variance $\sigma_\xi^2 = 0.026$. This model was obtained taking $\ell = 3$, $n_y = 10$, $n_u = 0$ and considering 20 linear noise terms as candidates, namely $\xi(k-i)\; i = 1, 2, \ldots, 20$. The *Akaike information criterion* (AIC) [41] was used to truncate the model. It is noted that the eight process terms above were chosen among 1440 candidate terms according to the ERR criterion described in § 2.4. The correlation plots of this model are shown in figure 8 and suggest that there are no significant unmodelled terms in the residuals thus qualifying this model as a prospective filter.

Disregarding the last term in equation (54), the entire data record was filtered as discussed in § 5.3. It is noted that this is equivalent to taking $G^\xi[\cdot] = 0$ in equation (52). The filtered sequence was subsequently decimated in order to produced data with an appropriate sampling interval. Thus 1750 data points of the decimated records were subsequently used to identify models for the double-scroll attractor.

Considering the same set of candidate terms as in example 1 in § 4.1, that is $\ell = 4$, $n_y = 5$, $n_u = 0$ and considering 20 linear noise terms, the following model was estimated from the filtered/decimated data

$$
\begin{aligned}
z(k) \;=\; & 2.0168 z(k-1) - 0.59826 z(k-2) - 0.19876 \times 10^{-1} z(k-1) z(k-3) z(k-4) \\
+ \; & 0.51415 z(k-5) - 0.43396 z(k-3) - 0.31361 z(k-4) \\
+ \; & 0.63095 \times 10^{-1} z(k-1) z(k-2) z(k-5) - 0.77961 \times 10^{-1} z(k-1)^3 \\
+ \; & 0.93462 \times 10^{-1} z(k-1) z(k-2) z(k-3) - 0.81817 \times 10^{-1} z(k-1) z(k-5)^2 \\
- \; & 0.14554 z(k-1)^2 z(k-5) - 0.52463 \times 10^{-1} z(k-5)^2 \\
+ \; & 0.97526 \times 10^{-1} z(k-3) z(k-5) - 0.50276 \times 10^{-1} z(k-3)^3 \\
+ \; & 0.66313 \times 10^{-1} z(k-1)^2 z(k-3) + 0.29260 \times 10^{-1} z(k-1) z(k-3) z(k-5) \\
+ \; & \Psi_\xi^T(k-1) \hat{\Theta}_\xi + \xi(k)
\end{aligned} \tag{55}
$$

where $\sigma_\xi^2 = 0.012$.

The double-scroll attractor reconstructed from the original noisy data is shown in figure 9a. Figure 9b shows the filtered data which were subsequently used in the identification. Finally, figure 9c shows the double-scroll attractor obtained by simulating the identified model of equation (55). The latter compares very well to the the attractor of the original system shown in figure 1. Moreover, the largest Lyapunov exponent and the correlation dimension for the identified model are respectively $\lambda_1 = 0.205$ and $D_c = 2.00 \pm 0.038$. Comparing these values to last row in table 1 and the attractors in figures 4 and 13 it is clear that the filtering of the noisy data has indeed enabled the identification of a model which reproduces fairly well some of the dynamical properties of the original system.

## 7.2 The Duffing-Ueda oscillator — the white noise case

The main objective in this example is to filter the data used in § 4.1.2 to estimate a model from the filtered data and assess the benefits due to filtering.

In order to illustrate that the structure of the filter is not necessarily critical (as long as it is unbiased in a nonlinear sense), the same model estimated in § 4.1.2 was be used to filter the data. The correlation tests for this model indicate that no unmodelled terms have been detected in the residuals. According to equations (52)–(53), the noise terms represented by $\Psi_\xi^T(k-1)\hat{\Theta}_\xi$ in equation (21) should be used to predict (filter) the data, that is $G^\xi[\cdot]=0$. In this case $\Psi_\xi^T(k-1)\hat{\Theta}_\xi$ consists of the twenty linear noise terms

$$\begin{aligned}
\Psi_\xi^T(k-1)\hat{\Theta}_\xi &= -0.67953\xi(k-1)+0.16246\xi(k-6)+0.4964\times10^{-1}\xi(k-20)\\
&- 0.18021\xi(k-3)-0.39983\times10^{-1}\xi(k-10)-0.48734\times10^{-1}\xi(k-4)\\
&- 0.31848\times10^{-1}\xi(k-18)+0.23195\times10^{-1}\xi(k-11)+0.2988\times10^{-1}\xi(k-12)\\
&+ 0.25606\times10^{-1}\xi(k-13)-0.25749\times10^{-1}\xi(k-15)+0.88453\times10^{-1}\xi(k-5)\\
&- 0.14239\xi(k-2)-0.12516\times10^{-1}\xi(k-14)-0.11704\times10^{-1}\xi(k-9)\\
&- 0.11033\times10^{-1}\xi(k-17)-0.10393\times10^{-1}\xi(k-19)-0.5001\times10^{-2}\xi(k-7)\\
&- 0.38589\times10^{-2}\xi(k-16)+0.36169\times10^{-2}\xi(k-8) \qquad (56)
\end{aligned}$$

Subsequently the input time series and the filtered output time series were used to identify the model

$$\begin{aligned}
y(k) &= 1.4205y(k-1)-0.28361y(k-2)-0.36429\times10^{-2}y(k-1)^3\\
&+ 0.10183\times10^{-1}u(k-2)-0.21327y(k-5)-0.99693\times10^{-3}y(k-1)y(k-5)^2\\
&+ 0.77163\times10^{-1}y(k-3)-0.64201\times10^{-2}u(k-3)+0.69093\times10^{-3}u(k-4)\\
&- 0.32431\times10^{-2}y(k-2)y(k-4)^2+0.15267\times10^{-2}u(k-5)\\
&+ 0.16366\times10^{-2}y(k-2)y(k-3)y(k-5)+\Psi_\xi^T(k-1)\hat{\Theta}_\xi+\xi(k) \qquad (57)
\end{aligned}$$

where $\xi(k)$ is zero-mean, white with variance $\sigma_\xi^2=0.0012$. It is noted that the noise model $\Psi_\xi^T(k-1)\hat{\Theta}_\xi$ in the latter equation is different from the one in equation (56) which corresponds to the filter in (21). The bifurcation diagram and the Poincaré section for $A=5.7$ of the deterministic part of this model are shown in figure 10. As can be seen, this model has all the bifurcation points as the original system and is clearly much better than the bifurcation diagram of the model estimated from the raw data. Similar comments hold for the respective Poincaré sections.

## 7.3 The Duffing-Ueda oscillator — the correlated noise case

In this example, the data were corrupted with correlated noise according to $y(t)=x(t)+v(t)$ where $v(t)=k[0.3e(t)+0.2e(t-1)+0.1e(t-2)]$ and $e(t)$ is a zero-mean white gaussian process and the constant $k$ was chosen such that the resulting SNR, that is $20\log_{10}(\sigma_v^2)/(\sigma_x^2)$, was close to the one considered in § 4.1.2 and § 7.2.

