# A Nonlinear Smoothing Algorithm
# For Chaotic and Non-chaotic Time Series

S A Billings and K L Lee
Department of Automatic Control and Systems Engineering
University of Sheffield
Mappin Street, Sheffield S1 3JD
United Kingdom

Research Report No. 816

February 2002

University of Sheffield

# A Nonlinear Smoothing Algorithm for Chaotic and Non-chaotic Time Series

S.A. Billings, K.L. Lee
Department of Automatic Control and Systems Engineering
University of Sheffield
Sheffield S1 3JD, UK

**Abstract**: A new NARMA based smoothing algorithm is introduced for chaotic and non-chaotic time series. The new algorithm employs a cross-validation method to determine the smoother structure, requires very little user interaction, and can be combined with wavelet thresholding to further enhance the noise reduction. Numerical examples are included to illustrate the application of the new algorithm.

## 1. Introduction

In the last decade, dynamical invariants such as the Lyapunov exponent and the correlation dimension have been developed to quantify chaos. But almost all the algorithms that are used to compute these invariants are sensitive to noise contamination and several can be rendered useless by small amounts of noise (Schreiber and Kantz 1996). Unfortunately, in many practical situations, regardless of how sophisticated the experimental setup is, noise will always be present in the measurements. Pre-processing the data to reduce the noise without distorting the dynamics of the underlying signal is therefore highly desirable.

Conventional methods of noise reduction such as linear low pass filtering do not work well for chaotic data since the signal and the noise often have overlapping bandwidths. Several nonlinear noise reduction methods have therefore been proposed for chaotic systems including Kostelich and Yorke (1990), Grassberger et al. (1993), Davies (1994, 1999), Bhowal and Roy (1999), and Shin et al. (1999). Some of these methods have been applied on experimental data and promising results have been obtained, see for example Kantz et al. (1993). However most of the noise reduction algorithms are based on local methods and require the choice of certain parameters to be set by the user. For example, the size of the neighbourhood of the local model, the number of noise reduction iterations to be performed and the embedding dimension to use. Detailed discussions on the issues of implementation and the choice of parameters in the local methods can be found in Hegger et al. (1999). However the task of properly selecting the values of these parameters is not trivial in most practical situations, and if the noise reduction methods are applied naively over smoothing may occur as reported by Mees and Judd (1993). Alternative noise reduction methods which require less user interaction are therefore needed.

Recently a relatively simple noise reduction method based on a global nonlinear smoother called the NARMAX (Nonlinear AutoRegressive Moving Average with eXogenous inputs) smoother was proposed by Aguirre et al. (1996) for input/output systems. The algorithm attempts to reduce noise by fitting NARMAX models with positive and negative lagged variables to compute the predictable part of the signal and discard the unpredictable part. In this study a NARMA smoother for time series is further investigated and it is shown that the smoothing algorithm introduced by Aguirre et al. (1996) can be improved. In Aguirre's smoothing algorithm the user has to define the smoother structure. No a priori knowledge of the underlying dynamics of the time series is assumed and hence the smoother structure may be defined incorrectly by the user and this can result in poor noise reduction. A new smoothing algorithm using a cross-validation approach to determine the smoother structure is introduced to overcome these limitations. The new algorithm requires very little user interaction and can be applied directly to smooth data corrupted by correlated noise. Numerical examples are given to show the application of the new smoothing algorithm. It is also shown that the new smoothing algorithm can be combined with a wavelet thresholding method to further enhance noise reduction.

The paper is organised as follows. Section 2 briefly introduces basic concepts and algorithms. The NARMA smoother is described in Section 3. The new smoothing algorithm is presented in Section 4. The application of the new NARMA smoothing algorithm is illustrated in Section 5. Wavelet thresholding is employed to reduce the high frequency noise in the signals, and simulation studies of several different noise reduction methods are conducted in Section 6. Conclusions are given in Section 7.

## 2. Preliminaries

### 2.1 Identification

Consider the Nonlinear AutoRegressive Moving Average model with eXogenous inputs (NARMAX) (Leontaritis and Billings 1985a, b)

$$y_t = f\left(y_{t-1},\ldots,y_{t-n_y},u_{t-1},\ldots,u_{t-n_u},e_{t-1},\ldots,e_{t-n_e}\right)+e_t \tag{1}$$

where $y_t$ is the output signal, $u_t$ is the input signal, $e_t$ is an unobserved unpredictable zero mean noise sequence with variance $\sigma_e^2$, $f(\cdot)$ is some nonlinear function and $n_y, n_u$ and $n_e$ are the maximum output, input and noise lags respectively. In this study only the time series case will be considered so the inputs are assumed to be zero to give a NARMA model

$$y_t = f\left(y_{t-1},\ldots,y_{t-n_y},e_{t-1},\ldots,e_{t-n_e}\right)+e_t \tag{2}$$

However the results and findings are still applicable to the general NARMAX case. Furthermore, in most parts of this study the nonlinear function $f(\cdot)$ will be assumed to be a polynomial function of degree $l$. Other choices of $f(\cdot)$ are available including rational functions, wavelet or neural network expansions. Equation (2) can be expressed as

$$y_t = \Psi^T_{y_{t-1}}\theta_y + \Psi^T_{(ye)_{t-1}}\theta_{ye} + \Psi^T_{e_{t-1}}\theta_e + e_t \tag{3}$$

where the superscript $T$ denotes transposition and $\Psi^T_{y_{t-1}}$ includes all the output terms of all possible combinations up to degree $l$ lagged in time from $(t-1)$. Similarly for the terms $\Psi^T_{(ye)_{t-1}}$ and $\Psi^T_{e_{t-1}}$. The parameters associated with these terms are $\theta_y,\theta_{ye}$ and $\theta_e$. Equation (3) can be written as

$$y_t = \Psi^T_{t-1}\theta + \xi_t \tag{4}$$

where $\theta$ represents the unknown parameters and the unobserved noise $e_t$ is replaced by the residuals $\xi_t$. The cost function

$$J = \left\|y_t - \Psi^T_{t-1}\theta\right\| \tag{5}$$

where $\|\cdot\|$ denotes the Euclidean norm, is minimised to estimate the unknown parameters $\theta$. The forward regression orthogonal least squares algorithm (Billings et al. 1988, Korenberg et al. 1988, Chen et al. 1989) can be employed to select the important terms from a set of candidate terms, which can include past and future lagged values, and to estimate the associated parameters.

## 2.2 Model Validation

Once a model has been identified the model should be validated. Billings and Zhu (1994) developed the statistical validation methods given below which check that the residuals are unpredictable.

$$\rho_{\xi\xi}\left(\tau_\rho\right) = \frac{\sum_{t=1}^{N-\tau_\rho}\xi_t\xi_{t+\tau_\rho}}{\sum_{t=1}^{N}\xi_t^2} = \delta\left(\tau_\rho\right) \tag{6}$$

$$\rho_{[y\xi][\xi\xi]}\left(\tau_\rho\right) = \frac{\sum_{t=1}^{N-\tau_\rho}[y_t\xi_t - \overline{y_t\xi_t}][\xi^2_{t+\tau_\rho} - \overline{\xi^2_t}]}{\sqrt{\sum_{t=1}^{N}[y_t\xi_t - \overline{y_t\xi_t}]^2}\sqrt{\sum_{t=1}^{N}[\xi_t^2 - \overline{\xi_t^2}]^2}} = k_\rho\delta\left(\tau_\rho\right) \tag{7}$$

where $N$ is the length of the time series, $k_p$ is a constant, $\delta$ denotes the Kronecker delta and overbar denotes time average. For sufficiently long data sets the 95% confidence intervals are approximately $\pm 1.96/\sqrt{N}$.

