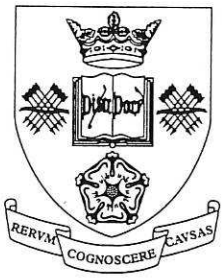
NONLINEAR SYSTEM IDENTIFICATION USING WAVELET MULTIRESOLUTION MODELS

D. COCA



S.A. BILLINGS

Department of Automatic Control and Systems Engineering,
University of Sheffield
Sheffield, S1 3JD,
UK

Research Report No.763
December 1999

Nonlinear System Identification using Wavelet Multiresolution Models

D. COCA S.A. BILLINGS
Department of Automatic Control and Systems Engineering,
University of Sheffield,
Mappin Street, Sheffield S1 3JD, UK

Abstract

A new system methodology for identifying nonlinear NARMAX models, from noise corrupted data, is introduced based on semi-orthogonal wavelet multiresolution approximations. An adaptive model sequencing strategy is introduced to infer model complexity from the data while reducing computational costs. This is used in conjunction with an iterative orthogonal-forward-regression routine coupled with model validity tests to identify sparse but accurate wavelet series representations of nonlinear processes. Experimental data from two real systems, a liquid level system and from a civil engineering structure are used to illustrate the effectiveness of the new identification procedure.

1 Introduction

Mathematical models are fundamental for the analysis of system behaviour, in controller design and many scientific and engineering studies. In every case an essential requirement of the model is an ability to reproduce the dynamical characteristics of the system as closely as possible.

A large class of dynamical systems can be represented with good approximation by linear models. But linear models cannot reproduce dynamical regimes which result from nonlinear interactions such as hysteresis, amplitude dependence, bifurcation or chaos. Nonlinear models will be required to capture these effects and this in turn leads to the complex problem associated with identifying accurate nonlinear models from noise corrupted data. If any a priori information about the system is known, this can be included by priming the algorithm with known model terms. But often the models are constructed purely from experimental data. This defines the black box identification problem.

The NARMAX model (Leontaritis and Billings, 1985a) can describe a wide range of nonlinear dynamical behaviours and includes many other nonlinear model types, including the Volterra, Hammerstein and Wiener models, as special cases (Pearson and Ogunnaike, 1997). However, identifying the relationship which describes some observed nonlinear interactions is a difficult problem, due to the immense number of possible model terms and combinations of terms. To overcome this problem many applications make use of methods for decomposing arbitrary functions into sums of special functions. Such methods of relating unknown functions or empirically derived variables to a set of fundamental *basis* functions are particularly useful in nonlinear system identification (Sjöberg et al., 1995; Juditski et al., 1995). The identification problem can then be reduced to a search for a nonlinear parametrisation of the NARMAX model

within a specific family of functions which can provide an exact or approximate representation of the observed dynamics.

The choice of approximating functions and the efficiency of the algorithm used to select the model structure are key elements in the identification procedure. Some classes of functions for example, are not rich enough to guarantee the convergence of the approximation to an arbitrary function. These can approximate efficiently only certain families of nonlinear input/output mappings. In principle, the more general the functions that can be approximated the more flexible the approach. Usually by increasing the approximation power, less parameters are necessary, hence the model becomes simpler. The drawback is that usually as the flexibility increases the method becomes more susceptible to errors present in the measurements which can lead to a false model. Finding the optimal parametrisation can be achieved by performing the model search successively, within function classes of increasing complexity until the correct model if found.

The necessary framework for implementing such an adaptive identification approach is provided by recent mathematical theories dealing with the representation of functions and distributions by multiresolution wavelet series. At the heart of the method is the concept of approximating an arbitrary nonlinear function in terms of dilates and translates of a single function, the 'mother' wavelet function. Unlike the Fourier basis functions the wavelet basis functions are localised both in space and frequency and provide an extremely versatile tool for synthetising the most "nasty" functions, very economically.

Various theoretical studies have proven that wavelet based approximation schemes are asymptotically near-optimal (DeVore et al., 1992), in the sense that the convergence rates are equal to the best attainable using general nonlinear approximation schemes, over a wide range of function spaces. In practice this means that wavelet approximations converge fast, in the appropriate norm, to any arbitrary function f regardless of the function class.

These properties make wavelets ideal model building blocks in nonlinear system identification. The crucial problems are choosing, among many possible wavelet representations, the best suited for this task and developing efficient wavelet selection algorithms which avoid the prohibitive computational costs mentioned by other studies (Zhang, 1997).

This paper presents a new approach to identifying NARMAX models from experimental input/output measurements based on a special class of wavelet multiresolution decompositions. A previous study has highlighted the efficiency of this method in the analysis of chaotic systems (Billings and Coca, 1999). The present work aims to demonstrate the applicability of this novel identification procedure to more general engineering systems. The method involves expanding the intrinsic nonlinearity in the NARMAX model as a multi-variable multiresolution wavelet series implemented with B-spline wavelets (Chui and Wang, 1992; Chui, 1992). Essential elements of this approach are the wavelet selection and model refinement algorithms, which were designed to minimise the computational cost and to provide a parsimonious representation of the dynamical process.

Inspired by the *projection pursuit* algorithm of Friedman and the ASMOD model introduced by Kavli (1993,1994), a new adaptive model sequencing strategy has been developed which exploits the hierarchical approximation structure provided by wavelet multiresolution approximations (Mallat, 1989) and the approximation properties of the B-spline wavelets. Starting with

a simple model form, the structure is refined iteratively until the selection algorithm produces a suitable model. A fast and efficient selection algorithm, based on the Orthogonal Forward Regression (OFR) (Chen et al., 1989), is then used to select only the relevant wavelet basis functions and to include these in the final model. Validation tests are used throughout to guide the model refinement process. The new procedures are illustrated using two real data sets from a liquid level system and a civil engineering structure.

2 Mathematical Realisations of Dynamical Systems

A dynamical system is a quadruple $\Sigma = (\mathcal{T}, \mathcal{X}, \mathcal{U}, \phi)$ where $\mathcal{T}$ is a time set $\mathcal{X}$ is a nonempty set called the *state space* of Σ, $\mathcal{U}$ is a nonempty set called the *input* or *control* space of Σ and $\phi : \mathcal{A}_\phi \to \mathcal{X}$ is a transition map Σ which is defined on a subset $\mathcal{A}_\phi$ of

$$\left\{(t, t_0, x, u) \mid t_0, t \in \mathcal{T},\ t_0 \leq t,\ x \in \mathcal{X},\ u \in \mathcal{U}^{[t_0,t)}\right\}, \tag{1}$$

where $\mathcal{U}^{\mathcal{I}} = \{u \mid u : \mathcal{I} \subset \mathcal{T} \to \mathcal{U}\}$ is a set of *input* or *control* maps.

The dynamical system may evolve continuously (the time set is real $\mathcal{T} = \mathbb{R}_+$) or for discrete instances in time (the time set is integer $\mathcal{T} = \mathbb{Z}_+$).

A dynamical system Σ can be augmented with a set $\mathcal{Y}$ called the *measurement* or *output space* and a map $h : \mathcal{T} \times \mathcal{X} \to \mathcal{Y}$ called the *measurement* or *observation map.*

The result is a dynamical system *with outputs*, which is denoted $\Sigma = (\mathcal{T}, \mathcal{X}, \mathcal{U}, \phi, \mathcal{Y}, h)$ with $u \in \mathcal{U}$, $x \in \mathcal{X}$, $y \in \mathcal{Y}$ the input, state and output variables respectively.

2.1 The NARMAX Input/Output Realisation

Dynamical systems are often characterised only in terms of the *input/output behavior* which illustrates the effects of the system's inputs $u \in \mathcal{U}^{\mathcal{I}}$ on the observed outputs $y \in \mathcal{Y}$. The assumption is that all the information about the state x should in principle be recoverable from the measurements y. In many cases it is more advantageous to have a direct characterisation of the output of the system as a function of the input only.

In system identification the input/output behaviour of a dynamical system is a discrete set of input and output samples

$$\mathcal{D}^t = \left\{(u(k), y(k))^t \mid u(k) \in \mathcal{U}^{\mathcal{I}},\ y(k) \in \mathcal{Y},\ k = 1, ..., t\right\} \tag{2}$$

which is often contaminated by noise.

The identification problem consists in determining the input/output equation that relates explicitly the sampled outputs of the system u^t to the sampled inputs y^t. Although it is still possible to identify continuous-time models from this data, the identification of discrete-time models is usually more convenient. Discrete-time models are widely used both for analysis as well as in the design of digital control systems.

The class of ARMAX models is a well known input/output representation in linear system identification ((Eykhoff, 1974; Ljung, 1987; Söderström and Stoica, 1989)). These models relate

the input and output sequences by linear difference equations,

$$y(t) = \sum_{i=1}^{n_y} a_i y(t-i) + \sum_{i=1}^{n_u} b_i u(t-i) + \sum_{i=1}^{n_e} c_i e(t-i) + e(t) \tag{3}$$

taking into account the combined effects of measurement noise, modelling errors and unmeasured disturbances which in (3) are represented by the variable $e(t)$. The usefulness of linear models to describe nonlinear dynamics is however limited. Most nonlinear phenomena can only be described in terms of nonlinear equations. The best known class of nonlinear discrete-time models used in identification, is the NARMAX family introduced by Leontaritis and Billings (1985a,b). These models are similar to the ARMAX representation, except that now the NARMAX model involves a nonlinear expression

$$y(t) = f(y(t-1), ..., y(t-n_y), u(t-1), ..., u(t-n_u), e(t-1), ..., e(t-n_e)) + e(t) \tag{4}$$

where f is a nonlinear function of $n_u + n_y + n_e$ variables.