Choosing $\ell=3$, $n_y=n_u=6$, $n_p=7$ and allowing the noise model to be composed of ten terms taken from $\xi(t-i)$ $i=1,2,\ldots,20$ a model was estimated

$$y_1(k)=\Psi_{yu}^{\tau}(t-1)\hat{\Theta}_{yu}+\Psi_{\xi}^{\tau}(t-1)\hat{\Theta}_{\xi}+\xi(k) \tag{58}$$

for which $\sigma_{\xi}^2=0.013$. The bifurcation diagram of this model was obtained by simulation of the purely deterministic part which is composed only by output and input terms and revealed that some features of the bifurcation structure of the original system had been lost. Because the correlation tests of this model indicated that no dynamics was present in the residuals, $y_1(k)-\xi(k)$ was used as a filter. This is equivalent to taking $G^{\xi}[\cdot]=0$ in equation (52). The resulting filtered data were subsequently used to estimate a model

$$y_2(k)=\Psi_{\hat{y}_1u}^{\tau}(t-1)\hat{\Theta}_{\hat{y}_1u}+\Psi_{\xi}^{\tau}(t-1)\hat{\Theta}_{\xi}+\xi(k) \tag{59}$$

for which $\sigma_{\xi}^2=0.011$. The bifurcation diagram of this model also revealed that the correct bifurcation pattern was not estimated properly. This inability of identifying a good model from the filtered data can be attributed to the following i) the noise reduction attained with the filter in equation (58) was insufficient, and ii) the noise dynamics are still in the data because the terms in $\Psi_{\xi}^{\tau}(k-1)\hat{\Theta}_{\xi}$ were used to filter the data.

In order to investigate this point further two experiments were performed on the data filtered using $\hat{y}_1=y_1(k)-\xi(k)$ as discussed above. In the first experiment the model of equation (59) was used to filter the data as before, that is including the terms in $\Psi_{\xi}^{\tau}(k-1)\hat{\Theta}_{\xi}$. This would usually provide further noise reduction. In the second experiment, the model of equation (59) was used to filter the data but without the terms in $\Psi_{\xi}^{\tau}(k-1)\hat{\Theta}_{\xi}$. This is in accordance with equations (52)–(53) and would tend to eliminate the noise dynamics from the data although the resulting SNR would probably not be as high as in the first experiment. These experiments have been included in Table 2.

It should be noted that in either case the model used as a filter, namely equation (59), must be unbiased. This was promptly verified from the correlation tests. Thus the first experiment yielded the following model

$$\begin{aligned}
y(k) &= 0.80966y(k-1)-0.31356y(k-4)-0.49992\times10^{-2}y(k-1)^2y(k-3)\\
&+ 0.11160\times10^{-1}u(k-1)-0.18290y(k-5)+0.68840y(k-2)\\
&- 0.35454\times10^{-2}y(k-1)y(k-2)y(k-4)+0.37183\times10^{-2}u(k-5)\\
&- 0.66180\times10^{-2}u(k-2)-0.15476\times10^{-2}y(k-3)\\
&+ \Psi_{\xi}^{\tau}(k-1)\hat{\Theta}_{\xi}+\xi(k)
\end{aligned} \tag{60}$$

for which $\sigma_{\xi}^2=0.0076$. Figures 11a–b show the bifurcation diagram and Poincaré section for $A=5.7$ of the deterministic part of the model in equation (60). The bifurcation quality index defined in equation (16) for this model is $J_b=0.161$. The improvements attained are obvious. On the other hand, in the second experiment the following model was estimated

$$\begin{aligned}
y(k) &= 0.12151\times10y(k-1)-0.46716y(k-3)-0.30016\times10^{-2}y(k-1)^2y(k-5)\\
&+ 0.10659\times10^{-1}u(k-1)-0.14507y(k-5)+0.19389y(k-2)\\
&- 0.43887\times10^{-2}y(k-1)y(k-2)^2+0.33268\times10^{-2}u(k-5)\\
&- 0.11251\times10^{-1}u(k-2)+0.46540\times10^{-2}u(k-3)-0.13899y(k-6)\\
&+ 0.34278y(k-4)-0.47384\times10^{-3}y(k-6)^3\\
&+ \Psi_{\xi}^{\tau}(k-1)\hat{\Theta}_{\xi}+\xi(k)
\end{aligned} \tag{61}$$

for which $\sigma_\xi^2 = 0.0082$. The bifurcation diagram and Poincaré section for this model are shown in figures 12a–b. The bifurcation quality index for this model is $J_b = 0.069$ indicating further improvement when compared to the model of equation (60).

## 7.4 Residual and parameter variances of the estimated models

The same procedure followed in § 4.3 to produce figure 6 was used to get a similar figure, see figure 13, for the model in equation (57). Comparison of such figures reveals that the filtering has improved the SNR but, most importantly, filtering has enabled the residual variance to converge. This suggests that the estimation algorithm performs better on the filtered data than on the raw data.

In the case of the double scroll attractor, the convergence properties of the residuals remained virtually unchanged. However, in this case a considerable SNR improvement was attained. Consequently the residual variance in the case of filtered data was much smaller and this enabelled improved parameter estimation.

A large noise variance will imply a large residual variance and some of the consequences of this are i) large uncertainty in the initial conditions of the model being estimated, and ii) large and sometimes 'diverging' residuals which are used to initialize the model during parameter estimation. These difficulties can also be analysed in the state space. A chaotic system is highly sensitive to small differences between two initial conditions in state space. This is referred to as *sensitive dependence on initial conditions* (SDIC). Consequently the basins of attraction in such a space are fractal.

A related aspect is that of *sensitive dependence on parameters* (SDP) as a consequence of which the basins of attraction in the parameter space are also fractal. Thus small variations in parameter values could result in a quantitatively and possibly qualitatively different dynamical properties of the corresponding attractors. Sensitive dependence on parameters has been known in the literature for some years [42] and seems to be a general property of chaotic systems [43].

It therefore becomes apparent that small variations in parameter values would be preferred to large variations because in the latter case the estimation algorithm would probably switch between qualitatively different dynamical regimes during parameter estimation in the attempt to find the best set of parameters and this would impair adequate parameter convergence.

It has been noted that "it is possible in principle to move from the deterministic to the random regime by varying $\epsilon$[11]" [32]. Although this has been stated for high dimensional systems, it is believed that a qualitatively similar situation may occur in parameter estimation of relatively low dimensional systems. In fact, for low dimensional systems it has been remarked that "small fluctuations (noise) in parameters ... will induce sizable changes in the dynamics " [44].

Furthermore, it is known that in the estimation of unstable systems the coefficient behaviour becomes very erratic [45]. This observation is relevant in the context of chaotic systems because such systems are on average *locally* unstable as indicated by the sign of the largest Lyapunov exponent.

[11] which is equivalent to $\sigma_\xi^2$ in this paper.

In order to investigate how the data filtered with the RF influences parameter variations, consider the following. The covariance of the parameter vector is defined as

$$\text{cov}\{\Theta\} \doteq \text{E}\{(\hat{\Theta} - \Theta)(\hat{\Theta} - \Theta)^{\text{T}}\} \tag{62}$$

where $\Theta$ is the true parameter vector and $\hat{\Theta}$ is the estimate. It is possible, however, to estimate $\text{cov}\{\Theta\}$ without knowing $\Theta$ which is indeed the case in most practical applications [16].

In order to compare the effects of filtering on the variance of the parameter vector, two things were done, namely i) the same model structure was assumed in each case, and ii) $\text{tr}[\text{cov}\{\hat{\Theta}_{yu}\}^{-0.5}]$ was used as a measure of the variation in the parameters of the deterministic part of the model, where $\text{tr}[\cdot]$ indicates the trace of a matrix.

Therefore the same structure of the filters were used to estimate parameters and the respective variances from the filtered data records. As can be seen from table 2, the parameter variance of the models[12] estimated from the filtered data is lower.