## 3. Global NARMA Smoother/Filter

The structure of the NARMA filter takes the form of eqn (2) with the unobserved unpredictable noise $e_t$ replaced by residuals $\xi_t$, whereas the smoother includes both past and future output values (Aguirre et al. 1996) as

$$y_t = f\left(y_{t+n_y}, \ldots, y_{t+1}, y_{t-1}, \ldots, y_{t-n_y}, \xi_{t-1}, \ldots, \xi_{t-n_\xi}\right) + \xi_t \tag{8}$$

The iterative smoothing procedure for noise reduction will be discussed below. The filtering procedure is exactly the same but only past information is allowed in $f(\cdot)$. Let $y_t^{(0)}$ denote the original noisy output data, then at the first iteration a model is fitted to the noisy output data to give

$$y_t^{(0)} = f^{(0)}\left(y_{t\pm}^{(0)}, \xi_{t-}^{(0)}\right) + \xi_t^{(0)} \tag{9}$$

where $(\cdot)_{t\pm}$ indicates that both future and past information is used and $(\cdot)_{t-}$ indicates that only past information is used. The output data set is then smoothed by retaining just $f^{(0)}\left(y_{t\pm}^{(0)}, \xi_{t-}^{(0)}\right)$ on the r.h.s. of eqn (9) to obtain the first smoothed data set as

$$y_t^{(1)} = f^{(0)}\left(y_{t\pm}^{(0)}, \xi_{t-}^{(0)}\right) \tag{10}$$

The second iteration begins by fitting a new smoother to the data set $y_t^{(1)}$. Therefore

$$y_t^{(1)} = f^{(1)}\left(y_{t\pm}^{(1)}, \xi_{t-}^{(1)}\right) + \xi_t^{(1)} \tag{11}$$

and the second smoothed data set is obtained as

$$y_t^{(2)} = f^{(1)}\left(y_{t\pm}^{(1)}, \xi_{t-}^{(1)}\right) \tag{12}$$

and this iterative procedure is continued.

The entire smoothing procedure is required to be statistically sound and unbiased. To check this condition, the overall discarded residuals $\xi_t^{(T)}$ which can be computed from

$$\xi_t^{(T)} = y_t^{(0)} - y_t^{(n_p)} \tag{13}$$

where $n_p$ is the number of smoothing iterations performed, must pass the correlation tests of eqns (6) & (7). The main objective of this approach is therefore to discard the unpredictable component of the noise sequence and this can be tested by applying the model validity tests of eqns (6) & (7) to the sequence $\xi_t^{(T)}$.

The NARMA smoothing procedure proposed in Aguirre et al. (1996) can be summarised as follows.

i) Set initial values of $l, n_y, n_\xi$ and the number of important terms $n_m$ to define the set of terms to search using the forward regression orthogonal least squares algorithm (Chen et al. 1989). Set $i = 0$.

ii) Use the forward regression orthogonal least squares algorithm to identify the smoother $f^{(i)}(\cdot)$ from the data set $y_t^{(i)}$ by selecting the best $n_m$ important terms from the candidate terms of the polynomial expansion and estimating the associated parameters.

iii) Use the identified smoother $f^{(i)}(\cdot)$ to obtain the smoothed data set $y_t^{(i+1)}$.

iv) If the signal to noise ratio of $y_t^{(i+1)}$ is satisfactory, proceed to step (v), otherwise increase $i$ by one and repeat steps (ii) to (iv) as many times as necessary.

v) Use the correlation tests to verify that the overall discarded residuals $\xi_t^{(T)}$ are unpredictable. If significant correlation is detected, return to step (i) and alter the smoother structure.

In Aguirre et al. (1996) the verification of the signal to noise ratio in step (iv) was made indirectly. This was accomplished by identifying models from the smoothed data set and if the identified model could successfully reproduce the true dynamical invariants of the time series, then step (v) was followed. Unfortunately in most practical situations the true dynamical invariants will usually be unknown and step (iv) has been replaced as follows.

(iv) Increase $i$ by one and repeat steps (ii) to (iv) until $i = n_p$, where $n_p$ is the number of smoothing iterations to be performed. The value of $n_p$ is to be set by the user.

Another unclear point in the smoothing algorithm is that no guidelines are provided on how to choose or alter the smoother structure. This was presumably because the authors thought that the smoother structure is unimportant as long as the smoother remains unbiased. Therefore the user has the freedom to choose any smoother structure as long as the overall residuals pass the correlation tests. This freedom may not be desirable because with a poorly chosen smoother structure, the amount of noise removed by the algorithm may be minimal. A simple non-chaotic time series will be used to illustrate this point in the section below.

### 3.1 Simulation of Duffing's Equation

Duffing's equation was chosen because it has been used as a benchmark in many studies. Consider Duffing's equation

$$\ddot{x} + \gamma\dot{x} - \alpha(x - x^3) = F\cos(wt) \tag{14}$$

with $w = 1, \gamma = 0.168, \alpha = 0.5$ and $F = 0.05$. The equation was integrated using a fourth order Runge-Kutta method with a fixed step size of 0.01s. A total of 2000 data points were collected with a sampling rate of 0.1s and this served as the noise-free time series data $x_t$. This time series exhibits non-chaotic limit cycle behaviour. Zero mean uncorrelated unpredictable noise $e_t$ was then added to $x_t$ to give

$$y_t = x_t + e_t \tag{15}$$

with a signal-to-noise ratio of 20dB. The results in Aguirre et al. (1996) suggest that filtering would achieve a higher noise reduction than smoothing for non-chaotic data and hence a polynomial NARMA filter was initially used to filter the output. The maximum degree of the polynomial function was chosen to be 4 and the maximum output and noise lags were chosen as 10 and 20 respectively. To limit the size of the candidate terms in the polynomial expansion, only linear noise terms were considered. By specifying different numbers of important terms $n_m$ to be selected by the forward regression orthogonal least squares algorithm (Chen et al. 1989), several filters were identified where the residuals passed the correlation tests. For example the following two filters can successfully filter the output and pass the correlation tests as illustrated in Figures 1 & 2. Filter 1 has only 2 output terms and 3 linear noise terms whereas filter 2 contains 10 output terms and 3 linear noise terms.

$$\text{Filter 1: } f_1(y_{t-}, \xi_{t-}) = 1.0635 y_{t-1} - 0.0643 y_{t-10}^2 - 0.7099 \xi_{t-1} + 0.0649 \xi_{t-4} + 0.0586 \xi_{t-10} \tag{16}$$

$$\begin{aligned}\text{Filter 2: } f_2(y_{t-}, \xi_{t-}) = {} & 1.2180 y_{t-1} + 0.0053 y_{t-10}^2 - 0.0053 y_{t-2} - 0.0738 y_{t-3} - 0.0097 y_{t-9} \\ & + 0.0643 y_{t-4} - 0.0015 y_{t-1}^2 y_{t-2}^2 - 0.0366 y_{t-1} y_{t-5} - 0.0761 y_{t-8} - 0.0852 y_{t-6} \\ & - 1.0749 \xi_{t-1} - 0.0881 \xi_{t-19} - 0.0766 \xi_{t-16}\end{aligned} \tag{17}$$

The phase portrait plots of the filtered data for the two filtered time series are shown in Figure 3. A measure of the effectiveness of the smoothing/filtering algorithm can be obtained by computing

$$D = \sqrt{\frac{\sum_{t=1}^{N} \left(y_t^{(0)} - x_t\right)^2}{\sum_{t=1}^{N} \left(y_t^{(n_p)} - x_t\right)^2}} \tag{18}$$

A higher value of the index $D$ means that more noise has been removed from the noisy output. Clearly this index $D$ is of theoretical interest only since, in practice, the noise-free data $x_t$ will not be available. The values of $D$ were 1.4934 and 2.6952 for filter 1 and 2 respectively. The performance of filter 1 does not improve by increasing the number of filtering iterations. This indicates that filter 2 should be chosen to filter the output. This example shows that in some cases the smoother/filter structure (in this case the value of $n_m$ defined by the user) may have a great influence on the amount of noise removed by the smoothing/filtering algorithm. Without a priori knowledge of the dynamics of the time series, filter 1 may have been chosen to filter the output when improved results could be obtained with another choice of filter, in this case filter 2 above. This suggests that it is important to investigate if a better method of determining the smoother structure can be developed.