In the above equations $f : \mathcal{Y}^{n_y} \times \mathcal{U}^{n_u} \times \Xi^{n_e} \rightarrow \mathcal{Y}$ is an unknown nonlinear mapping, $u \in \mathcal{U}$ an m-dimensional input vector, $y \in \mathcal{Y}$ an l-dimensional output vector and n_y and n_u are the maximum output and input lags. The unobserved stochastic variable $e(t) \in \mathcal{X})$ is assumed to be bounded $|e(t)| < \delta$ and uncorrelated with the input and n_e is the maximum noise lag. The random variable $e(t)$ is the prediction error or innovation at time t for the stochastic dynamical system (4).

2.2 The Realisation Problem for a Nonlinear Function f

The methodology of identifying a NARMAX representation provides for estimating both the structure and the the parameters of the unknown nonlinear system from the input/output data. This is quite a formidable task since the number of possible nonlinear realisations of f is theoretically infinite. Finding f which best agrees with the experimental data according to some adequacy criterion can be formulated as a nonlinear approximation problem.

Unless some a priori knowledge of the nonlinear equations is available most methods use nonparametric regression (Juditski et al., 1995; Sjöberg et al., 1995) to estimate the nonlinear function f from data. This involves searching for the best nonlinear parametrisation of $f \in \mathcal{F}$

$$f = f(\, g(y(t-1), ..., u(t-1), ..., e(t-1), ...), \theta\,) \in \mathcal{F} \tag{5}$$

where $g = \{g_k\}$ is a finite dimensional vector of nonlinear regressors $g_k \in \mathcal{G}$, and θ is the associated parameter vector. The set $\mathcal{G}$ consists of a finite number of candidate regressors sometimes referred to as *basis functions* since often they form a *basis* of a functional space.

For example, f can be implemented as a linear expansion in terms of the (basis) functions of regressors g_k selected from $\mathcal{G}$ such that

$$f = \sum_{k \in \mathcal{K}} \theta_k g_k \in \mathcal{F} \tag{6}$$

minimises a given adequacy criterion relative to the identification data

$$J(f(y^t, u^t, e^t, \theta), y^t, u^t, e^t) \tag{7}$$

subject to some additional complexity criterion

$$Q(\theta, n_\theta, t) \tag{8}$$

which penalises models with a large number of parameters. Although (6) is a parametric model, the approach is nonparametric because the number of parameters and the elements of the regression vector g are not known in advance.

Practical implementations of the candidate regressor set $\mathcal{G}$ include polynomial (Billings and Chen, 1989), radial basis functions (Chen et al., 1990), B-Splines (Kavli, 1993; Kavli and Weyer, 1994) or radial wavelets (Zhang, 1997). So far none of the existing approaches can be viewed as universally optimal and the choice of the model implementation is usually dictated by the application.

3 The Wavelet Model Structure

For years researchers have looked for a means to represent functions and distributions in terms of elementary basis functions that would characterise better the underlying properties, transitory phenomena or isolated singularities for example, than the Fourier series. Wavelet theory provides an elegant solution to this problem.

Wavelets provide the means to represent functions or distributions in terms of the translations and dilations of a unique prototype function, the *mother wavelet function*. Several properties, which make wavelet approximations particularly suitable for system identification, are briefly summarised below.

3.0.1 Approximation Rates

Wavelet-based approximation schemes compare favourably and often outperform many other approximation methods. This aspect has been emphasised by theoretical studies (DeVore et al., 1992) which have shown that wavelet approximations are asymptotically near-optimal in the sense that the convergence rates are equal to the best attainable using general nonlinear approximation schemes.

An important feature is that wavelets provide bases for a large variety of function spaces and are equally effective in each case. If f belongs to a functional space, such as the familiar Sobolev or Besov spaces for example, the wavelet approximation will converge automatically to f in the appropriate norm and, most importantly, provides similar rates of approximation in each case. In addition, wavelets are particularly effective in approximating functions characterised by sparse singularities or functions that are not uniformly smooth or regular which are notoriously difficult to approximate by other means (DeVore et al., 1992; Juditski et al., 1995).

3.0.2 Localisation

Unlike the Fourier basis functions wavelets are localised both in space and frequency. As a result the coefficients of the wavelet series translate precisely the local properties of the function that is approximated. Few wavelet coefficients are usually sufficient to approximate accurately a nonlinear function over a smooth interval. More wavelet coefficients are required only to describe sharp variations or singularities of the nonlinear function. The magnitude of the wavelet coefficients will reflect directly the properties of the underlying function which makes this approximation scheme an excellent singularity detection tool.

It follows that the approximation is spatially adaptive and hence it can be tuned in each subinterval of the input domain without interfering with the rest of the model. This allows fitting the isolated features (discontinuities or sharp variations) of the underlying function while keeping the complexity of the model low, thus reducing the risk of overfitting the data.

The wavelet bases provide approximating structures of arbitrary high order. In practice, increasing the order leads to better approximation rates. Ideally however the order (i.e. regularity) of the basis functions should closely match the regularity of the underlying function f to avoid overfitting.

3.0.3 Lacunarity

The ultimate goal of any approximation method is to use the fewest number of expansion coefficients to represent the nonlinear function f. Wavelets provide one of the most efficient methods to encode information which explains the large number of image and signal compression applications that use wavelets. In general the decomposition of regular functions with sparse singularity (belonging to the inhomogeneous Besov spaces $\mathcal{B}_p^{s,q}$ for example) is lacunary in the sense that very few coefficients of its decomposition will be non-negligible. The wavelet coefficients are significant near singularities and in the regions where f has large variations and small where the function is regular.

As noted by Meyer (1993), wavelet decompositions with plenty of non-negligible coefficients are characteristic of pathological functions. The 'normal' functions correspond to wavelet representations which are sparse or lacunary. The problem is deriving an efficient algorithm to select the essential wavelet basis functions and to provide a parsimonious representation of the system that has good generalisation properties.

3.1 Multiresolution Analysis

Wavelet multiresolution approximations (MRA) (Mallat, 1989) are constructed as a scale of nested finite-dimensional subspaces

$$...V_0 \subset V_1 \subset ... \subset V_j \subset ... \tag{9}$$

of a Hilbert space V. Usually V is the space $L^2(\mathbb{R})$ of square-summable functions.

The subspaces $\{V_j\}_{j\in\mathbb{Z}}$ are dense in V

$$\bigcup_{j=0}^{\infty} V_j = V \tag{10}$$

which means that any function $f \in V$ can be approximated with desired accuracy by its projection $f_j = P_j f$ on V_j that is, $\lim_{j\to\infty} f_j = f$.

The multiscale structure (9) can be associated with suitable detail spaces $W_j \subset V_j$ such that $V_j = V_{j-1} \oplus W_j$ provide a stable orthogonal split of V_j into low and high frequency parts V_{j-1} and W_j respectively. Using this two-level decomposition recursively V_j can be written as

$$V_j = V_l \oplus W_{l+1} \oplus W_{l+2} \oplus ... \oplus W_j \tag{11}$$

where all these subspaces are orthogonal and $l \in \mathbb{Z}$. By virtue of (10) and (11) this implies

$$V = \bigoplus_{i\in\mathbb{Z}} W_j \tag{12}$$

Finding a wavelet basis for $V = L^2(\mathbb{R})$ involves finding a basis for each subspace W_j. The most attractive bases are obviously orthonormal ones. However, finding a L^2-orthonormal wavelet basis in W_j which is also computable is very hard. Therefore, the orthogonality requirement of the basis in W_j is often relaxed whilst maintaining the orthogonality between subspaces W_j.

Most commonly, wavelet bases are derived using shift-invariance and dyadic dilation. The one-dimensional bases are constructed in terms of the dilations and translations of two prototype functions the scaling $\phi(\cdot)$ and the wavelet $\psi(\cdot)$ functions such that $V_j = \{\phi_{j,k}\}_{k\in\mathbb{Z}}$ and $W_j = \{\psi_{j,k}\}_{k\in\mathbb{Z}}$ with $\phi_{j,k}(x) = 2^{j/2}\psi(2^j x - k)$ and $\psi_{j,k}(x) = 2^{j/2}\phi(2^j x - k)$.

It follows that any square-summable function $f(x)$ in V can be expressed as a wavelet series expansion

$$f(x) = \sum_k c_{l,k}\, \phi_{l,k}(x) + \sum_{j=l+1}^{\infty} \sum_k d_{j,k}\, \psi_{j,k}(x) \tag{13}$$

where $\{c_{l,k}\}$ and $\{d_{j,k}\}$ are the coefficients of the expansion.

In equation (13) j is an integer parameter representing the scale (or dilation) while k is the translation parameter, an integer indicating the position of the basis function.

Multi-variable bases can be constructed using the tensor product method. A d-dimensional multiresolution approximation can be implemented using basis functions $\{\Phi(x)\}$ and $\{\Psi^{(l)}(x)\}_{l=1,2^d-1}$ constructed as tensor products of scalar basis functions. Assuming $x = (x_1, x_2)$ for example, the multiresolution decomposition can be implemented in terms of the translates and dilates of the following two-dimensional basis functions

$$\begin{aligned} \Phi(x) &= \phi(x_1)\phi(x_2) & \Psi^{(1)}(x) &= \phi(x_1)\psi(x_2) \\ \Psi^{(2)}(x) &= \psi(x_1)\phi(x_2) & \Psi^{(3)}(x) &= \psi(x_1)\psi(x_2) \end{aligned} \tag{14}$$

The choice of basis functions to implement the MRA is relatively wide and usually is

dictated by the application. The wavelet multiresolution expansion proposed here is defined in terms of B-spline scaling and wavelet basis functions, originally introduced by Chui and Wang (1992), which define a class of semi-orthogonal wavelets.

This implementation was chosen because it is particularly suitable in system identification. Some practical considerations are summarised below.