Table 2. Parameter variance measure for estimated models

| Data | System | Structure | Filter | Noise | $J^a$ |
|---|---|---|---|---|---|
| raw | DS[b] | eq. (54) | — | white | 0.430 |
| filtered | DS | eq. (54) | eq. (54) | white | 0.374 |
| raw | DU[c] | eq. (21) | — | white | 0.875 |
| filtered | DU | eq. (21) | eq. (21) | white | 0.488 |
| raw | DU | eq. (58) | — | correlated | 0.333 |
| filtered | DU | eq. (58) | eq. (58) | correlated | 0.321 |
| filtered | DU | eq. (58) | eqs. (58) & (59) | correlated | 0.318 |
| filtered | DU | eq. (58) | eqs. (58) & (59)[d] | correlated | 0.316 |

[a] $J = \text{tr}[\text{cov}\{\hat{\Theta}_{yu}\}^{-0.5}]$
[b] DS – Double-scroll attractor
[c] DU – Duffing-Ueda oscillator
[d] Without $\Psi_{\xi}^{\text{T}}(t-1)\hat{\Theta}_{\xi}$

# 8 Final remarks and conclusions

In all the examples above the models estimated from filtered sequences were chosen among a family of models. The criteria used to choose such models were dynamical invariants such as $\lambda_1$, $D_c$, the geometry of the reconstructed attractors, Poincaré sections and bifurcation diagrams.

The search was performed by varying the number of process terms in each model, $n_p$, and the maximum lags allowed, $n_y$ and $n_u$, which are the degrees of freedom of the models. It is believed that in some cases the subregion of the model structure space in which good models are found is limited and, in fact, is relatively small when compared with the entire space

[12] In order to enable comparison, these models have the same structure of the respective filter and are *not* the final models estimated in § 7.1–§ 7.3.

of all possible model structures of a given representation[6]. This is in accordance with the sometimes neglected fact that, once the nonlinearities in the data are well represented in the model structure, very little is to be gained by overparametrizing the model. On the contrary, overparametrization in nonlinear systems usually induces a number of spurious dynamical regimes.

By contrast, the structure of the filters is not so important because the filters are reset at each iteration. However, this does not mean to say that different filters will perform in exactly the same way and thereby yield similar results. In the examples above, two criteria were used as guidelines in the selection of the filters, namely i) the filters must be unbiased, and ii) it is desirable that the residual variance of the filters be relatively close to the noise variance. Apart from these criteria, the filter structures were chosen rather freely in order to illustrate a certain degree of liberty present in the design. However, it is believed that further improvements could be attained by optimizing in some way the choice of the filters.

Because the dynamics of the filter are learned from the noisy data, it becomes apparent that the noise variance poses limits on the performance of the filter. It is rather tempting to establish a link between this limit and the noise variance at which the estimated values of $\lambda_1$ begin to fall sharply, that is the 'knees' of the curves in figures 7a–b. In the case of the Duffing-Ueda oscillator, it is easy to verify that the variance of the noise remaining in the filtered data ($\approx 10^{-3}$) is below the critical value in figure 7b ($\approx 7 \times 10^{-3}$) as opposed to the original noise variance ($\approx 1.5 \times 10^{-2}$).

However, in the case of the double-scroll attractor, both the noise variances in the raw and filtered data are less than the critical value for $\lambda_1$, which seems to be around $10^{-1}$. It should be noted that increasing the maximum lag allowed from $n_y = 5$ to $n_y = 7$ enabled the identification of a model directly from the raw data for which $\sigma_\xi^2 = 0.037$, $\lambda_1 = 0.225$ and $D_c = 2.08 \pm 0.0219$. This model faithfully reproduces faithfully the attractor geometry.

It is interesting to note that the noise variance in this case is on the flat part of the curve in figure 7a. Could this be viewed as a reason why the identification was possible?

This example agrees with the results of Casdagli *et al.* who have defined a distortion matrix which describes the noise amplification and have argued that increasing the dimension of the reconstructed space (which is equivalent to increasing $n_y$ in the example above) reduces the distortion but also increases the estimation error. Consequently in a practical situation $n_y$ may be increased to a certain extent in order to estimate models from noisy data. If this is not possible or if it does not produce satisfactory results, filtering the data seems to be a viable alternative.

In all the examples using the Duffing-Ueda oscillator the increase in the maximum lags allowed did not enable the identification of good models from the raw data. Thus, in conclusion, it should be said that the critical point in the estimation of $\lambda_1$ could be viewed only as a gross indication of the critical variance beyond which accurate identification is precluded. Moreover, such a critical point seems to depend not only on the system, as it would be expected, but also on the number of degrees of freedom of the estimated models.

This paper has investigated the identification of NARMAX polynomial models for chaotic systems from noisy data. The main motivation for such an investigation was that the effects of noise on the quality of the identified models seems to be more deleterious when the system is chaotic.

Simple experiments performed in the estimation of the largest Lyapunov exponent, $\lambda_1$, from noisy data has also suggested that the estimates are more robust when $\lambda_1 < 0$. Moreover,

such exponent can be accurately estimated in some cases from noisy data as long as the noise variance is not larger than a certain critical value beyond which the accuracy of the estimates falls sharply. A similar phenomenon has been observed with some chaotic systems for which good models could be estimated from noisy data provided the noise variance did not exceed the critical point. It has been shown that a slight increase in the noise variance was enough to completely preclude the estimation of a valid model.

Such similarities between the estimation of $\lambda_1$ and of NARMAX polynomial models seem to suggest that some of the difficulties might be related to the limited capability of a chaotic model to accurately predict a given time series when the uncertainty (noise) in the initial conditions is too high. This seems to account for the fact that in both cases the estimation is heavily based on predictions. For low noise levels good estimates are possible because such predictions are made over relatively short periods of time. Although this period is maintained when the data are noisy, because the uncertainty in the data is much higher, even accurate short-term predictions seem to be affected. Such effects appear in the residuals and this is relevant because the residuals are responsible for a kind of feedback during parameter estimation. Consequently, if the residual sequence is inadequate the estimated model is likely to be inaccurate.

The aforementioned difficulties seem to be related to the *sensitive dependence on initial conditions* (SDIC) which is a well known feature of chaotic systems. A *dual* characteristic is the *sensitive dependence on parameters* (SDP) and although this latter feature has not been investigated in detail in this paper it is believed that some of the difficulties encountered in the identification of chaotic models from noisy data might also be related to the SDP. This is mainly because during estimation the parameters are adjusted depending on the quality of short-term prediction and the variance of such parameters is usually proportional to the noise variance.

A procedure for filtering chaotic data using NARMAX polynomial models has been suggested. The noise which is mostly unpredictable, is left out by 'predicting the predictable'. Moreover, the resetting effect inherent to the filter guarantees that the filtered data remains close to the raw data. This distinguishes the resetting filter from other prediction-based filtering techniques.

Some advantages of the suggested filtering procedure are i) no *a priori* knowledge is required, thus the filter is learned from the original data, ii) the procedure for estimating both the filter and the final model is the same, thus avoiding extra algorithms, iii) the filtering procedure is iterated only a few times (one or two) unlike other techniques which require four to twenty iterations [39], iv) as a consequence of the resetting effect, the filtering procedure is very robust with respect to several choices of the design parameters such as the sturcture of the filter, and v) parameter estimation is performed only once unlike piecewise linear method which estimate the parameters in each neighbourhood for each filtering iteration. On the other hand, the latter techniques achieve a higher noise reduction. However, the main objective of the filtering procedure suggested in this paper is to enable the identification of good models from the filtered data and not necessarily to attain a high increase in the SNR.

Examples using the autonomous double-scroll attractor and the Duffing-Ueda driven oscillation have been provided to illustrate the main points of the paper. Such examples show that the models estimated from filtered data sequences had lower residual and parameter variances. Such results seem to lend support to some of the conjectures made in the paper.

It has not been advocated that such conjectures account for all the problems found in the

identification of chaotic systems. Furthermore it is recognised that, for the sake of clarity, rather simple systems have been used in illustrating the main ideas. Nonetheless it is believed that some of the conjectures and procedures suggested in the paper are applicable to a wider class of chaotic models.

ACKNOWLEDGMENTS

LAA gratefully acknowledges financial support from the Brazilian Council of Scientific and Technological Development - CNPq, under grant 200597/90-6. SAB gratefully acknowledges that this work was supported by SERC under grant GR/H 35286.