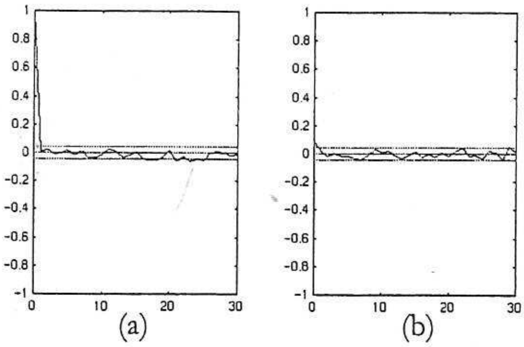


Figure 1 Correlation plots of the overall discarded residual $\xi_t^{(T)}$, (a) $\rho_{\xi\xi}(\tau_\rho)$ and (b) $\rho_{[y\xi][\xi\xi]}(\tau_\rho)$ for filter 1 in Section 3.1.

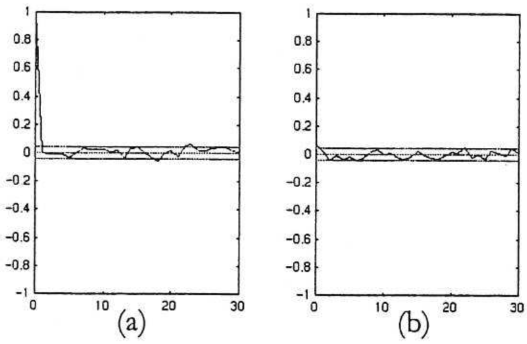


Figure 2 Correlation plots of the overall discarded residual $\xi_t^{(T)}$, (a) $\rho_{\xi\xi}(\tau_\rho)$ and (b) $\rho_{[y\xi][\xi\xi]}(\tau_\rho)$ for filter 2 in Section 3.1.

# 4. A new NARMA Smoothing Algorithm

## 4.1 A Cross-Validation Approach to Determine the Smoother Structure

The NARMA smoothing algorithm attempts to smooth the data by determining the predictable components in the noisy output so that the unpredictable noise can be isolated and removed. However, by employing a high degree of nonlinearity and using a large number of terms in the model, the noisy signal can become highly predictable because this can produce overfitting with all the consequent problems that this introduces. A common method to prevent overfitting to the noise is to use cross-validation, where the data set is divided into training and testing data sets. A model is

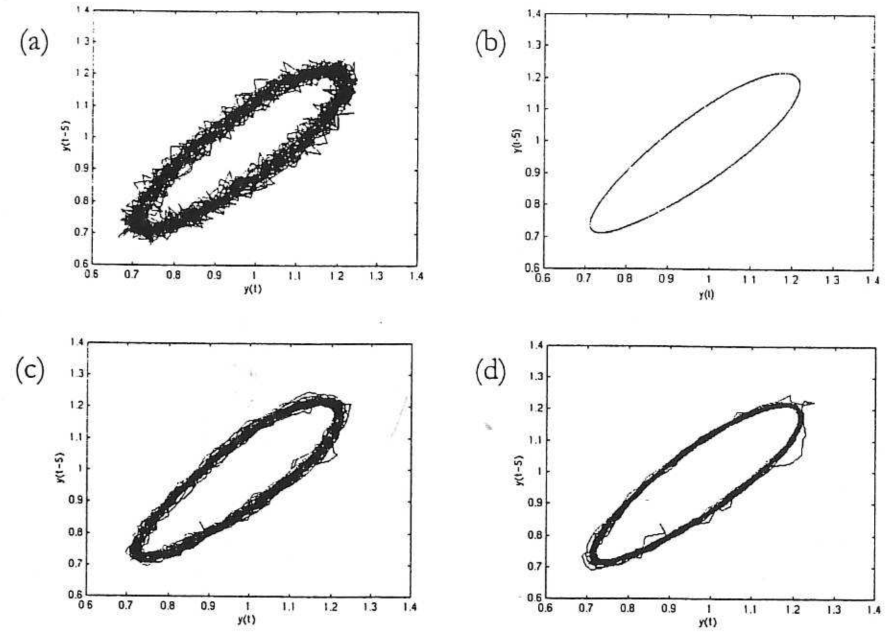

Figure 3 The phase portraits plots of the non-chaotic Duffing equation for (a) noisy output signal $y_t$, (b) noise-free signal $x_t$, (c) filtered signal using filter 1 and (d) filtered signal using filter 2 in Section 3.1.

identified from the training data set and the mean squared prediction errors are calculated over the testing data set. This is out-of-sample prediction because the testing data set is unknown and is not used during the learning/training stage of model estimation. The structure of the model (the values of $l, n_y, n_\xi$ and $n_m$) which give the smallest mean squared prediction errors over the testing data set is interpreted to mean that the model has captured most of the predictable part of the signal. Therefore the structure of the smoother/filter can be chosen to be the same as the structure of the model with the smallest mean squared prediction errors over the testing data set.

Consider the simplest and idealistic case where zero mean uncorrelated unpredictable noise $e_t$ is added to the noise-free data $x_t$ to obtain the noisy output data $y_t$ as in eqn (15). The output data set is divided into training $(y_t)_{train}$ and testing $(y_t)_{test}$ data sets. By specifying a model structure, a model $\hat{f}(\bullet)$ can then be identified from the training data set using the forward regression orthogonal least squares algorithm (Chen et al. 1989) and the model is tested on the testing data set to give

$$(y_t)_{test} = \hat{f}\left(y_{t+n_y}, \ldots, y_{t+1}, y_{t-1}, \ldots, y_{t-n_y}, \xi_{t-1}, \ldots, \xi_{t-n_\xi}\right)_{test} + (\xi_t)_{test} \tag{19}$$

$$(y_t)_{test} = \left(y_t^{(1)}\right)_{test} + (\xi_t)_{test} \tag{20}$$

where $(\xi_t)_{test}$ and $(y_t^{(1)})_{test}$ are the prediction errors and the predictions of the identified model on the testing data set. The noise in the testing data set $(e_t)_{test}$ is unpredictable from eqn (15). Since $(e_t)_{test}$ is not used in the arguments of $\hat{f}(\bullet)$ in eqn (19) to make predictions, therefore $(y_t^{(1)})_{test}$ the predictions of the model $\hat{f}(\bullet)$ should not be correlated with the noise $(e_t)_{test}$. This gives

$$\frac{1}{N_{test}}\sum_{t=1}^{N_{test}}\left(y_t^{(1)}e_t\right)_{test} \approx 0 \tag{21}$$