- B-spline wavelets are piecewise polynomial functions that can be computed easily for data which are not necessarily spaced uniformly. Some wavelets are only defined on a uniform grid, see for example Daubechies's orthonormal wavelets (Daubechies, 1993).
- B-spline wavelets of Chui and Wang have local support (i.e. the basis functions are zero outside a closed interval $\mathcal{B} = \text{Supp}(\psi(x))$). These functions also provide near-optimal time-frequency localisation (Chui, 1992).
- Spline wavelets outperform other wavelet decompositions in terms of approximation rate (Sweldens and Piessens, 1994). This means that fewer resolution levels are required to approximate a function with a given degree of accuracy. Since each extra level doubles the amount of work, the choice of wavelet is clearly important.

3.2 Semi-Orthogonal B-spline wavelet bases

The multiresolution approximation is implemented using as scaling function $\phi(x) = \phi^m(x) = \beta^m(x)$ the m-th order cardinal B-spline function given by the recursive relation (de Boor, 1978)

$$\beta^m(x) = \frac{x}{m-1}\beta^{m-1}(x) + \frac{m-x}{m-1}\beta^{m-1}(x-1) \tag{15}$$

where $\beta^1(x)$ is the indicator function

$$\beta^1(x) = \begin{cases} 1 & \text{if } x \in (0,1) \\ 0 & \text{otherwise} \end{cases} \tag{16}$$

Alternatively, each polynomial piece of a B-spline function can be computed separately (de Boor, 1978).

The wavelet function is defined as a linear combination of scaling functions (Chui and Wang, 1992)

$$\psi^m(x) = \sum_{l=0}^{3m-2} q_l \phi^m(2x-l) \tag{17}$$

with the coefficients given by

$$q_l = \frac{(-1)^k}{2^{m-1}} \sum_{l=0}^{m} \binom{m}{l} \phi^{2m}(k-l+1), \quad k = 0, \ldots, 3m-2 \tag{18}$$

The cubic B-spline scaling and wavelet functions ($m = 4$) are illustrated in Figs. (1a,b).

The B-spline wavelet multiresolution approximations are structured as a class indexed by the order (regularity) parameter m which is related to the order of polynomial pieces $(m-1)$ which make up the B-spline wavelets.

3.3 Wavelets in Identification

Practical wavelet model implementations derived in the past in system identification did not exploit the advantages of the multiresolution structure. The paper of Sjöberg *et al* proposes an identification scheme based on orthogonal wavelets introduced by Daubechies (1993). These wavelets however are defined only on a uniformly spaced grid over the input space and for that reason the original identification data must be used to generate a synthetic input first. The synthetic input is used to determine the wavelet coefficients. It is not clear how the resulting model will be used in practice and how well it performs since the algorithm was introduced only in principle and no simulation examples were given.

A similar orthogonal wavelet MRA model was suggested by Sureshbabu and Farell (1999). Again the approach was mainly presented in principle, the only practical implementation suggested using Haar wavelets which are known to have poor frequency localisation and lack regularity.

The Wavelet Networks introduced by Zhang and Beneviste (1992), Zhang (1997) are a powerful approximation device. The model implementation involves radial wavelet basis functions which do not generate a multiresolution approximation. The resulting wavelet model therefore resembles the more familiar RBF networks with the Gaussian or thin-spline functions replaced by radial wavelet basis functions. The wavelet basis was defined by discretisation of the continuous wavelet transform using either adaptive or fixed dilation/translation sampling. This implementation allowed a more systematic approach to selecting the model structure and determining the parameters associated with it.

The system identification approach introduced in the present study is centered around a new model implementation within the wavelet multiresolution approximation framework. Although this paper is not concerned with making a comparison with the existing approaches some advantages of the model structure advocated here over the wavelet network of Zhang and Beneviste are summarised below:

- The B-spline wavelet basis defines a hierarchical multiresolution structure with fixed dilation/translation sampling. Thus, the location of each basis function is fixed and does not have to be optimised by a separate algorithm as in Zhang (1997).

- While the radial wavelets are only 'localised' i.e. vanish rapidly as $x \rightarrow \pm\infty$, the B-spline wavelets are compactly supported. Hence, only the the B-spline wavelets which cover the data range need to be considered during identification.

- The multi-variable B-spline wavelet basis functions are implemented using the tensor product method as opposed to the radial construction used by Zhang. As noted in Sjöberg *et al* (1995) the tensor-product construction leads to more versatile multivariable basis functions as opposed to the radial construction which imposes some directional homogeneity.

- The wavelet model proposed here does not require the inclusion of all the process variables in every regressor as in the radial construction. This gives more flexibility in selecting the correct model structure and avoids model overfitting.

- The regularity of the radial wavelets of Zhang and Beneviste is fixed. The resulting approximation structure is very powerful but prone to overfitting. In contrast, the B-spline wavelets are structured as a class indexed by the regularity parameter m. This can be selected adaptively to reduce the risk of overfitting.

The main advantage of the wavelet model proposed by Zhang and Beneviste is that the radial construction leads to a smaller candidate regressor set compared with the tensor-product approach. However, the methodology of building wavelet multiresolution models proposed here, has been designed to deal efficiently with this problem.

3.4 The Wavelet Multiresolution Model

The wavelet model proposed here involves expanding the nonlinear function $f(\cdot)$ in (4) as a multiresolution wavelet series using B-spline scaling and wavelet functions

$$y(t) = \sum_{i=1} \theta_i g_i(t) + e(t) \tag{19}$$

where

$$g_i(t) = g_{j,k}(\, y(t-1), ..., y(t-n_y), u(t-1), ..., u(t-n_u), e(t-1), ..., e(t-n_e)\,) \tag{20}$$

is a multi-variable scaling or wavelet basis function $g_{j,k} \in \{\phi_{j,k}, \psi_{j,k}\}$ of past outputs, inputs and noise and $\theta = \{\theta_i\}$ is the parameter vector.

The family of nonlinear basis functions (or regressors) is doubly indexed according to scale and location and for the one-dimensional case can be written as follows

$$g_{j,k}(x) = \begin{cases} 2^{j/2}\phi(2^j x - k) & j = j_0 \\ 2^{j/2}\psi(2^j x - k) & j > j_0 \end{cases} \tag{21}$$

where $\phi(x)$ and $\psi(x)$ are the scaling and wavelet functions respectively and j_0 is an arbitrary starting scale. The multi-variable basis functions are constructed by tensor products of scalar basis functions preserving in this way the multiresolution structure in the multidimensional case

The approximation (19) converges in the standard space $L^2(\mathbb{R}^{n_y+n_u+n_e})$. In practice, f may not be square-summable in $\mathbb{R}^{n_y+n_u+n_e}$. However, the restriction of f on the closed interval $D \subset \mathbb{R}^{n_y+n_u+n_e}$ which contains the identification data belongs to $L^2(D)$ hence (19) converges in the L^2 norm.

As pointed out in Section 3 wavelet multiresolution approximations are just as effective and converge in the corresponding norm over a wide variety of function spaces including Besov spaces, Hölder and Hardy spaces (see Meyer 1993). This combined with the fact that for each

function space wavelet approximations provide approximation rates comparable with the best attainable by any other approximation scheme, makes the proposed wavelet structure a very powerful, near-universal implementation in nonlinear system identification.

4 The Identification Algorithm

The model terms in (19) are selected from a set $\mathcal{G}$ of candidate wavelet regressors implemented according to a given model structure $\mathcal{M}$. The model set corresponding to a complete multi-dimensional wavelet multiresolution approximation structure could contain hundreds or even thousands of basis functions. As noted earlier only a small fraction of the initial candidate wavelets are useful to implement the model. Also, the complete multi-dimensional wavelet multiresolution implementation is far too complex for normal identification tasks since often the nonlinear function $f(\cdot)$ in (4) can be represented as a superposition of low-dimensional functions. Hence, building a complete multi-dimensional wavelet basis is not always necessary.

The identification algorithm should be able to perform efficiently three essential tasks:

- Formulate a sequence of evolving models $\mathcal{M}_n$ and implement in a stepwise manner the corresponding sets $\mathcal{G}_n$ of candidate regressors
- At each step search for and select the best model realisation according to the adequacy criterion and the parsimony principle
- Model testing/validation

The algorithm is stopped when the selected model meets all the identification requirements.

4.1 Model Sequencing Strategy

The model proposed in Section 3.4 can describe virtually any realisable nonlinear input/output behaviour if the full multi-dimensional ($d = n_u + n_y + n_e$) wavelet multiresolution approximation is implemented. However, in order to avoid problems related to dealing with large regression sets $\mathcal{G}$, an "intelligent" identification algorithm should attempt to approximate f using simpler model structures first and avoid implementing the full multi-dimensional multiresolution approximation structure when the application does not require this.

The solution is to postulate a sequence of nested model architectures

$$\mathcal{M}_1, ..., \mathcal{M}_n \tag{22}$$

of increased complexity which corresponds to a sequence $\mathcal{G}_1$, ..., $\mathcal{G}_n$ of candidate regressor sets. Thus, the identification procedure becomes an iterative process of searching for a model within nested model families of increasing complexity until a suitable representation is found. This results in a sequence of nonlinear models $\mathbf{m_1}$, $\mathbf{m_2}$, ..., $\mathbf{m_n}$, the last one being the model sought.

The sequencing strategy proposed here is inspired by the *Projection Pursuit Regression Algorithm* developed by Friedman and Stueltze (1981) and used in the ASMOD (Adaptive

Spline Modelling of Observational Data) method introduced by Kavli (1994). A variant of this algorithm is the *Matching Pursuit Algorithm* proposed by Mallat and Zhang (1993) to perform adaptive time-frequency decompositions of speech signals. The projection or matching pursuit philosophy is integrated here into a new identification methodology which exploits the advantages offered by the multiresolution approximation structure and the efficiency of the orthogonal forward regression algorithm in performing subset selection and parameter estimation.