# References

[1] S. A. Billings. Identification of nonlinear systems — a survey. *IEE Proceedigs Pt. D*, 127(6):272–285, 1980.

[2] I. J. Leontaritis and S. A. Billings. Input-output parametric models for nonlinear systems part I: deterministic nonlinear systems. *Int. J. Control*, 41(2):303–328, 1985.

[3] I. J. Leontaritis and S. A. Billings. Input-output parametric models for nonlinear systems part II: stochastic nonlinear systems. *Int. J. Control*, 41(2):329–344, 1985.

[4] S. A. Billings, M. J. Korenberg, and S. Chen. Identification of nonlinear output affine systems using an orthogonal least squares algorithm. *Int. J. Systems Sci.*, 19(8):1559–1568, 1988.

[5] H. Haber and H. Unbehauen. Structure identification of nonlinear dynamic systems — A survey on input/output approaches. *Automatica*, 26(4):651–677, 1990.

[6] L. A. Aguirre and S. A. Billings. Relationship between the structure and performance of identified nonlinear polynomial models. *(Submitted for publication)*, 1993.

[7] J.D. Farmer and J.J. Sidorowich. Exploiting chaos to predict the future and reduce noise. In Y.C. Lee, editor, *Evolution, Learning and Cognition*. World Scientific, Singapore, 1988.

[8] M. Casdagli. Nonlinear prediction of chaotic time series. *Physica D*, 35:335–356, 1989.

[9] P. Grassberger, J. Schreiber, and C. Schaffrath. Nonlinear time sequence analysis. *Int. J. Bif. Chaos*, 1(3):521–547, 1991.

[10] B. R. Haynes and S. A. Billings. Global analysis and model validation in nonlinear system identification. *J. of Nonlinear Dynamics*, (in press), 1993.

[11] J. P Crutchfield and B. S. McNamara. Equations of motion from a data series. *Complex Systems*, 1:417–452, 1987.

[12] R. Brown, P. Bryant, and H. D. I. Abarbanel. Computing the Llyapunov spectrum of a dynamical system from an observed time series. *Phys. Rev. A*, 43(6):2787–2806, 1991.

[13] S. Chen and S. A. Billings. Representations of nonlinear systems: the NARMAX model. *Int. J. Control*, 49(3):1013–1032, 1989.

[14] S. Chen, S. A. Billings, and W. Luo. Orthogonal least squares methods and their application to nonlinear system identification. *Int. J. Control*, 50(5):1873–1896, 1989.

[15] M. J. Korenberg and L. D. Paarmann. Orthogonal approaches to time-series analysis and system identification. *IEEE Signal Processing Magazine*, 8(3):29–43, 1991.

[16] M. J. Korenberg, S. A. Billings, Y. P. Liu, and P. J. McIlroy. Orthogonal parameter estimation algorithm for nonlinear stochastic systems. *Int. J. Control*, 48(1):193–210, 1988.

[17] T. Söderström and P. Stoica. *System Identification.* Prentice Hall, London, 1989.

[18] S. A. Billings and W. S. F. Voon. Correlation based model validity tests for non-linear models. *Int. J. Control*, 44(1):235–244, 1986.

[19] S. A. Billings and Q. H. Tao. Model validation tests for nonlinear signal processing applications. *Int. J. Control*, 54:157–194, 1991.

[20] H. D. I. Abarbanel, R. Brown, and J. B. Kadtke. Prediction in chaotic nonlinear systems: Methods for time series with broadband Fourier spectra. *Phys. Rev. A*, 41(4):1782–1807, 1990.

[21] J. C. Principe, A. Rathie, and J. M Kuo. Prediction of chaotic time series with neural networks and the issue of dynamic modeling. *Int. J. Bif. Chaos*, 2(4):989–996, 1992.

[22] L. A. Aguirre and S. A. Billings. Validating identified nonlinear models with chaotic dynamics. *(Submitted for publication)*, 1993.

[23] T. Matsumoto, L.O. Chua, and K. Tokumasu. Double scroll via two-transistor circuit. *IEEE Trans. Circuits Syst.*, 33(8):828–835, 1986.

[24] L.O. Chua, M. Komuro, and T. Matsumoto. The double scroll family. *IEEE Trans. Circuits Syst.*, 33(11):1072–1118, 1986.

[25] T. Matsumoto, L.O. Chua, and M. Komuro. The double scroll. *IEEE Trans. Circuits Syst.*, 32(8):698–718, 1985.

[26] Y. Ueda. Random phenomena resulting from nonlinearity in the system described by Duffing's equation. *Int. J. Non-Linear Mech.*, 20(5/6):481–491, 1985.

[27] F. C. Moon. *Chaotic Vibrations - an introduction for applied scientists and engineers.* John Willey and Sons, New York, 1987.

[28] Y. Ueda. Steady motions exhibited by Duffing's equation: A picture book of regular and chaotic motions. In P. J. Holmes, editor, *New approaches to nonlinear problems in dynamics*, pages 311–322. SIAM, 1980.

[29] R. Shaw. Strange attractors, chaotic behavior and information flow. *Z. Naturforsch*, 36a:80–112, 1981.

[30] A. Wolf, J. B. Swift, H. L. Swinney, and J. A. Vastano. Determining lyapunov exponents from a time series. *Physica D*, 16:285–317, 1985.

[31] T. S. Parker and L. O. Chua. *Practical numerical algorithms for chaotic systems.* Springer Verlag, Berlin, 1989.

[32] M. Casdagli, S. Eubank, J. D. Farmer, and J. Gibson. State space reconstruction in the presence of noise. *Physica D*, 51:52–98, 1991.

[33] R. Badii, G. Broggi, B. Derighetti, M. Ravani, S. Ciliberto, A. Politi, and M. A. Rubio. Dimension increase in filtered chaotic signals. *Phys. Rev. Lett.*, 60(11):979–982, 1988.

[34] F. Mitschke. Acausal filters for chaotic signals. *Phys. Rev. A*, 41(2):1169–1171, 1990.

[35] A. Chennaoui, K. Pawelzik, W. Liebert, G. H. Schuster, and G. Pfister. Attractor reconstruction from filtered chaotic time series. *Phys. Rev. A*, 41(8):4151–4159, 1990.

[36] J. D. Farmer and J. J. Sidorowich. Optimal shadowing and noise reduction. *Physica D*, 47:373–392, 1991.

[37] E. J. Kostelich and J. A. Yorke. Noise reduction in dynamical systems. *Phys. Rev. A*, 38(3):1649–1652, 1988.

[38] S. A. Billings and W. S. F. Voon. Least squares parameter estimation algorithms for nonlinear systems. *Int. J. Systems Sci.*, 15(6):601–615, 1984.

[39] T. Schreiber and P. Grassberger. A simple noise-reduction method for real data. *Phys. Lett.*, 160 A(5):411–418, 1991.

[40] S. Chen and S. A. Billings. Modelling and analysis of non-linear time series. *Int. J. Control*, 50(6):2151–2171, 1989.

[41] H. Akaike. A new look at the statistical model identification. *IEEE Trans. Automat. Contr.*, 19(6):716–723, 1974.

[42] J. D. Farmer. Sensitive dependence on parameters in nonlinear dynamics. *Phys. Rev. Lett.*, 55(4):351–354, 1985.

[43] R. Brown, L. Chua, and B. Popp. Is sensitive dependence on initial conditions nature's sensory device? *Int. J. Bif. Chaos*, 2(1):193–199, 1992.

[44] J. A. C. Gallas, C. Grebogi, and J. Yorke. Vertices in parameter space: Double crises which destroy chaotic attractors. *Phys. Rev. Lett.*, 71(9):1359–1362, 1993.