The predictions $(y_t^{(1)})_{test}$ can be further broken down into

$$\left(y_t^{(1)}\right)_{test} = (x_t)_{test} + \left(\eta_t^{(1)}\right)_{test} \tag{22}$$

where $(\eta_t^{(1)})_{test}$ are the deviations of the predictions $(y_t^{(1)})_{test}$ from the true noise-free data $(x_t)_{test}$. One objective therefore would be to design the smoother so that $\sum_{t=1}^{N_{test}}\left(\eta_t^{(1)}\right)_{test}^2$ is as small as possible, where $N_{test}$ is the length of the testing set. Then the smoothed data or predictions $(y_t^{(1)})_{test}$ of the testing data set are close to the noise-free data $(x_t)_{test}$. Substituting $(y_t^{(1)})_{test}$ from eqn (22) into eqn (21) gives

$$\frac{1}{N_{test}}\sum_{t=1}^{N_{test}}\left(x_t e_t + \eta_t^{(1)}e_t\right)_{test} \approx 0 \tag{23}$$

Given the assumption that the noise is uncorrelated with the dynamics of the time series, $\frac{1}{N_{test}}\sum_{t=1}^{N_{test}}(x_t e_t)_{test} = 0$, therefore eqn (23) becomes

$$\frac{1}{N_{test}}\sum_{t=1}^{N_{test}}\left(\eta_t^{(1)}e_t\right)_{test} \approx 0 \tag{24}$$

Manipulation of eqns (15), (20) & (22) gives

$$(e_t)_{test} - \left(\eta_t^{(1)}\right)_{test} = (\xi_t)_{test} \tag{25}$$

Taking the average of the sum of the squares of eqn (25) gives

$$\frac{1}{N_{test}}\sum_{t=1}^{N_{test}}(e_t)_{test}^2 - \frac{2}{N_{test}}\sum_{t=1}^{N_{test}}\left(e_t\eta_t^{(1)}\right)_{test} + \frac{1}{N_{test}}\sum_{t=1}^{N_{test}}\left(\eta_t^{(1)}\right)_{test}^2 = \frac{1}{N_{test}}\sum_{t=1}^{N_{test}}(\xi_t)_{test}^2 \tag{26}$$

Using eqn (24), eqn (26) can be simplified as

$$\frac{1}{N_{test}}\sum_{t=1}^{N_{test}}(e_t)_{test}^2 + \frac{1}{N_{test}}\sum_{t=1}^{N_{test}}\left(\eta_t^{(1)}\right)_{test}^2 = \frac{1}{N_{test}}\sum_{t=1}^{N_{test}}(\xi_t)_{test}^2 \tag{27}$$

From eqn (27) a lower value of $\frac{1}{N_{test}}\sum_{t=1}^{N_{test}}(\xi_t)_{test}^2$ implies a smaller value of $\frac{1}{N_{test}}\sum_{t=1}^{N_{test}}\left(\eta_t^{(1)}\right)_{test}^2$. One possibility therefore is to select the best values of $l, n_y, n_\xi$ and $n_m$ to give a smoother with the smallest mean squared prediction errors $\frac{1}{N_{test}}\sum_{t=1}^{N_{test}}(\xi_t)_{test}^2$ over the testing data set. Here the word smallest is used with respect to the mean squared prediction errors found using other smoother structures. The absolute value of the smallest mean squared prediction error may be large since this will depend on the unknown noise variance, $\frac{1}{N_{test}}\sum_{t=1}^{N_{test}}(e_t)_{test}^2$.

The applicability of this approach to select the structure of the filter/smoother was tested on the non-chaotic Duffing time series studied in Section 3.1. The two thousands noisy data samples generated in Section 3.1 were used, the first thousand were used as the training data set and the last thousand were used as the testing data set. Filter 1 with the best 2 output terms and the best 3 linear noise terms and filter 2 with the best 10 output terms and the best 3 linear noise terms were each identified from the training data set using the forward regression orthogonal least squares algorithm (Chen et al. 1989). The mean squared prediction errors calculated on the testing data set were 0.000506 and 0.000387 respectively. Filter 2 should therefore be chosen to filter the data.

The noise reduction capability of the smoother on the same non-chaotic noisy data set of the Duffing example was also investigated. For the smoother structure the maximum values of $l$ and $n_y$ were set to be 4 and 10 respectively. Note that both positive and negative lags were allowed in the smoother and no noise terms were used. When smoothing the data it has been found that the noise terms in the smoother structure are not critical and can be ignored (Aguirre et al. 1996). This is because although all the noise will not be removed in the first iteration it will be reduced to an acceptable level after repeated smoothing operations. For the filtering case the noise terms of the filter play an important role to prevent the filtered data from biased and hence cannot be ignored in most cases (Aguirre et al 1996). By specifying different numbers of terms in the smoother structure, many different smoothers were identified using the forward regression orthogonal least squares algorithm (Chen et al. 1989) and different numbers of smoothing iterations were performed. The best result based on the highest value of $D$ obtained in eqn (18) for smoothing was 3.9858 after 10 iterations with 30 terms in the smoother. The same procedure was repeated for the filtering case except that now linear noise terms with a maximum lag of 20 were allowed in the filter. The best filter only achieved a value of $D = 3.4799$ at iteration 9 with 21 output terms and 8 linear noise terms. In this example therefore smoothing is beneficial for a non-chaotic time series. This is contrary to the findings in Aguirre et al. (1996). Since the noise reduction method is based on prediction, utilising both future and past information in a smoother should give better predictions compared to just using past information in the filtering case. Therefore smoothing would be expected to perform as well as if not better than filtering for most time series.

## 4.2 Number of Smoothing Iterations

A cross-validation approach has been employed to identify the smoother structure, the next task is to determine the number of smoothing iterations to be performed. One method is to stop the smoothing procedure when $d_i$, defined as the mean squared discarded residuals at iteration $i$

$$d_i = \frac{1}{N}\left(\left(\xi_t^{(i)}\right)^2\right) = \frac{1}{N}\left(y_t^{(i)} - y_t^{(i+1)}\right)^2 \tag{28}$$

is small and changes very little compared to $d_{i-1}$. A graph of $d_i$ plotted against iteration $i$ can be used to determine the required number of smoothing iterations to be performed, as shown in Section 5. When $d_i$ is small and changes very little compared to $d_{i-1}$, the smoothed data set is much closer to deterministic than the noisy output and there are minimal changes between the smoothed data sets obtained at iteration $i-1$ and iteration $i$.

The recommended smoothing procedure can now be summarised below. No noise terms have been used so as to reduce the computation time in searching for the structure of the smoother, which was selected based on the smallest mean squared prediction errors calculated over the testing data set as described in Section 4.1.

i) Divide the raw noisy output equally into training and testing data sets.

ii) Set the initial values of $l$ and $n_y$ to be 3 and 1.

iii) Set the number of important term $n_m = 4$ initially.

iv) Use the forward regression orthogonal least squares algorithm (Chen et al. 1989) to identify the smoother from the training data set by selecting the best $n_m$ important terms from the candidate terms in the polynomial expansion and estimating the associated parameters.

v) Calculate the mean squared prediction errors over the testing data set using the identified smoother in step (iv).

vi) Increase $n_m$ by one and repeat steps (iv) to (vi). Go to step (vii) when $n_m = 50$.

vii) Increase $n_y$ by one and repeat steps (iii) to (vii). Go to step (viii) when $n_y = 15$.

viii) Record the values of $(l, n_y, n_m)$ of the smoother with the smallest mean squared prediction errors calculated over the testing data set. Set $i = 0$.

ix) Re-identify a global smoother $f^{(i)}(\bullet)$ from all the data $y_t^{(i)}$ (training +testing data sets) using the values of $(l, n_y, n_m)$ obtained in step (viii).

x) Use the smoother $f^{(i)}(\bullet)$ to compute the smoothed data set $y_t^{(i+1)}$.

xi) Calculate the value of $d_i$.

xii) Increase $i$ by one and repeat steps (ix) to (xii) until $d_i$ becomes small and changes very little compared with $d_{i-1}$.