The simplest structure $\mathcal{M}_1 = \mathcal{M}_1(n_u, n_y, n_e)$ in the model sequence (22) is defined as a superposition of one-dimensional submodels

$$f = \sum_{i=1}^{n_u} f_i(u(t-i)) + \sum_{i=1}^{n_y} f_{n_u+i}(y(t-i)) + \sum_{i=1}^{n_e} f_{n_u+n_y+i}(e(t-i)) \tag{23}$$

In this equation each of the one-dimensional submodels f_i is implemented as a one-dimensional multiresolution expansion

$$f_i = \sum_k c^i_{j_i,k}\phi_{j_i,k} + \sum_k \sum_{l=j_i+1}^{j_i+N_s(i)} d^i_{l,k}\psi_{l,k} \tag{24}$$

j_i is the initial or starting scale and $N_s(i)$ is the number of approximation scales involved. Initially $N_s(i) = 1$. The model structure can further be indexed $(\mathcal{M}_1(n_u, n_y, n_e, N_s(i), m))$ in terms of the number of scales $N_s(i)$ and order m. The approximation can be refined for each of the additive submodels by considering additional wavelet subspaces W_j, W_{j+1}, W_{j+2}, ... The number of wavelet subspaces that can be considered is limited to those for which the support of the corresponding wavelet basis functions is large enough to cover a minimum number of data samples.

More complex model structures $\mathcal{M}_k$ are derived by replacing a combination of two or more submodels with a multi-dimensional tensor product submodel. This accounts for possible non-additive interactions between the corresponding variables. In each case, the candidate regressor set is revised according to the improved model structure $\mathcal{M}_k$. Again, $\mathcal{G}_i$ will include only the relevant regressors, i.e. the basis functions which cover the interval spanned by the data samples.

In most practical cases the nonlinear interactions can be modelled using a structure of moderate complexity. A typical model would be a superposition of low-dimensional submodels, similar to the ASMOD model introduced by Kavli (1993,1994)

$$y(t) = \sum_{i=1}^{n_f} f_i^{d(i)}(y(t-1), ..., u(t-1), ...e(t-1)) \tag{25}$$

where n_f is the number of additive submodels and $d(i)$ is the dimension of each submodel. Each submodel can include several wavelet subspaces W_j in addition to the initial coarse scaling subspace V_{j_0}. The selected model subset $\mathcal{G}_n(\Lambda) = \{g_{j,k}\}_{j,k\in\Lambda}$ will contain only a small number of low-dimensional basis functions selected from the candidate regressor set $\mathcal{G}_n$.

4.2 The Structure Selection and Parameter Estimation Algorithm

In theory, the wavelet multiresolution approximation is an infinite series expansion. However, even a finite dimensional wavelet basis is very large although, as pointed out earlier, only a small number of basis functions are needed to approximate typical nonlinear functions. Finding these basis functions, also known as structure selection, is a very important task in system identification. Formally this problem can be stated as follows:

Given a Hilbert space H spanned by a basis $\{g_j\}_{j\in\mathbb{Z}_+}$ which is not necessarily finite or orthogonal and the input/output data $(u(t), y(t))_{t=1}^N$, find the smallest partial set $\{g_1, ...g_k\}$ of $\{g_j\}_{j\in\mathbb{Z}_+}$ such that the input/output equations

$$y(t) = \sum_{j=1}^{k} c_j g_j(y(t-1), ..., u(t-1), ..., e(t-1), ...) + e(t) \tag{26}$$

approximate the input/output behaviour with a given accuracy $\|e\| \leq \rho_e$, for some parameters $c_j \in \mathbb{R}, j = 1, ..., k$.

In practice the basis $\{g_j\}_{j\in\mathbb{Z}_+}$ needs to be finite which means that any infinite dimensional basis has to be truncated in some way. A multiresolution wavelet basis $\mathcal{G}_n = \{g_j\}$ is truncated for practical reasons by including only wavelet subspaces W_j up to a certain scale $j \leq j_{\max}$ and only basis functions which cover the data range. The highest resolution or scale is such that the corresponding wavelets have support which is large enough to include several observations. Statistical laws of the log log type provide an estimate of the upper bound $j_{\max}$ such that

$$j_{\max} = \max\left\{ j | 2^{dj} \leq \frac{2N}{logN} \right\} \tag{27}$$

which assumes that the data is uniformly distributed on $[0, 1]^d$. Since in practice this is rarely the case, the estimate provides only a rough indication of the upper scale $j_{\max}$. A more practical approach is to select $j_{\max}$ which guarantees a minimum number of observations hit the support of every basis function $\psi_{j_{\max},k}$ that covers the data range.

A fast and efficient structure selection approach can be implemented using the Orthogonal Forward Regression algorithm (OFR) (Chen et al., 1989). This is one of the best solutions next to the optimal algorithm that would require testing *all* possible combinations and for that reason is computationally prohibitive.

The least-squares based OFR algorithm involves a stepwise orthogonalisation of the regressors and a forward selection of the relevant terms in (26) based on the Error Reduction Ratio criterion (ERR) (Billings et al., 1988). The algorithm provides the optimum least-squares estimate of the parameter vector $\theta = \{\theta_k\}$.

If $\mathcal{G} = \{g_j\}_{j=1}^M$ is the candidate regressor set, the orthogonal term selection can be summarised as follows:

- Step 1

$$I_1 = I_M = \{1, ..., M\} \tag{28}$$

$$w_i(t) = g_i(t), \quad \hat{b}_i = \frac{w_i^T y}{w_i^T w_i} \tag{29}$$

$$l_1 = \arg\max_{i \in I_1} \left(\hat{b}_i \frac{w_i^T w_i}{y^T y} \right) = \arg\max_{i \in I_1} (\mathrm{err}_i) \tag{30}$$

$$w_1^0 = w_{l_1}, \quad c_1^0 = \frac{{w_1^0}^T y}{{w_1^0}^T w_1^0} \tag{31}$$

$$\alpha_{1,1} = 1 \tag{32}$$

Equation (30) shows that the basis function which achieves the best ERR score i.e. contributes to the largest extent to the reduction in the modelling error, is selected first from the candidate set.

- **Step** j, $j > 1$

$$I_j = I_{j-1} \backslash \{l_{j-1}\} \tag{33}$$

$$w_i(t) = g_i(t) - \sum_{k=1}^{j-1} \frac{{w_k^0}^T g_i}{{w_k^0}^T w_k^0} w_k^0, \tag{34}$$

$$Q_j = \left\{ \arg_{i \in I_j} (w_i^T w_i < \rho) \right\}, \quad I_j = I_j \backslash Q_j \tag{35}$$

$$l_j = \arg\max_{i \in I_1} \left(\hat{b}_i \frac{w_i^T w_i}{y^T y} \right) = \arg\max_{i \in I_1} (\mathrm{err}_i) \quad \forall i \in I_j \tag{36}$$

$$w_j^0 = w_{l_j}, \quad c_j^0 = \frac{{w_j^0}^T y}{{w_j^0}^T w_j^0} \tag{37}$$

$$\alpha_{k,j} = \frac{{w_k^0}^T g_{l_j}}{{w_k^0}^T w_k^0}, \quad k = 1, j-1 \tag{38}$$

The basis function candidate terms which are not chosen in the first steps are orthogonalised with respect to all previously selected basis functions. Because of the orthogonality the jth basis function can be selected in the same way as the first. In equation (37) w_j^0 is the j-th orthogonal model term selected and c_j^0 is the corresponding weight (parameter) associated with the orthogonalised regressor.

Any numerical ill conditioning is avoided by eliminating the candidate basis functions for which $w_i^T w_i$ is less than a predetermined threshold (35).

The procedure is terminated at the M_s step when

$$1 - \sum_{i=1}^{M_s} \mathrm{err}_i < \rho_e \tag{39}$$

where ρ_e is a desired error tolerance.

The wavelet basis parameters are calculated from the following equation:

$$\begin{bmatrix} c_{l_1} \\ c_{l_2} \\ \vdots \\ c_{l_{M_s}} \end{bmatrix} = \begin{bmatrix} 1 & \alpha_{1,2} & \cdots & \alpha_{1,M_s} \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \vdots & 1 & \alpha_{M_s-1,M_s} \\ 0 & 0 & 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} c_1^0 \\ c_2^0 \\ \vdots \\ c_{M_s}^0 \end{bmatrix} \tag{40}$$

Further improvements of this algorithm, due to Zhu and Billings (1995), provide significant reductions in the computation time.

4.3 Some Implementation Issues

The multi-variable multiresolution wavelet basis was defined in Section 3.1 using the tensor-product implementation. This method of constructing multi-variable basis functions is usually associated with the so called curse of dimensionality, that is, for large input dimensions the number of candidate regressors can be quite high. This does not automatically mean that the final model needs to have more parameters that a polynomial model for example. As noted in Section 3 the wavelets provide lacunary series expansions for most nonlinear functions. The key to finding this economical representation is the implementation of a good model sequencing strategy and the use of an efficient basis selection algorithm.

The main issues that have to be addressed in practice are related to the computational resources needed to implement and store the candidate regressor set and perform structure selection.

A solution to reduce the amount of memory used to store the regressors is to permanently keep in the memory of the computer only the regressors which correspond to the scalar basis functions. The regressors corresponding to multi-dimensional basis functions can be computed fast during structure selection using only regressors which are permanently stored.

Assuming for example that in equation (34) $g_i(u(t-1), y(t-1)) = \phi(u(t-1))\psi(y(t-1))$, it follows that

$$w_i(t) = (\phi\psi) - \sum_{k=1}^{j-1} \frac{w_k^{0T}(\phi\psi)}{w_k^{0T} w_k^0} w_k^0, \tag{41}$$

can be computed during selection using $\phi(u(t-1))$ and $\psi(y(t-1))$ so there is no need to store g_i. The computation of the additional tensor product $\phi(x)\psi(y)$ during selection has little impact on the overall performance of the OFR routine. This however, significantly reduces the amount of memory required for storing the candidate regressors.