[45] V. J. Mathews. Adaptive polynomial filters. *IEEE Signal Processing Magazine*, 8(3):10–26, 1991.

# Captions

Figure 1. Double-scroll attractor reconstructed from the $z$ component. Data sampled at $T_s = 0.15$ and $T_p = 2$.

Figure 2. (a) bifurcation diagram and (b) Poincaré section of the attractor at $A = 5.7$ for the Duffing-Ueda oscillator.

Figure 3. Reconstructed attractor for the estimated model of equation (20).

Figure 4. (a) bifurcation diagram and (b) Poincaré section of the attractor at $A = 5.7$ for the estimated model of equation (21).

figure 5. Schematic diagram of the prediction error estimation algorithm. Note that whereas the deterministic part of the model, $\Psi_{yu}^{T}(t-1)\hat{\Theta}_{yu}$, is totally reset with the measured data $y(t)$ and $u(t)$, the stochastic parts, $\Psi_{yu\xi}^{T}(t-1)\hat{\Theta}_{yu\xi}$ and $\Psi_{\xi}^{T}(t-1)\hat{\Theta}_{\xi}$, are subject to feedback via the residuals $\xi(t)$.

Figure 6. Asymptotic behaviour of the residuals.

Figure 7. Estimated values of the largest Lyapunov exponent, $\lambda_1$, for increasing values of noise variance (a) the double-scroll attractor (b) the Duffing-Ueda oscillator in a chaotic regime for $A = 11$ indicated (o), and in a periodic regime for $A = 4.5$ indicated by (*).

Figure 8. Correlation tests for the RF in equation (54) (a) $\Phi_{\xi\xi}(\tau)$, (b) $\Phi_{\xi\xi^2}(\tau)$ and (c) $\Phi_{\xi^2\xi^2}(\tau)$. Note that, because the functions are confined to the confidence bands, it can be considered that no significant dynamics were left in the residuals.

Figure 9. Double-scroll attractors reconstructed from (a) the raw data, (b) the filtered data, and (c) a time series generated by the estimated model of equation (55).

Figure 10. (a) bifurcation diagram with $J_b = 0.168$, and (b) Poincaré section of the attractor at $A = 5.7$ for the estimated model of equation (57).

Figure 11. (a) bifurcation diagram with $J_b = 0.161$, and (b) Poincaré section of the attractor at $A = 5.7$ for the estimated model of equation (60).

Figure 12. (a) bifurcation diagram with $J_b = 0.069$, and (b) Poincaré section of the attractor at $A = 5.7$ for the estimated model of equation (61).

Figure 13. Asymptotic behaviour of the residuals for models estimated from filtered data (compare with figure 6).

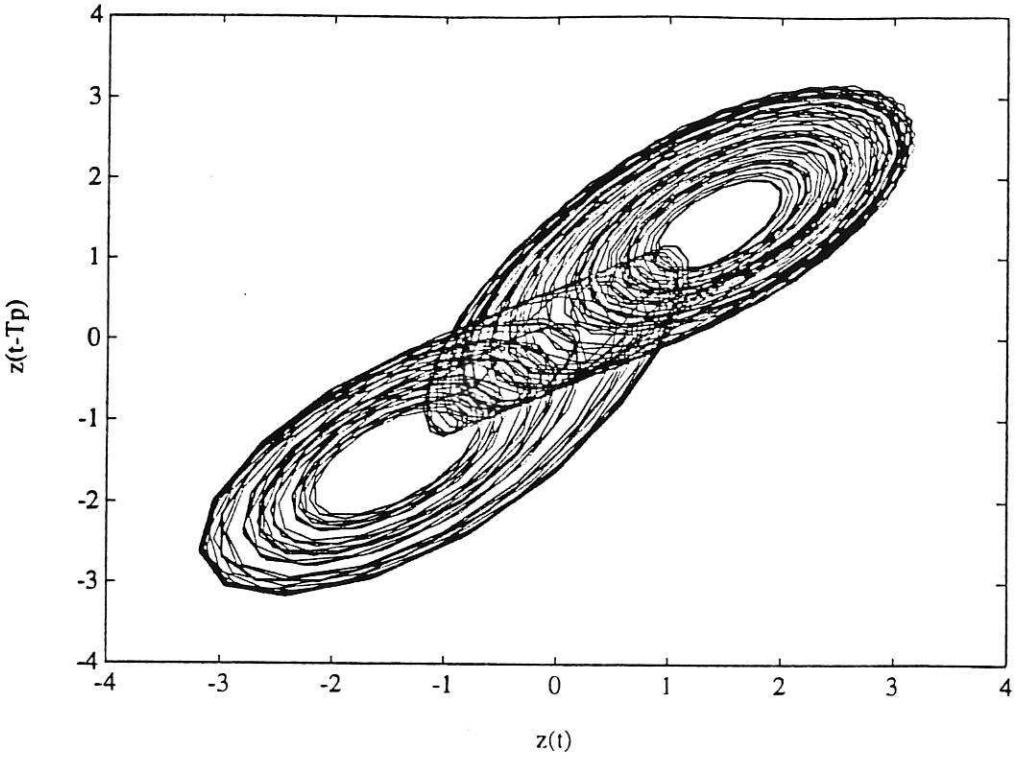
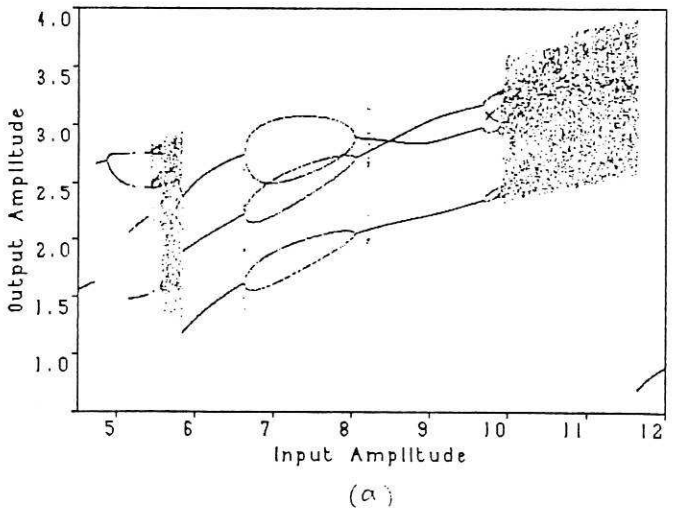
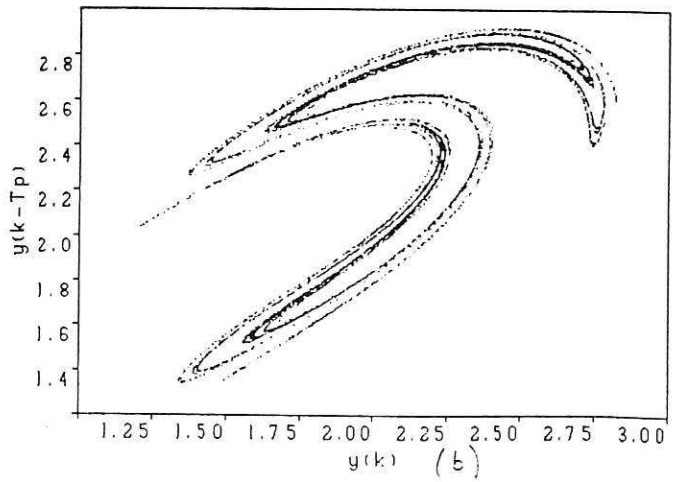

z(t-Tp)
z(t)
4
3
2
1
0
-1
-2
-3
-4
-4
-3
-2
-1
0
1
2
3
4

Fig. 1


Output Amplitude
Input Amplitude
4.0
3.5
3.0
2.5
2.0
1.5
1.0
5
6
7
8
9
10
11
12

(a)


y(k-Tp)
y(k)
2.8
2.6
2.4
2.2
2.0
1.8
1.6
1.4
1.25
1.50
1.75
2.00
2.25
2.50
2.75
3.00