To save computation time a large maximum output lag value, such as $n_y = 15$, may be selected directly in step (ii). The forward regression orthogonal least squares algorithm (Chen et al. 1989) can then be used to select the best $n_m$ terms from the large candidate term set in the polynomial expansion. This avoids the iteration in step (vii). However for some other choices of nonlinear function $f(\bullet)$ in eqn (8) such as the Gaussian radial basis function, it may be advantageous to vary the values of $n_y$, $n_m$ and the Gaussian width to determine the smoother structure. In step (vi) the maximum allowable terms in the smoother has been set to be 50 because simulations suggest that the smallest mean squared prediction errors calculated over the test data usually occur at values of $n_m$ much smaller than 50. Similarly the maximum allowable $n_y$ which has been set to 15 and $l = 3$ have been chosen based on practical application of the algorithm. In step (ix) a global smoother is re-identified using all the data instead of just the training data in step (iv). This is because the simulation studies suggest that a higher noise reduction can be achieved using the global smoother identified in this way.

## 5. Simulation Studies of the New NARMA Smoothing Algorithm

The new smoothing procedure discussed in Section 4.2 will be tested on two other examples, the Lorenz and the Duffing chaotic time series.

### 5.1 The Lorenz and the Duffing Chaotic Time Series

The Lorenz equations are

$$\dot{x} = \sigma(y - x)$$

$$\dot{y} = rx - y - xz \qquad (29)$$

$\dot{z} = -bz + xy$

with $\sigma = 10, b = 8/3$ and $r = 28$. The Duffing example in eqn (14) was also used but with the parameters set to $w = 1, \gamma = 0.4, \alpha = 1$ and $F = 0.4$ to give a chaotic time series.

Both the Lorenz and the Duffing equations were integrated using a fourth order Runge-Kutta method with fixed step sizes of 0.001s and 0.01s respectively. Two thousand data points were collected from the $x$-coordinate with a sampling rate of 0.01s for the Lorenz and 0.1s for the Duffing time series respectively. These time series served as the noise-free data $x_t$. Zero mean uncorrelated unpredictable noise was then added to obtain the noisy output $y_t$ as defined in eqn (15). The resulting data sets had signal-to-noise ratios equal to 20dB. The smoothing procedure described in Section 4.2 was then implemented.

The smoother structures of $l = 3, n_y = 13, n_m = 16$ and $l = 3, n_y = 15, n_m = 8$ gave the smallest mean squared prediction errors on the testing data sets for the Lorenz and the Duffing time series respectively. These values were used to construct the smoothers to smooth the corresponding noisy output data. To determine the number of smoothing iterations to perform, the values of $d_i$ were plotted against the iteration $i$ as shown in Figure 4.

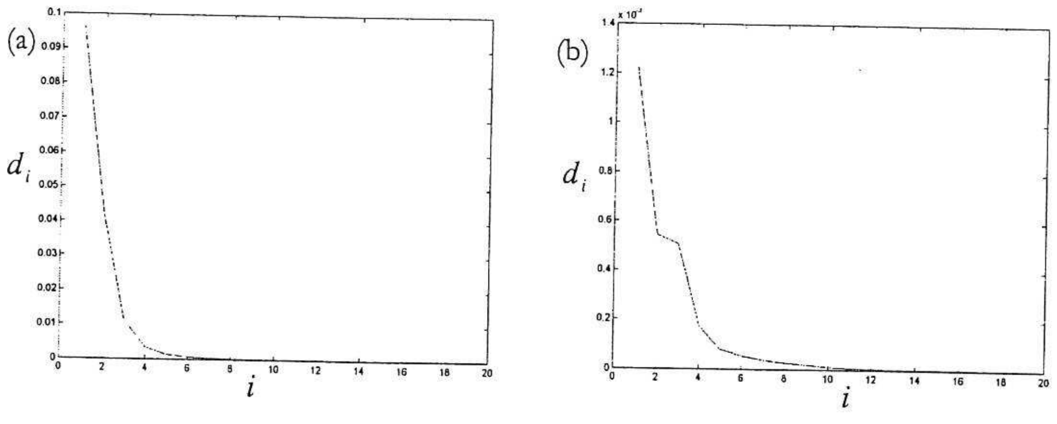


Figure 4 The values of the mean squared discarded residuals $d_i$ plotted against iteration $i$ for (a) the Lorenz and (b) the Duffing chaotic time series.

The values of $d_i$ were small and close to zero at iteration 6 and 10 for the Lorenz and the Duffing time series respectively. Therefore the smoothed data sets obtained at those iterations were taken as the final smoothed data sets. The values of $D$ in eqn (18) obtained for the corresponding smoothed data sets were 3.5169 and 2.9765 for the Lorenz and the Duffing examples respectively. The time series and phase portrait plots of the Lorenz and the Duffing examples are shown in Figure 5 and Figure 6 respectively. The new NARMA smoothing algorithm certainly seems to have reduced the noise while retaining the underlying dynamics of the time series.

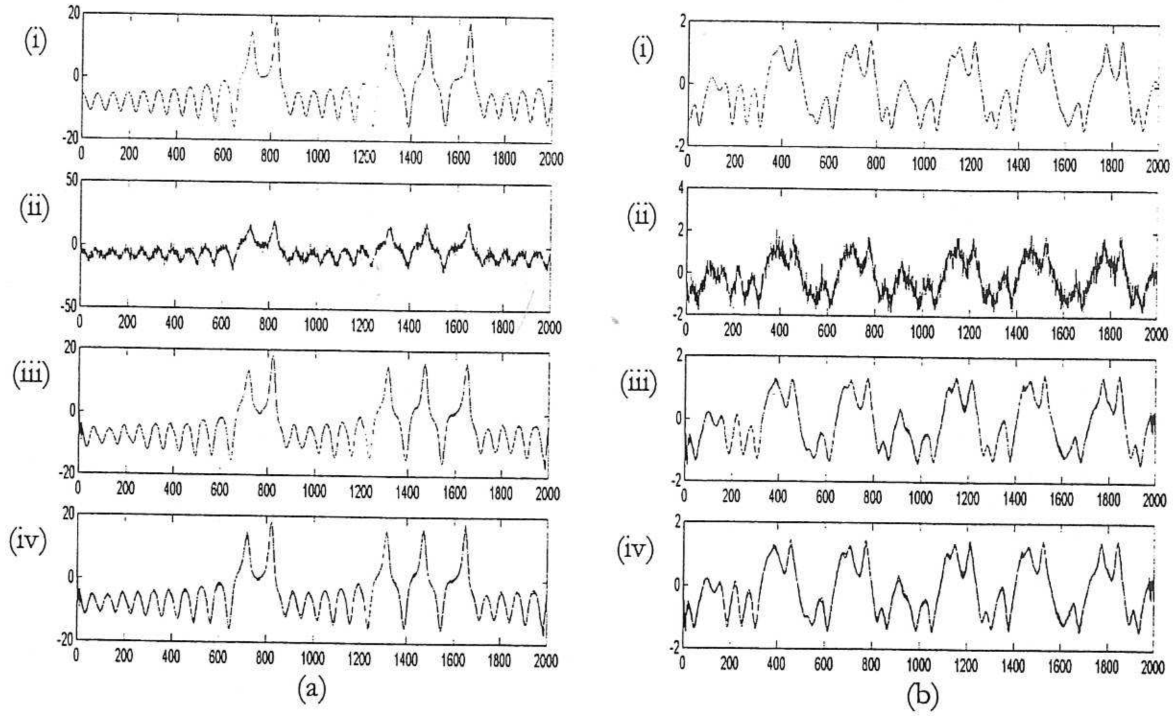


Figure 5 The time series plots for (a) the Lorenz and (b) the Duffing examples. The time series plots show (i) the noise-free data, (ii) the noisy output, (iii) the obtained smoothed data and (iv) a plot of (i) & (iii) superimposed.