4.4 The Stochastic Model

The stochastic perturbations due to unobserved noise can be modelled as additional input variables in the NARMAX model (4). Because the stochastic variable $e(t)$ cannot be measured directly, the unobserved noise sequence has to be estimated during identification using a prediction error method.

The sequence $e(t)$ is assumed to be independent, bounded and uncorrelated with the input u. The following prediction error equations

$$\varepsilon(t) = y(t) - \hat{f}\left(y(t-1),\ ...,\ u(t-1),...,\varepsilon(t-1),...\right) \tag{42}$$

provide an estimate of the stochastic variable vector $e(t)$. Equation (42) defines a stochastic dynamical system with inputs u and y and outputs the prediction error $\varepsilon(t)$. The identification problem can be formulated in this context as the determination of the nonlinear mapping f such that $\varepsilon(t)$ converges to $e(t)$ as $t \to \infty$. The prediction error is used during the estimation to implement an additional regression set which involves the prediction error variable $\varepsilon(t)$.

4.5 Model Validation

Model validation ensures that the final model is an accurate description of the original nonlinear system. For example, if the correct model subset and parameters have been estimated, the predicted noise sequence $\varepsilon(t)$ should be unpredictable, that is, uncorrelated with all linear and nonlinear combinations of past inputs and outputs.

This hypothesis can be tested in practice for SISO nonlinear systems using nonlinear correlation tests (Billings and Voon, 1986a)

$$\begin{cases} \Phi_{\varepsilon\,\varepsilon}(\tau) & \tau \neq 0 \\ \Phi_{u\,\varepsilon}(\tau) & \forall\tau \\ \Phi_{\varepsilon\,(\varepsilon u)}(\tau) & \tau \geq 0 \\ \Phi_{u^2\,\varepsilon}(\tau) & \forall\tau \\ \Phi_{u^2\,\varepsilon^2}(\tau) & \forall\tau \end{cases} \tag{43}$$

The null hypothesis, that the data was generated by the model, is accepted if the correlation functions are within the 95% confidence intervals, which for large N are approximately $\pm 1.96/\sqrt{N}$. These correlation tests can be extended to validate MIMO nonlinear models

Alternative correlation tests, which exploit the information in the system outputs, were introduced to further enhance the efficiency of the correlation-based validation approach (Billings and Zhu, 1994b).

$$\begin{cases} \Phi_{\varepsilon^2\,(y\varepsilon)}(\tau) & \tau \neq 0 \\ \Phi_{u^2\,\eta}(\tau) & \forall\tau \end{cases} \tag{44}$$

These new correlation tests have also been extended for MIMO systems (Billings and Zhu, 1994a).

The correlation tests however may not always reflect discrepancies between the observed dynamical behaviour and the model dynamics. Additional evidence about the adequacy of the model is provided by simulating the model and comparing the model predicted dynamics with the data available, especially the data not used in the identification.

The (deterministic) model predicted output of the model is defined as

$$\hat{y}(t) = \hat{f}(\hat{y}(t-1),...,\hat{y}(t-n_y),u(t-1),...,u(t-n_u),0,...,0) \tag{45}$$

This can be used to calculate the model prediction error

$$\hat{e}(t) = y(t) - \hat{y}(t) \tag{46}$$

In practice the data set available $\mathcal{D}$ is divided into two datasets $\mathcal{D}_1$ and $\mathcal{D}_2$ such that

$$\mathcal{D}_2 \bigcap \mathcal{D}_2 = \emptyset \tag{47}$$

The model is estimated usually using only the first data set $\mathcal{D}_1$ also known as the estimation set. The models found to be a best fit for $\mathcal{D}_1$ are then tested using the second data set $\mathcal{D}_2$, known as the validation or test set. The model which performs best with respect to both data sets is selected as the final model. This provides a practical means to test the null hypothesis, that the entire data

$$\mathcal{D} = \mathcal{D}_1 \bigcup \mathcal{D}_2 \tag{48}$$

was generated by the model. Note that comparing the model predicted output with the data provides a much more severe test than the commonly used one-step-ahead predictions which are computed by replacing $\hat{y}(\cdot)$ (previous model output) with the actual measurement $y(\cdot)$ on the right-hand side of equation (45).

4.6 The Iterative Identification Procedure

This section sumarises the complete iterative procedure for identifying a wavelet multiresolution model for a linear or nonlinear dynamical system.

- **Step** 1 Select n_y, n_u and n_e. Because the noise sequence is not available at this stage the model structure is defined using only the deterministic process variables

$$\{y(t-1), ..., y(t-n_y), u(t-1), ..., u(t-n_u)\}. \tag{49}$$

  For each process variable initialise the following parameters:

  - The order of B-spline wavelet and scaling functions used to implement the model
  - The starting resolution (scale) j
  - The number of resolutions (scales) N_s

  Usually the model is initialised with $m(i) = 2$ and $N_s(i) = 1$ in order to try to keep it as simple as possible. The model however can subsequently be refined if necessary. Selecting the starting resolution level can be simplified if the input/output data is normalised first to the unit interval $[0, 1]$. In this case the starting resolution $j(i) = 0$ can be chosen for all sub-models. If the data is not normalised, $j(i)$ takes into account the range of each input and output. The initial model structure $\mathcal{M}_1$ is a superposition of one-dimensional sub-models corresponding to each lagged input and output variable.

- **Step** 2

Implement the wavelet model set $\mathcal{G}_k$ corresponding to the current model structure $\mathcal{M}_k(n_u, n_y, n_\epsilon$
For $k = 1$, the candidate regressor set consists of scalar basis functions only.

Because the B-spline wavelet and scaling functions are compactly supported only a finite number of basis functions, that have data points inside their support, have to be considered

$$\mathcal{G}_i = \{g_k \,|\, (\exists)\,(y(t), u(t)) \in \mathcal{D}\text{such that } g_k(y(t), u(t)) \neq 0\} \tag{50}$$

where $\mathcal{D}$ is the data set used in identification.

- **Step 3**

  Select a subset wavelet model $\hat{f}^{(0)}$ from the current model set $\mathcal{G}_i$ and compute the initial one-step-ahead prediction error sequence

$$\varepsilon^0(t) = y(t) - f^{(0)}(t) \tag{51}$$

- **Step 4**

  Augment $\mathcal{G}_i$ with regressors corresponding to the one-step-ahead prediction errors. A new subset wavelet model is selected from this extended model set and used to generate a fresh residual sequence. This will be used to update the model iteratively. At the jth iteration the subset model $\hat{f}^{(j)}(t)$ is selected from the extended model set and this gives rise to the error $\{\varepsilon^j(t)\}$.

  Without updating the model further, the residual sequence and the parameters are re-evaluated. Typically 10 iterations are sufficient to achieve convergence of the parameter estimates.

- **Step 5**

  Apply model validity tests to evaluate the model.

If no valid models are found, the model structure is refined by

- increasing the number of scales $N_s(i)$ (resolutions).
- replacing two or more sub-models by a sub-model of higher dimension.
- increasing the order $m(i)$ of the basis functions in the current approximation.

After each refinement, the algorithm is restarted at **Step 2**.

5 Applications

The identification procedure, outlined in the previous sections has been successfully applied in the identification of several real life nonlinear dynamical systems. Two practical applications of the wavelet multiresolution models are presented in the following sections to illustrate the new algorithm.

5.1 A Liquid Level System

This system has been previously analysed in a number of papers (Billings and Voon, 1986b; Chen et al., 1990; Chen and Billings, 1988). The data was sampled from a large pilot scale liquid level system using a zero mean gaussian signal as input. The system consists of a DC water pump feeding a conical flask connected to a square tank. The input is the voltage to the pump motor and the output is the water level in the conical flask. A description of this process can be found in Billings and Voon (1986).

The data set consisted of 1000 input and output samples which were divided into a estimation set $\mathcal{D}_1$ consisting of 500 data samples and a test set $\mathcal{D}_2$ of 500 data samples. The input and output data used in identification are plotted in Figs 2 and 3.

Basis Functions	Estimates	$[ERR]_i$	Std. Dev.
$\phi_{0,-4}(u(t-1))$	$-0.39508E+0$	$0.77181E-4$	$0.73740E+0$
$\phi_{0,-1}(u(t-1))$	$0.59043E+0$	$0.24418E-3$	$0.82814E-1$
$\psi_{0,-5}(u(t-1))$	$0.49211E+0$	$0.46586E-5$	$0.19835E+0$
$\phi_{0,-4}(u(t-2))$	$1.3291E+0$	$0.82637E-5$	$0.34643E+0$
$\phi_{0,-1}(u(t-2))$	$-0.17719E+0$	$0.19013E-4$	$0.40267E-1$
$\psi_{0,-3}(u(t-2))$	$-0.28193E+0$	$0.85420E-5$	$0.77464E-1$
$\phi_{0,-1}(y(t-1))$	$2.6426E+0$	$0.99660E+0$	$0.17961E+0$
$\psi_{0,-2}(y(t-1))$	$-2.0207E+1$	$0.36509E-3$	$0.24980E+1$
$\psi_{1,-5}(y(t-1))$	$0.39676E-1$	$0.19588E-4$	$0.22742E-1$
$\psi_{1,-3}(y(t-1))$	$-0.21376E+0$	$0.40363E-5$	$0.85349E-1$
$\phi_{0,-4}(y(t-2))\psi_{0,-5}(y(t-3))$	$-1.6629E+2$	$0.18149E-3$	$0.40623E+2$
$\phi_{0,-1}(y(t-2))\phi_{0,-1}(y(t-3))$	$-2.0638E+0$	$0.20104E-2$	$0.70047E+0$
$\psi_{0,-5}(y(t-2))\phi_{0,-3}(y(t-3))$	$0.11734E+0$	$0.27393E-4$	$0.83137E-1$
$\psi_{0,-4}(y(t-2))\psi_{0,-5}(y(t-3))$	$1.0591E+1$	$0.78159E-4$	$0.10130E+2$
$\psi_{0,-3}(y(t-2))\phi_{0,-4}(y(t-3))$	$-1.7671E+2$	$0.39549E-5$	$0.11990E+2$
$\psi_{0,-3}(y(t-2))\phi_{0,-2}(y(t-3))$	$0.62131E+0$	$0.34950E-5$	$0.22855E+0$
$\psi_{0,-3}(y(t-2))\psi_{0,-4}(y(t-3))$	$5.0889E+0$	$0.27329E-5$	$0.63196E+0$
$\psi_{0,-2}(y(t-2))\phi_{0,-3}(y(t-3))$	$1.4245E+2$	$0.93664E-5$	$0.48995E+0$
$\psi_{0,-2}(y(t-2))\phi_{0,0)}(y(t-3))$	$2.0710E+2$	$0.22676E-4$	$0.14790E+1$
$\psi_{0,-2}(y(t-2))\psi_{0,-2}(y(t-3))$	$6.8221E+2$	$0.44247E-5$	$0.43642E+1$
$\sum_{i=1}^{49} a_i\xi_i(t)$	-	$9.7835E-5$[a]	-