(b)

Fig. 2

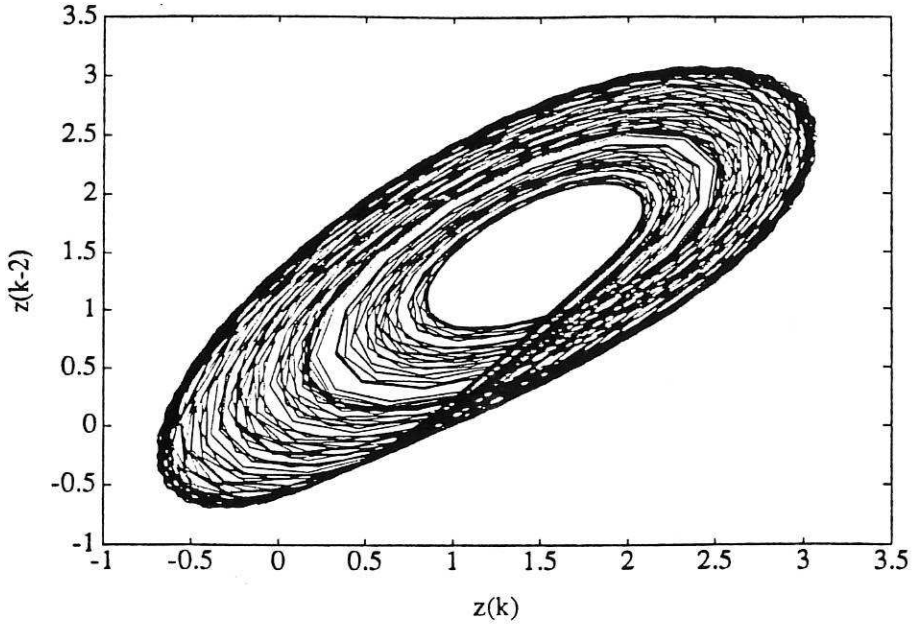
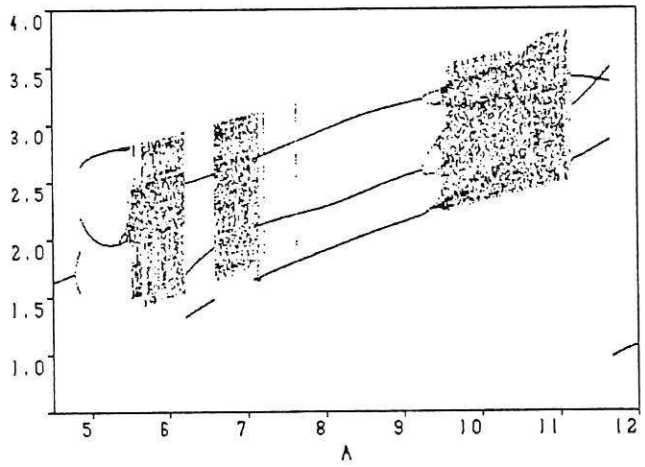
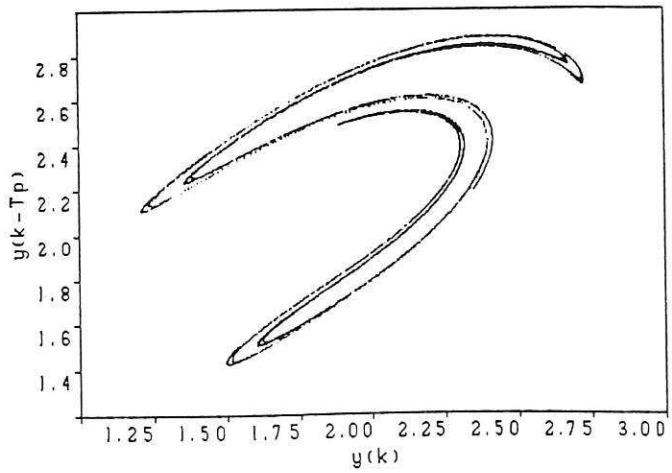

3.5
3
2.5
2
1.5
1
0.5
0
-0.5
-1
z(k-2)
-1
-0.5
0
0.5
1
1.5
2
2.5
3
3.5
z(k)

Fig. 3


4.0
3.5
3.0
2.5
2.0
1.5
1.0
5
6
7
8
9
10
11
12
λ

(a)


2.8
2.6
2.4
2.2
2.0
1.8
1.6
1.4
y(k-Tp)
1.25
1.50
1.75
2.00
2.25
2.50
2.75
3.00
y(k)

(b)