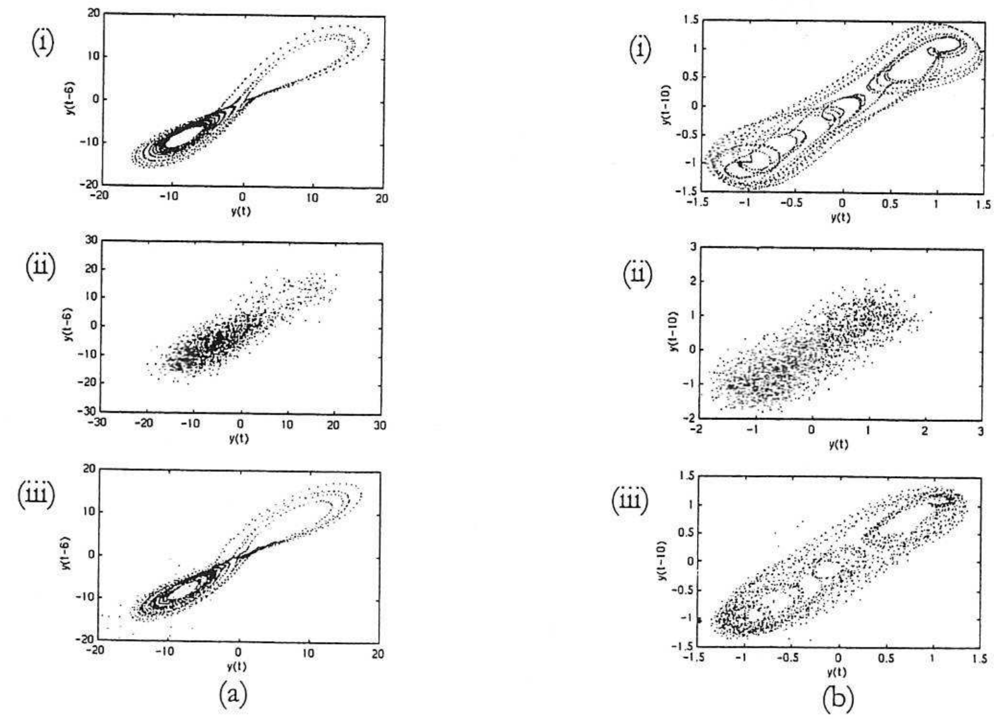


Figure 6. The phase portrait plots for (a) the Lorenz and (b) the Duffing chaotic time series with (i) the noise-free data $x_t$, (ii) the noisy output $y_t$ and (iii) the final smoothed data.

## 5.2 Correlated Noise Examples for the Lorenz and the Duffing Time Series

In Section 5.1, the noise-free data set was corrupted additively by uncorrelated noise $e_t$. In the current section, correlated noise examples will be considered. The noise-free data set is assumed to be corrupted additively by the correlated noise $\zeta_t$ to give the noisy output $y_t = x_t + \zeta_t$, where

$$\zeta_t = f\left(e_{t-1},\dots,e_{t-n_e}\right) + e_t \tag{30}$$

In the original smoothing algorithm described in Section 3, after $n_p$ numbers of smoothing operations have been performed (step (iv)), then the correlation tests are applied to check that the overall discarded residuals $\xi_t^{(T)}$ are unpredictable (step (v)). For correlated noise examples, the overall discarded residuals $\xi_t^{(T)}$ are likely to be correlated as well. Hence it is not an easy task to find the appropriate smoother structure so that the overall discarded residuals $\xi_t^{(T)}$ satisfy the correlation tests stated in eqns (6) & (7). The new smoothing algorithm discussed in Section 4 can still be applied directly to smooth data corrupted by correlated noise. To investigate this the noise-free data sets of the Lorenz and the Duffing examples in Section 5.1 were used. A correlated noise model

$$\zeta_t = 0.3e_{t-1} + e_t \tag{31}$$

was used and the noise-free data sets were corrupted additively to give noisy output data sets. Two signal-to-noise ratios of 20dB and 40dB were considered, and the new procedure described in Section 4.2 was employed to smooth the noisy output data sets without using any noise terms in the smoother structure. The results are presented in Table 1 and suggest that the new smoothing procedure can still perform satisfactorily for correlated noise examples, even though no noise terms were used in the smoother structure.

Table 1 The values of the index $D$ in eqn (18), which measures the effectiveness of the new NARMA smoothing algorithm on the Lorenz and the Duffing chaotic time series corrupted with correlated noise.

| *Time series* | *Signal-to-Noise ratio* | $D$ |
|---|---|---|
| Lorenz Time Series | 40dB | 2.9773 |
| | 20dB | 2.7686 |
| Duffing Time Series | 40dB | 2.3816 |
| | 20dB | 3.0166 |

## 6. Comparison Studies of Different Noise Reduction Methods

### 6.1 High Frequency Noise Reduction Using Wavelets

The multi-scale feature of the wavelet transform allows the decomposition of a signal into a number of scales. This decomposition breaks the signal down into a set of signals of varying "coarseness" ranging from low frequency components progressively to higher frequency components. If the measurement noise has high frequency components these can be removed by deleting the set of wavelet coefficients at higher scales. The main problem is deciding which scales of the wavelet decomposition represent mostly noise so that the wavelet coefficients at those scales can be set to zero. This process is known as wavelet thresholding. One property of the wavelet coefficients is that the variance of the coefficients is scale invariant when uncorrelated random noise is considered (Coca and Billings 2001). This behaviour yields a very simple and effective method for reducing high frequency random noise. If the noisy data is approximated in terms of a wavelet series, the wavelet coefficients tend to remain within a certain band of fairly constant amplitude above a certain scale $j$. These coefficients tend to account for the high frequency components of the signal which are assumed to be caused mainly by uncorrelated random noise. By simply setting to zero and thus omitting the corresponding wavelet coefficients for parts of the series that represents the random noise, the signal can be smoothed. Fast algorithms are available to decompose the signal into different scales. In this paper, the B-spline scaling and wavelet functions proposed by Chui (1992) and developed by Coca and Billings (2001) were employed to decompose the signal.