Table 5.1

[a] Combined ERR contribution of the selected noise terms

The initial parameters for the model were $n_u = n_y = n_e = 3$, $m = 5$, initial scale $j(k) = 0$ and $N_s(k) = 1$ for all variables. To simplify the implementation of the candidate regressor set $\mathcal{G}$ the data was normalised on the interval $[0\,,1]$. Starting with a basic model structure composed only of one-dimensional additive submodels the iterative identification procedure described in Section 4.6 led to the identification of the model subset shown in Table 5.1. The final model structure tested was as follows

$$y(t) = \sum_{i=1}^{3} f_i^1(u(t-i)) + f_4^1(y(t-1)) + f_5^2(y(t-2), y(t-3)) + f_{noise}(e(t-1), ..., e(t-5)) \quad (52)$$

where f_{noise} is the noise model.

The model predicted output, the one-step-ahead prediction error and the model prediction error are shown in Figs 4a,b,c. In Fig. 5 is shown how the model predicted output compares with the test data set. The correlation tests are shown in Figs.(6a,b,c,d,e). The model performance is comparable with that of the RBF model identified by Chen *et al* (1990) which required 34 parameters (not including the parameters needed to specify the centers). The wavelet model identified here (the deterministic part) has only 20 parameters. The noise model, which normally is not used in simulation consisted of 49 additional terms.

5.2 A Civil Engineering Structure

The second example is related to the application of the wavelet identification methodology in the detection and localisation of defects in civil engineering structures. The system analysed in this case was a large reinforced concrete beam from a bridge. A persistent excitation was applied at one end of the structure. The response was recorded using five accelerometers equally spaced along the beam. The experimental data used here to illustrate the proposed identification method consists of two sets of measurements recorded by two of the accelerometers mounted on the bridge. 500 data samples of the measured signals, assumed to be the input and respectively the output of a nonlinear dynamical system, are plotted in Figs. (7) and (8).

The initial parameters for the model were $n_u = 10$, $n_y = n_e = 5$, $m = 3$, initial scale $j(k) = 0$ and $N_s(k) = 1$ for all variables. The initial model structure

$$y(t) = \sum_{i=1}^{10} f_i^1(u(t-i)) + \sum_{i=1}^{5} f_{i+10}^1(y(t-i)) + f_{noise}(e(t-1), ..., e(t-10)) \quad (53)$$

led to the identification of a model consisting of 20 basis functions detailed in Table (5.2). An additional noise model which included 29 terms was also identified. The model was found to be sufficiently accurate and no further refinements were carried out.

The model predicted output, the one-step-ahead prediction and the model prediction errors which are plotted in Figs (9a,b,c) show very good agreement between the model and the data. The model has been simulated using data not used in identification. The model output and the original data are plotted in Fig. (10a) for comparison. The model predicted error over the same data interval is shown in Fig. (10b) while the correlation tests are shown in Figs (11).

Basis Functions	Estimates	$[ERR]_i$	Std. Dev.
$\phi_{0,0}(u(t-1))$	0.1386570	$0.13258E-4$	$0.14369E-1$
$\phi_{0,0}(u(t-2))$	0.1987395	$0.71995E-3$	$0.18328E-1$
$\psi_{0,-3}(u(t-2))$	−0.1660639	$0.25415E-4$	$0.26766E-1$
$\phi_{0,-2}(u(t-3))$	−0.1657839	$0.10865E-1$	$0.15813E-1$
$\phi_{0,0}(u(t-3))$	0.1176670	$0.93189E-4$	$0.18249E-1$
$\phi_{0,0}(u(t-4))$	0.2096010	$0.74452E-4$	$0.21499E-1$
$\phi_{0,-2}(u(t-5))$	−0.4773314	$0.78793E-3$	$0.60412E+0$
$\phi_{0,0}(u(t-5))$	0.3101863	$0.36144E-3$	$0.17001E-1$
$\phi_{0,-2}(u(t-6))$	0.0537393	$0.21384E-4$	$0.20849E-1$
$\phi_{0,-2}(u(t-7))$	0.1333856	$0.35242E-2$	$0.16553E-1$
$\phi_{0,-1}(u(t-7))$	0.1692944	$0.34938E-4$	$0.21593E-1$
$\phi_{0,-2}(u(t-8))$	−0.2975717	$0.27241E-3$	$0.18101E-1$
$\phi_{0,-1}(u(t-8))$	−0.0986084	$0.25339E-4$	$0.13686E-1$
$\phi_{0,-2}(u(t-9))$	−0.2060172	$0.83471E-3$	$0.16063E-1$
$\phi_{0,0}(u(t-9))$	0.3650497	$0.38811E-3$	$0.19404E-1$
$\phi_{0,-1}(y(t-1))$	0.4331915	$0.68501E-1$	$0.18041E-1$
$\phi_{0,0}(y(t-1))$	0.5793608	$0.91283E+0$	$0.51700E-1$
$\phi_{0,-2}(y(t-2))$	−0.1394088	$0.24531E-4$	$0.19797E-1$
$\phi_{0,0}(y(t-2))$	0.3105835	$0.17448E-4$	$0.39293E-1$
$\phi_{0,0}(y(t-5))$	0.0478836	$0.35880E-4$	$0.17698E-1$
$\sum_{i=1}^{29} a_i\xi_i(t)$	-	$2.0686E-4^a$	-

Table 5.1

[a]Combined ERR contribution of the selected noise terms

6 Conclusions

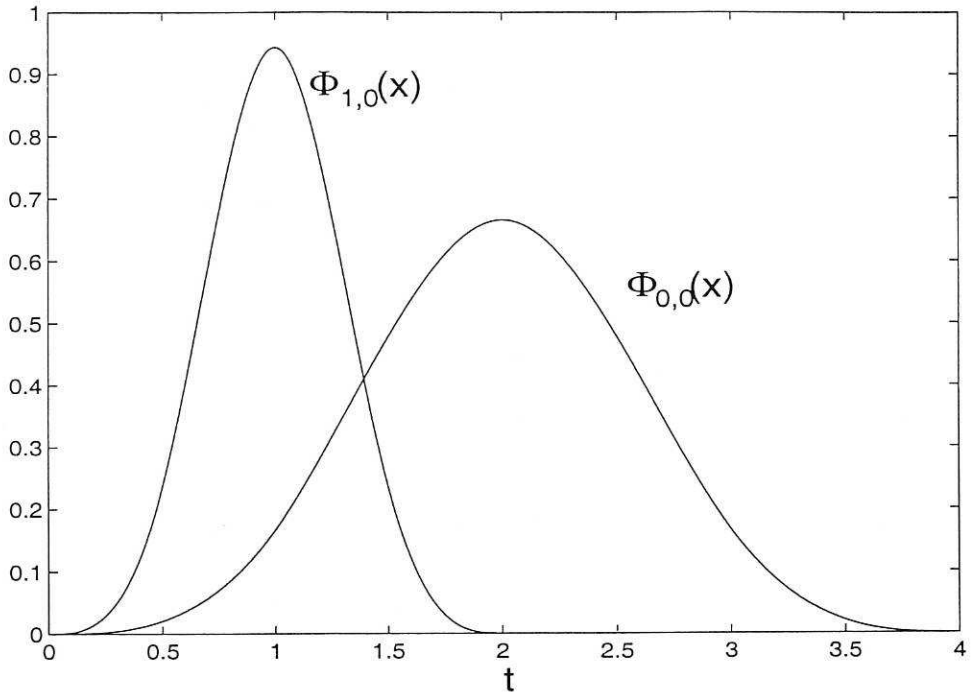
A new identification methodology, which makes use of wavelet multiresolution approximations to expand the unknown nonlinearity in the NARMAX input/output representation, has been introduced and applied in the identification of two real life nonlinear systems.