Fig. 4

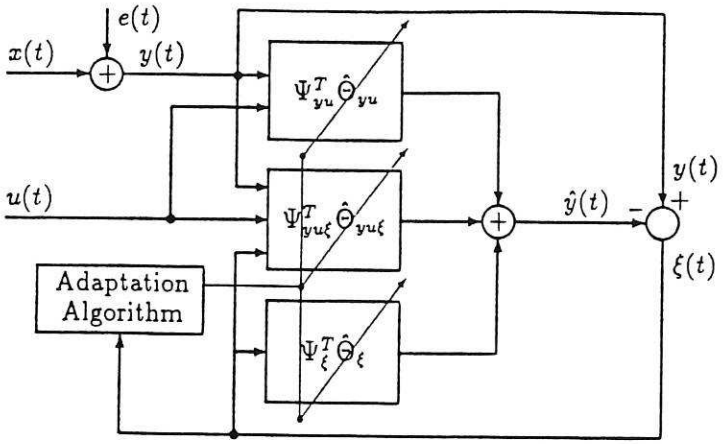



Fig. 5

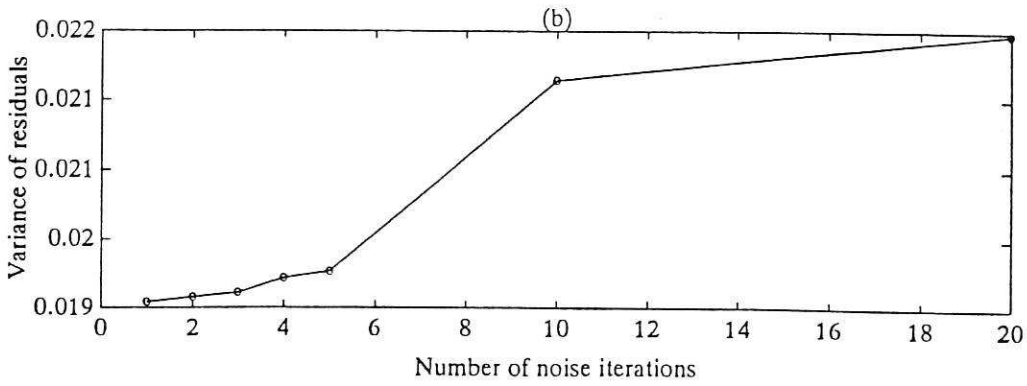



Fig. 6

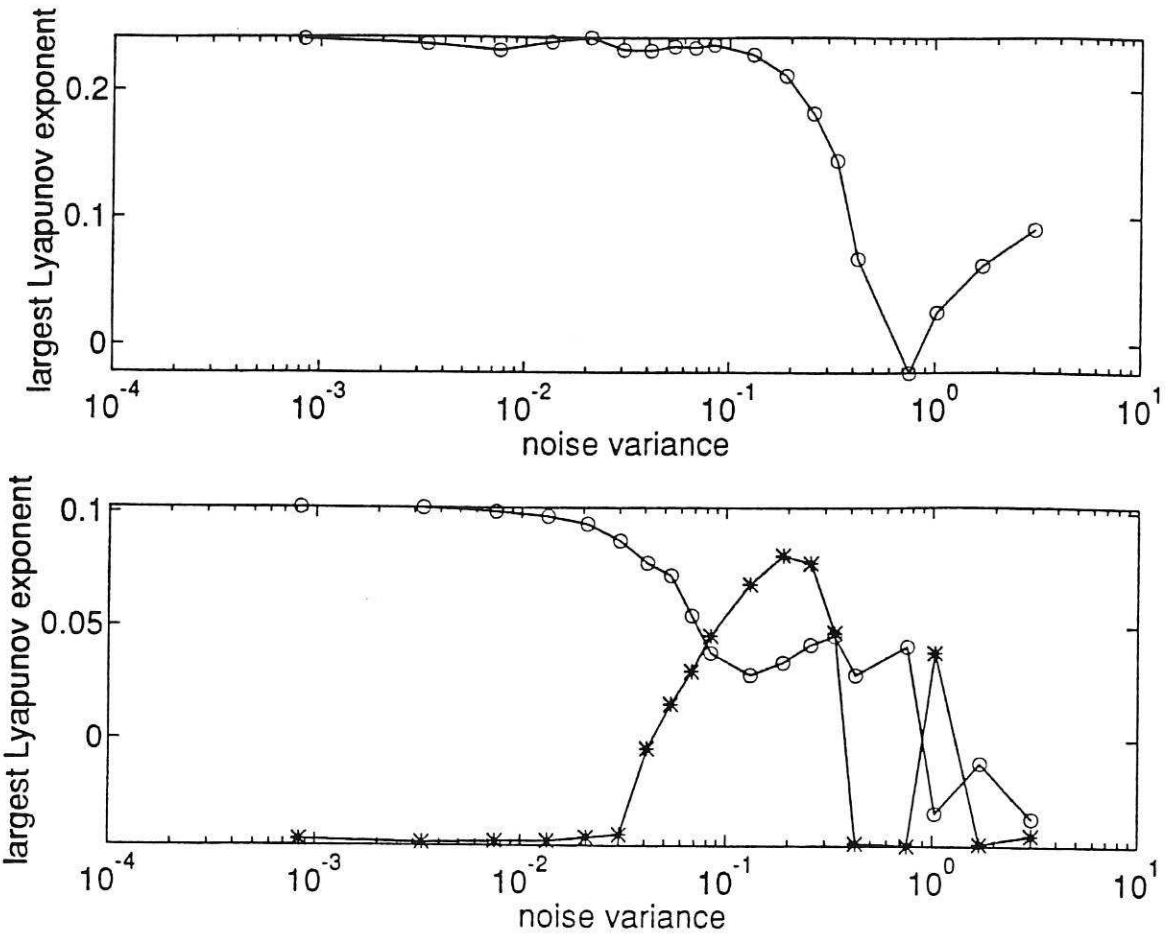

largest Lyapunov exponent
0.2
0.1
0
$10^{-4}$
$10^{-3}$
$10^{-2}$
$10^{-1}$
$10^{0}$
$10^{1}$
noise variance
largest Lyapunov exponent
0.1
0.05
0
$10^{-4}$
$10^{-3}$
$10^{-2}$
$10^{-1}$
$10^{0}$
$10^{1}$
noise variance

Fig. 7

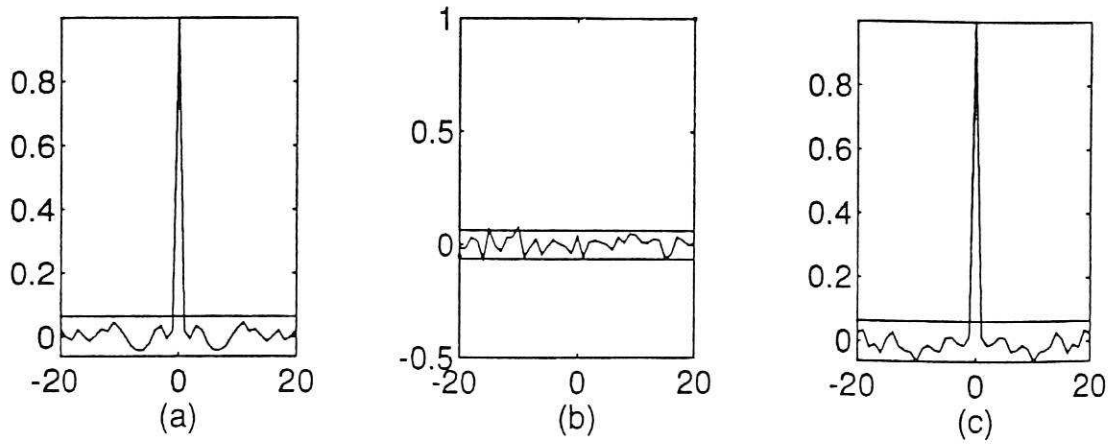
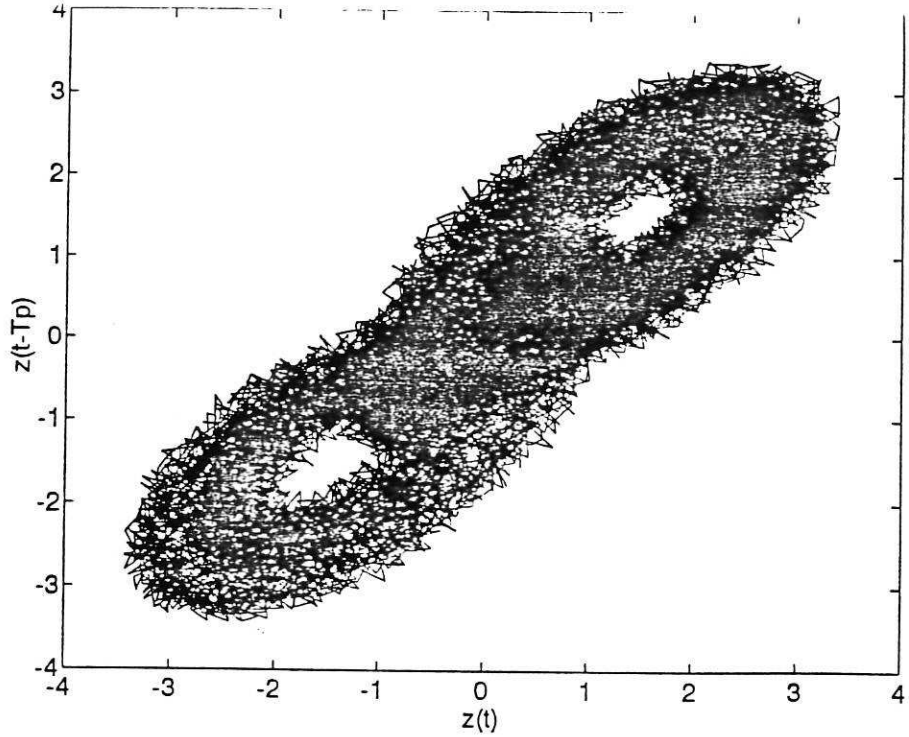
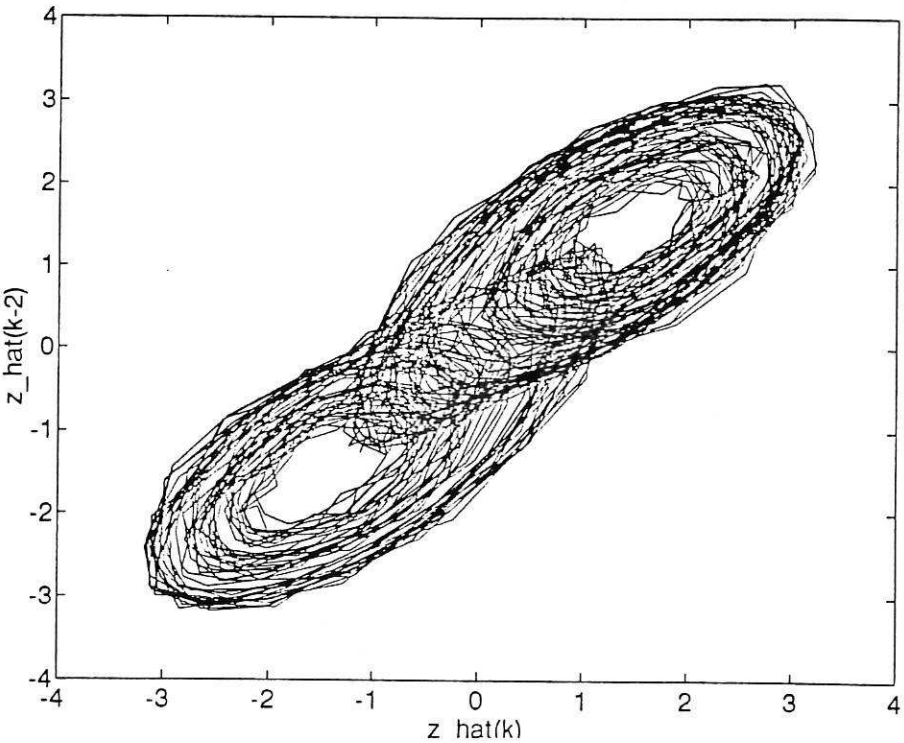