### 6.2 The Iterative Singular Value Decomposition Noise Reduction Method

Recently another simple algorithm called the iterative Singular Value Decomposition (SVD) was introduced by Shin et al. (1999) for noise reduction of low dimensional chaotic time series. Given a noisy output data set, a sequence of vector $\{\mathbf{y}_i \in \Re^{dd}, i = 1,\ldots,N_{dd}\}$ can be generated and a matrix $Y_{dd}$ can be constructed as

$$Y_{dd} = \begin{bmatrix} \mathbf{y}_1^T \\ \mathbf{y}_2^T \\ \vdots \\ \mathbf{y}_{N_{dd}}^T \end{bmatrix} = \begin{bmatrix} y_1 & y_2 & \cdots & y_{dd} \\ y_2 & y_3 & \cdots & y_{dd+1} \\ \vdots & & \ddots & \vdots \\ y_{N_{dd}} & y_{N_{dd}+1} & \cdots & y_{N_{dd}+dd-1} \end{bmatrix} \tag{32}$$

where $N_{dd} = N - (dd - 1)$. Performing singular value decomposition of the matrix $Y_{dd}$ gives

$$Y_{dd} = USV^T = \begin{bmatrix} u_1 & u_2 & \cdots & u_{dd} \end{bmatrix} \begin{bmatrix} s_1 & & & \\ & s_2 & & \\ & & \ddots & \\ & & & s_{dd} \end{bmatrix} \begin{bmatrix} v_1^T \\ v_2^T \\ \vdots \\ v_{dd}^T \end{bmatrix} \tag{33}$$

where $U$ is the $N_{dd} \times dd$ matrix of eigenvectors of $Y_{dd} Y_{dd}^T$ and $N_{dd} >> dd$, $V$ is the $dd \times dd$ matrix of eigenvectors of $Y_{dd}^T Y_{dd}$ and $S$ is the $dd \times dd$ diagonal matrix consisting of singular values, that is $diag(s_1, s_2, ..., s_{dd})$. The noise reduction of the iterative SVD method is obtained by using only the first singular value to construct the noise-reduced matrix $Y_{rd}$ as

$$Y_{dd} = Y_{rd} + N_{noise} = \begin{bmatrix} u_1 & U_2 \end{bmatrix} \begin{bmatrix} s_1 & 0 \\ 0 & S_2 \end{bmatrix} \begin{bmatrix} v_1^T \\ V_2^T \end{bmatrix} \tag{34}$$

where $N_{noise}$ is the noise dominated part. Details of the algorithm can be found in Shin et al. (1999). The procedure of the iterative SVD noise reduction method can be summarised as follows.

i) Construct the matrix $Y_{dd}$ from the noisy outputs.

ii) Apply the singular value decomposition, $Y_{dd} = USV^T$ in eqn (33).

iii) Construct the noise-reduced matrix $Y_{rd}$ using only the first singular value, see eqn (34).

iv) Obtain the noise-reduced data by averaging each column of $Y_{rd}$.

This procedure can be repeated a few times. It has to be pointed out that although the noise reduction method is very simple, however this method still requires the user to choose the value of $dd$ so that the matrix $Y_{dd}$ can be constructed and the noise reduction procedure can be started.

### 6.3 Simulation Results of the Different Noise Reduction Methods

The application of the NARMA smoother with cross-validation, wavelet noise reduction, a combination of the two and the iterative SVD method were studied. The combination method consisted of applying the NARMA smoother with cross-validation to smooth the noisy data for only one iteration. This was followed by a wavelet decomposition of the smoothed data to investigate whether a further reduction in the noise was possible by removing any high frequency noise which remained in the smoothed data. The noise-free data sets of the Lorenz and the Duffing examples in Section 5.1 were used. A zero mean uncorrelated unpredictable noise was added additively to the noise-free data to obtain the noisy output. Three different signal-to-noise ratios of 40dB, 20dB and 0dB were considered, and the four noise reduction methods were applied to smooth the noisy output. To show that other nonlinear functions can also be used to perform the NARMA smoothing algorithm, a thin-plate-spline radial basis function was employed for the representation of $f(\cdot)$ in eqn (8).

The wavelet decompositions of the smoothed data obtained after one iteration of the NARMA smoother with cross-validation are shown in Figure 7(b) for the Lorenz time series and the 20dB signal-to-noise ratio case. From Figure 7(b), the wavelet coefficients at the first three scales ($j-1, j-2, j-3$) can be seen to be quite small (mostly within the values of ±5) compared to that at the fourth and fifth scales. Hence the wavelet coefficients at the first three scales were assumed to be caused by the remaining high frequency noise in the smoothed data and were set to zero. The wavelet decomposition of the original output data set are shown in Figure 7(a). In this case it is not as clear as in the smoothed data case which wavelet coefficients represent noise. Under or over smoothing may occur if the cut off in the wavelet decomposition is not chosen properly. Under smoothing will result in only a marginal improvement compared to the original signal whereas over smoothing may cause some of the dynamics of the time series to be lost. Applying the NARMA smoother with cross-validation for one iteration removes a certain amount of noise and this often facilitates the identification of the wavelet coefficients which represent the noise. Similar findings were obtained for the 40dB and 0dB examples. The values of $D$, calculated using eqn (18), for the four different noise reduction methods are shown in Table 2. The signal $y_t^{(n_p)}$ in eqn (18) represents the final smoothed signal obtained from the different noise reduction methods. For the iterative SVD method, the value of $dd$ in eqn (32) has to be selected before the noise reduction process described in Section 6.2 can begin. It was assumed that the noise-free data were available so that the value of $D$ can be computed for the iterative SVD method. Hence the value of $dd$ was chosen such that the best results (the highest value of $D$ attainable) for the iterative SVD method was reported in Table 2. Clearly this comparison approach would favour the iterative SVD method. However the best results was still obtained by the combination approach which does not employ any priori knowledge of the time series. Similar findings were observed for the Duffing chaotic time series and the values of $D$ for this case are presented in Table 3. Again the combination approach has performed the best in all the cases considered.

Table 2 The values of D obtained for the four noise reduction methods for the Lorenz time series with three different signal-to-noise ratios (40 dB, 20 dB and 0 dB), (a) the NARMA smoother with cross-validation method $D_{nar}$, (b) the wavelet thresholding method $D_{wav}$, (c) the combination of NARMA smoothing and the wavelet method $D_{com}$ and (d) the iterative SVD method $D_{svd}$.

| *Lorenz time series* | $D_{nar}$ | $D_{wav}$ | $D_{com}$ | $D_{svd}$ |
|---|---|---|---|---|
| Case (i) 40dB | 3.3986 | 3.1351 | 3.7009 | 2.6203 |
| Case (ii) 20dB | 3.9117 | 3.1597 | 4.4215 | 3.3999 |
| Case (iii) 0 dB | 4.1492 | 4.2274 | 4.7623 | 4.5816 |

Table 3 The values of D obtained for the four noise reduction methods for the Duffing chaotic time series with three different signal-to-noise ratios (40 dB, 20 dB and 0 dB), (a) the NARMA smoother with cross-validation method $D_{nar}$, (b) the wavelet thresholding method $D_{wav}$, (c) the combination of NARMA smoothing and the wavelet method $D_{com}$ and (d) the iterative SVD method $D_{svd}$.

| *Duffing time series* | $D_{nar}$ | $D_{wav}$ | $D_{com}$ | $D_{svd}$ |
|---|---|---|---|---|
| Case (i) 40 dB | 3.0905 | 3.1336 | 3.5749 | 2.7408 |
| Case (ii) 20 dB | 2.8446 | 3.2477 | 3.7417 | 3.5222 |
| Case (iii) 0 dB | 3.8613 | 4.3053 | 4.7134 | 4.7802 |