The particular semi-orthogonal wavelet multiresolution approximation structure used in this study, provides a powerful approximation tool for an extremely wide range of nonlinear behaviours. This model implementation is coupled with an iterative identification algorithm which uses a model sequencing approach to determine the model architecture, an efficient OFR algorithm for model subset selection and model validation tools to decide model adequacy. This approach has been successfully applied in the identification of two practical systems from noisy measurements. In one case, the present method is shown to lead to a simpler model that an

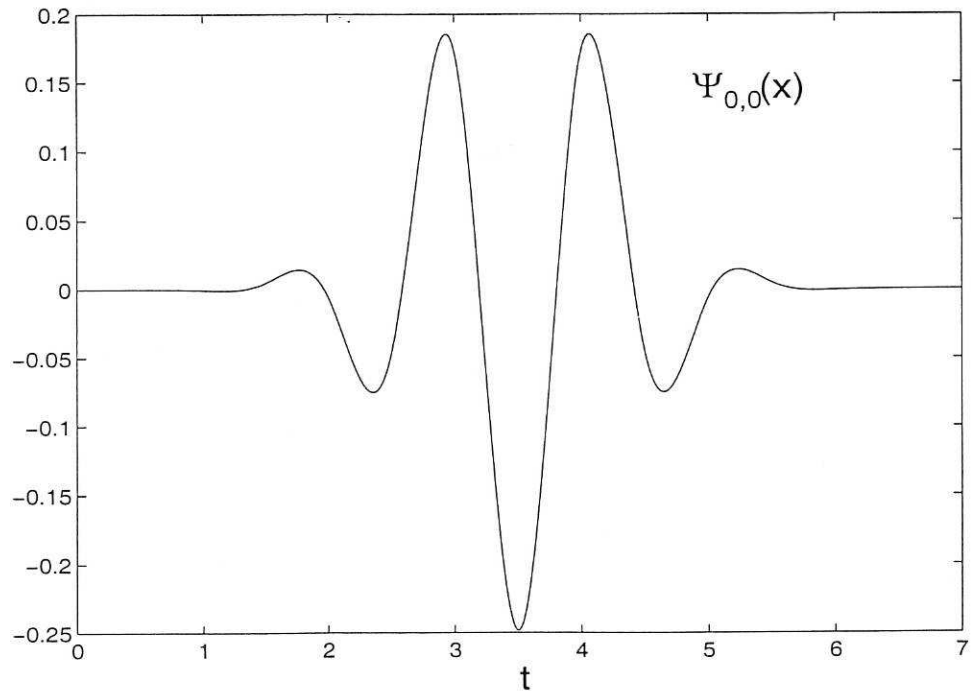
equivalent RBF model, previously estimated using the same data set.

7 Acknowledgment

The authors gratefully acknowledge that this research was supported by the UK Engineering and Physical Sciences Research Council. The authors would also like to thank Prof. F. Brancaleoni, University of Rome and Dr. C. Valente, University of Pescara, Italy, who peformed the experiments and the data collection on the civil engineering structure.



(a)



(b)

Figure 1: Examples of B-Spline (a) scaling and (b) wavelet functions

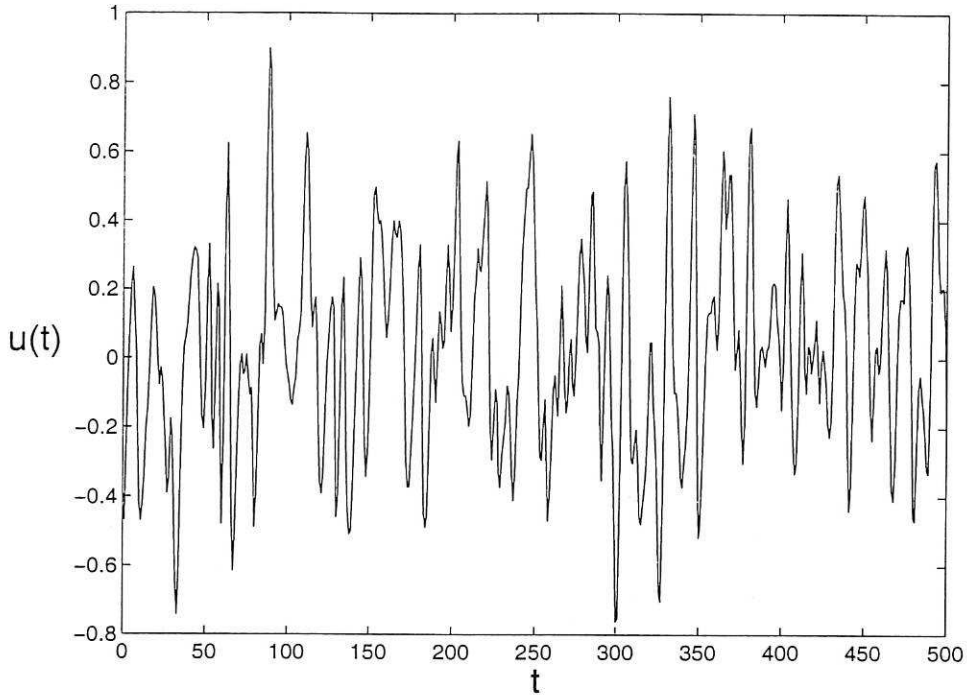


Figure 2: **Example 1**: Input data $u(t)$ - Voltage measurements from the pump DC motor

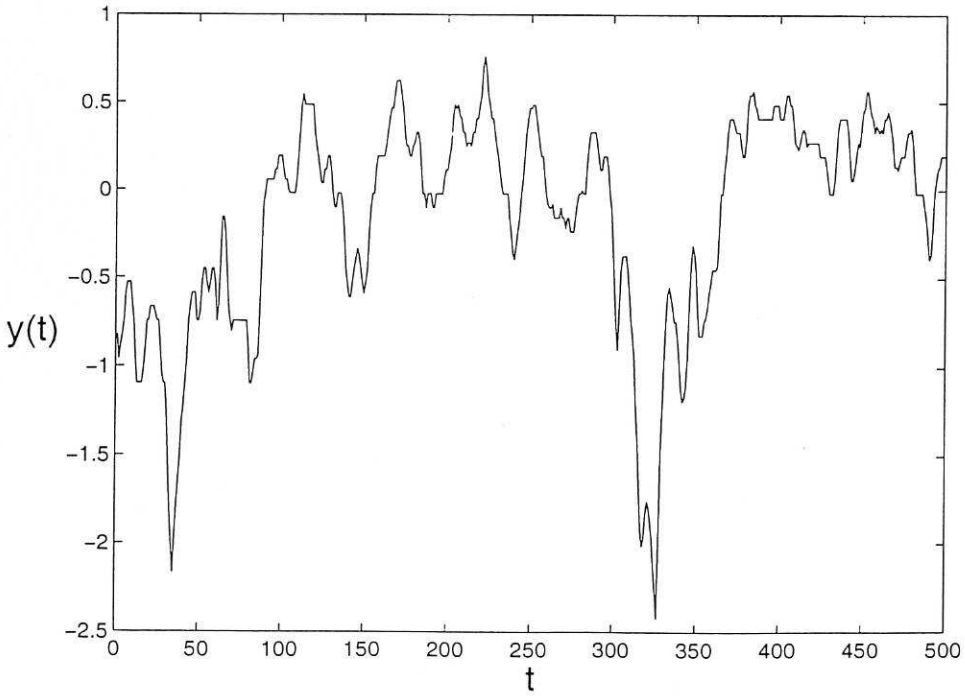


Figure 3: **Example 1**: Output Data $y(t)$-Water level measurements from the conical flask