0.8
0.6
0.4
0.2
0
-20
0
20
(a)
1
0.5
0
-0.5
-20
0
20
(b)
0.8
0.6
0.4
0.2
0
-20
0
20
(c)

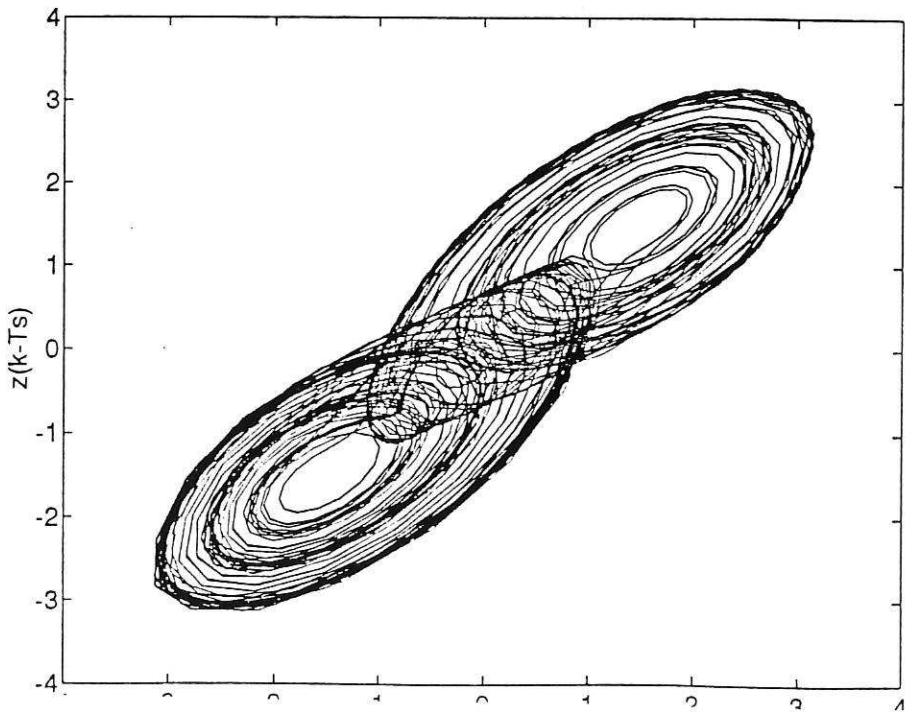
Fig. 8




(a)

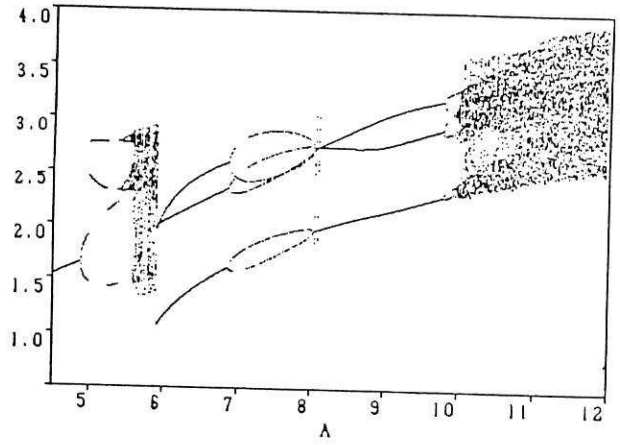



(b)

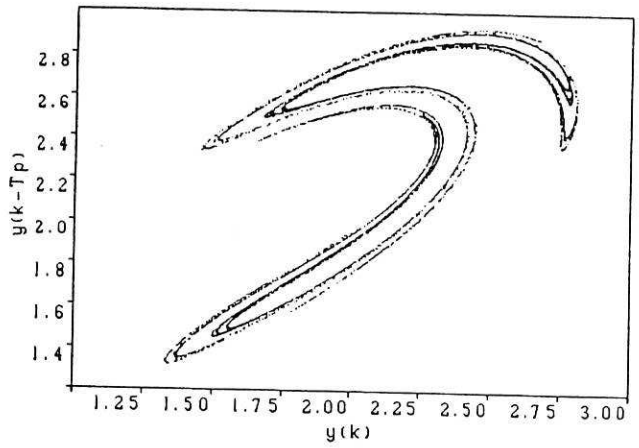



(c)

Fig. 9

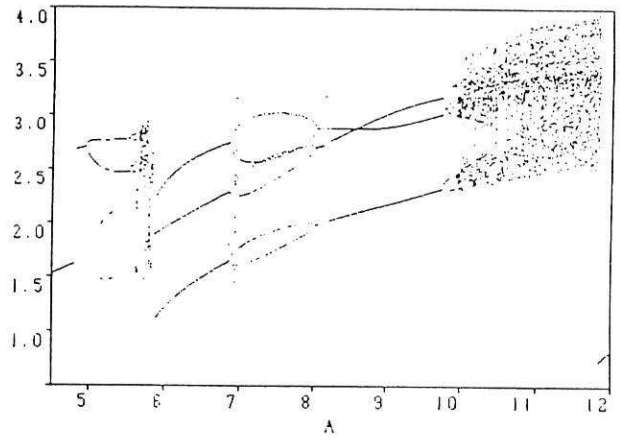



(a)

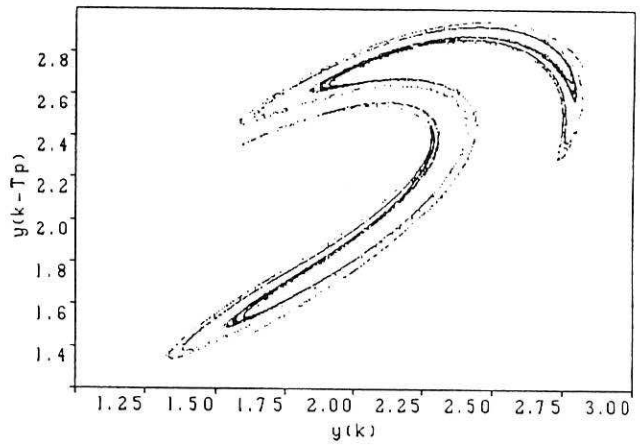



(b)

Fig. 10




(a)




(b)

Fig. 11