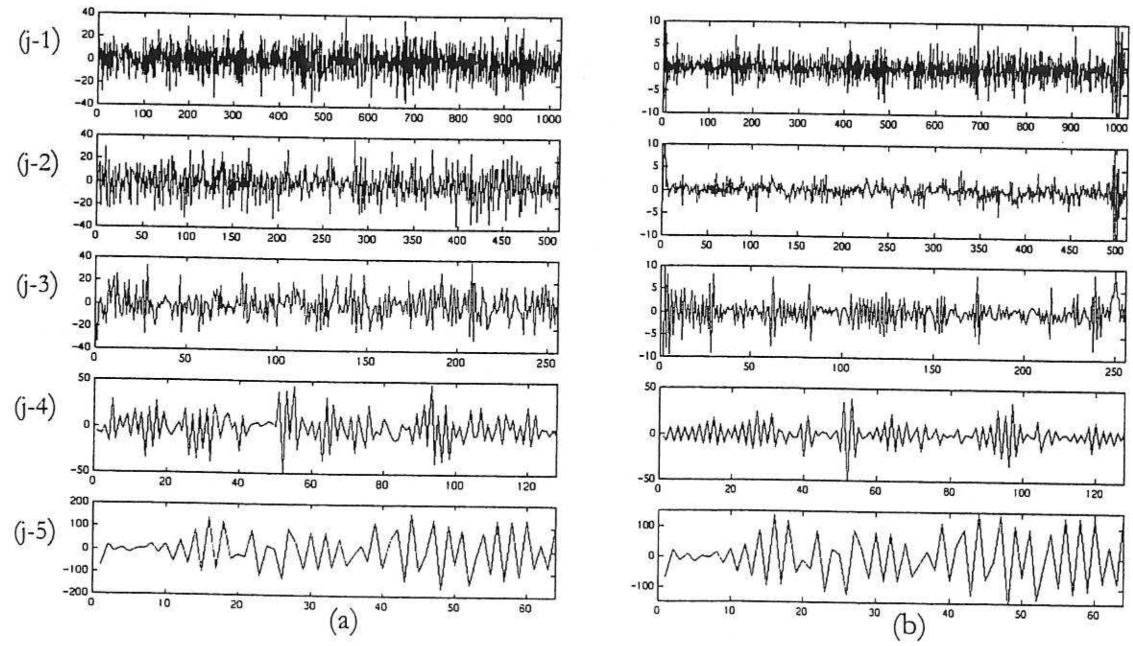


Figure 7 Wavelet coefficients corresponding to a 20dB signal-to-noise ratio for (a) the raw noisy signal and (b) the smoothed signal of the Lorenz time series. Only the first five scales are shown.

### 6.4 Discussions on the NARMA Smoother and the Combination Noise Reduction Method

In most parts of this paper, the Nonlinear AutoRegressive (NAR) structure was considered rather than the more complicated Nonlinear AutoRegressive Moving Average (NARMA) model structure. The main reason for this choice was the increase in computation that would be involved in finding the unknown parameters in a NARMA structure. The increase in computation time is not simply due to the increase in the search space due to the extra Moving Average (MA) terms. The identification of the MA terms requires the use of an iterative approach, and hence the identification of a pure autoregressive model is much simpler and quicker than for a pure moving average model. Furthermore a pseudo-linear approach is usually employed to identify the moving average terms. In the case of linear models it has been proven that strict conditions have to be satisfied otherwise the model may become unstable (Ljung and Söderström 1983). The same effect has been observed in the nonlinear case. Hence the more time-consuming prediction error method may have to be employed to identify NARMA models. Therefore even if the NARMA models could achieve a higher noise reduction ratio, the increase in computation time may not justify the usage of NARMA models and hence only NAR models were considered in this paper.

The main advantage of the NARMA smoother with cross-validation is that most of the unknown parameters in the algorithm can be estimated using the smallest mean squared prediction errors calculated over the testing data set. Hence the method is robust, requires very little user interaction and no knowledge of the dynamics of the time series are required to start the procedure. The use of the combination approach based on NARMA smoothing followed by wavelet thresholding to reduce high frequency noise seems to always produce the best results in all the examples that were investigated.

A crucial factor, which affects the noise reduction capability of all the different noise reduction methods, is the sampling rate. For the NARMA smoother with cross-validation, increasing the sampling rate means that the prediction time step is smaller. Hence better predictions can usually be obtained and a higher noise reduction ratio can usually be achieved.

## 7. Conclusions

A new smoothing algorithm which uses a cross-validation approach to determine the smoother structure has been introduced. The new algorithm requires very little user interaction and can be applied directly to smooth data corrupted by correlated noise. Numerical examples were included to illustrate the application of the new smoothing algorithm. It was also shown that wavelet thresholding to remove any high frequency noise remaining in the smoothed data after one iteration of the NARMA smoother can often further enhance the noise reduction.


### Acknowledgement

SAB gratefully acknowledges that part of this work was supported by EPSRC.


### References


Aguirre, L.A., Mendes, E.M. and Billings, S.A. (1996) Smoothing data with local instabilities for the identification of chaotic systems. International Journal of Control, vol 63, no 3, p483-505.

Billings, S.A., Korenberg, M.J. and Chen, S. (1988) Identification of nonlinear output affine systems using an orthogonal least squares algorithm. International Journal of Systems Science, 19, p1559-1568.

Billings, S.A. and Zhu, Q.M. (1994) Nonlinear model validation using the correlation tests. International Journal of Control, vol 60, no 6, p1107-1120.

Bhowal, A. and Roy, T.K. (1999) Noise reduction in chaotic time series data. PRAMANA-Journal of Physics, vol 52, no 1, p25-37.

Chen, S, Billings, S.A. and Luo, W. (1989). Orthogonal least squares methods and their application to nonlinear system identification. International Journal of Control, no 50, p1873-1896.

Chui, C. (1992) An introduction to Wavelets. Academic Press, New York.

Coca, D. and Billings, S.A. (2001) Nonlinear system identification using wavelet multiresolution models; International Journal of Control, vol 74, no18, p1718-1736.

Davies, M. (1994) Noise reduction schemes for chaotic time series. Physica D 79, p174-192.

Davies, M (1999) Nonlinear noise reduction through Monte Carlo sampling. Chaos, vol 8, no 4, p775-781.

Grassberger, P., Hegger, R., Kantz, H., Schaffrath, C. and Schreiber, T. (1993) On noise reduction methods for chaotic data. Chaos, 3, p127-141.

Kantz, H., Schreiber, T., Hoffmann, I., Buzug, T., Pfister, G., Flepp, L.G., Simonet, J., Badii, R. and Brun, E. (1993) Nonlinear noise reduction: A case study on experimental data. Physical Review E, vol 48, no 2, p1529-1538.

Korenberg, M.J., Billings, S.A., Liu, Y.P. and Mcilroy, P.J. (1988) Orthogonal parameter estimation algorithm for nonlinear stochastic systems. International Journal of Control, 48, p193-210.

Kostelich, E.J. and York, J.A. (1990) Noise reduction: finding the simplest dynamical system consistent with the data. Physica D 41, p183-196.

Leontaritis, I.J. and Billings, S.A. (1985a) Input-output parametric models for nonlinear systems part I: deterministic nonlinear systems. International Journal of Control, 41, p303-328; (1985b) Input-output parametric models for nonlinear systems part II: stochastic nonlinear systems. International Journal of Control, 41, p329-344.

Ljung, L. and Söderström, T. (1983) Theory and practice of recursive identification. MIT press, Cambridge, MA.

Schreiber, T. and Kantz, H. (1996) Observing and predicting chaotic signal: Is 2% noise too much? In Kravtsw, Y. and Kadtke, J. editors, Predictability of Complex Dynamical Systems, Springer Series in Synergetics No.69, 1-24, Springer New York.

Shin, K., Hammond, J.K. and White, P.R. (1999) Iterative SVD method for noise reduction of low-dimensional chaotic time series. Mechanical Systems and Signal Processing, 13(1), p115-124.