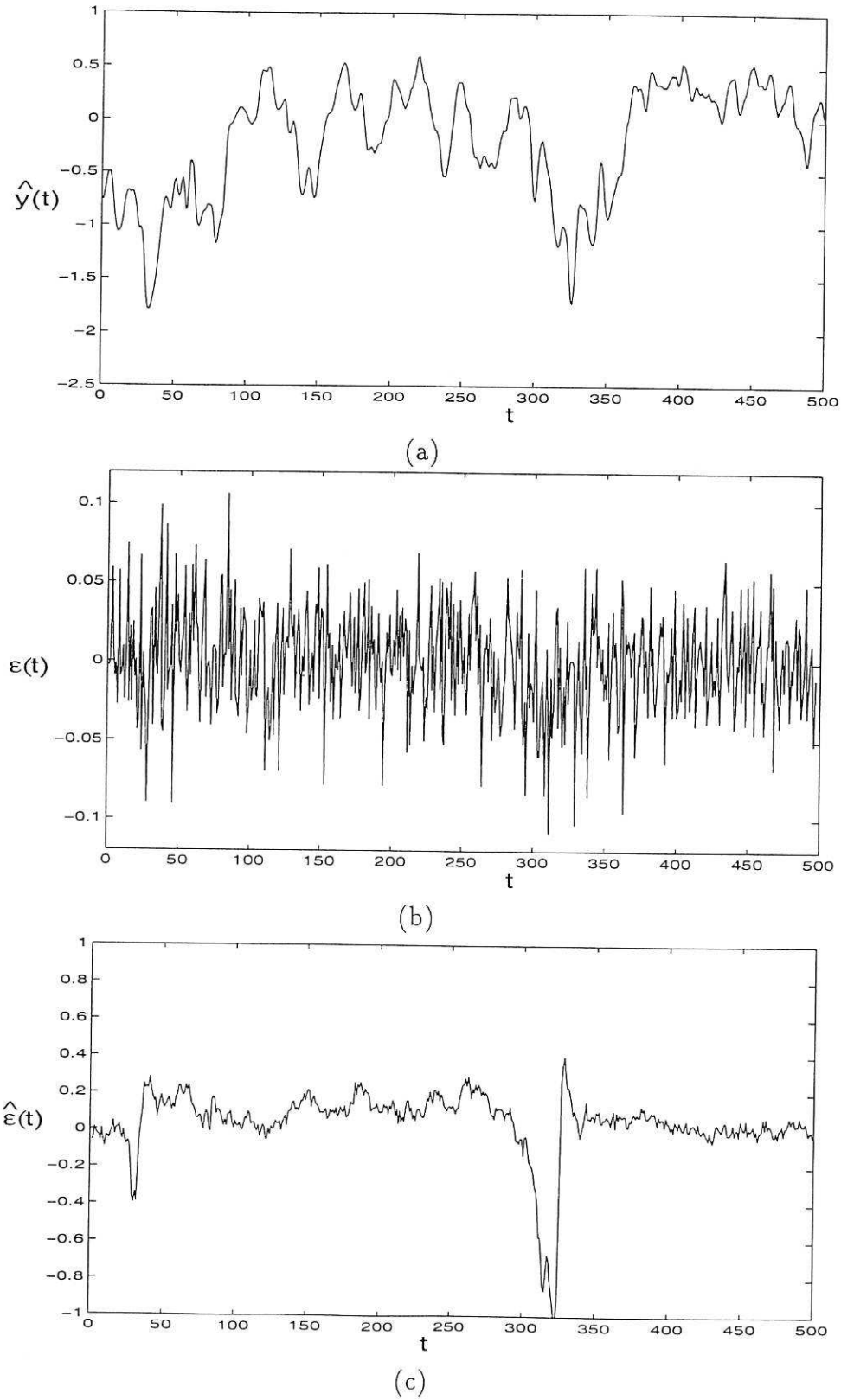
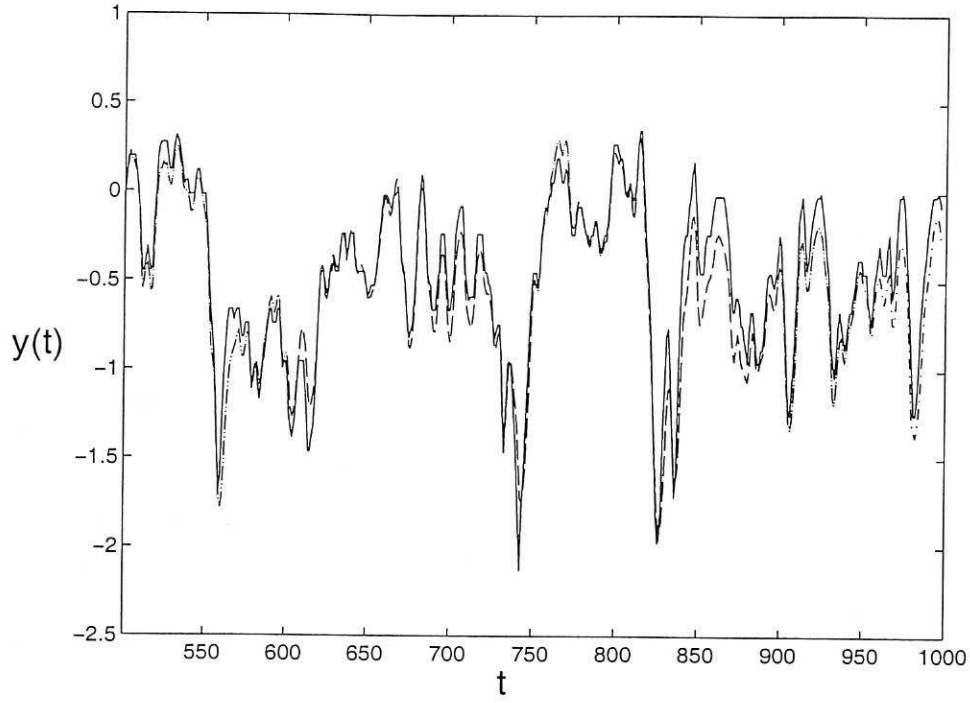
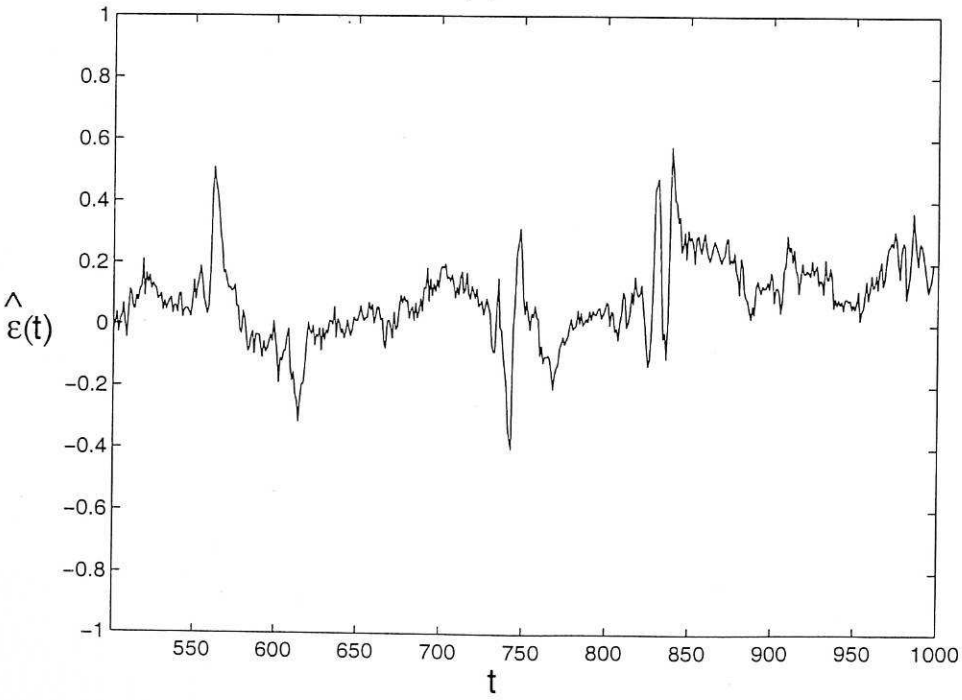


Figure 4: **Example 1**: Model responses: (a) Model Predicted Output $\hat{y}(t)$, (b) One Step Ahead Prediction Error $\varepsilon(t)$ and (c) Model Prediction Error $\hat{\varepsilon}(t)$ over the estimation interval



(a)



(b)

Figure 5: **Example 1**: (a) Measured Output (cont.) and Model Predicted Output (dash-dot) over the test set and (b) the Model Prediction Error over the test set

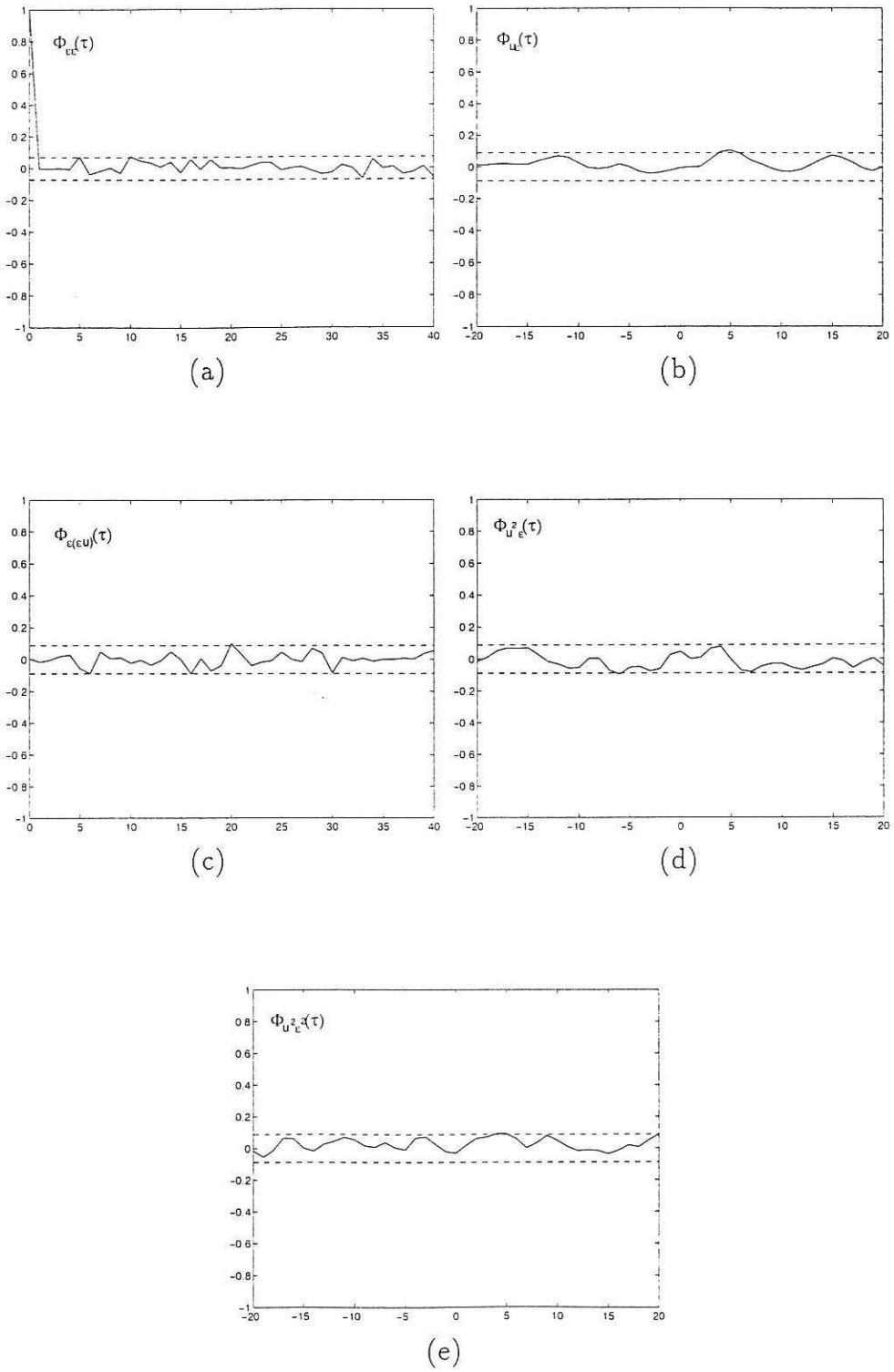


Figure 6: **Example 1**: Correlation Tests - (a) $\Phi_{\varepsilon,\varepsilon}(\tau)$, (b) $\Phi_{u,\varepsilon}(\tau)$, (c) $\Phi_{\varepsilon,(\varepsilon u)}(\tau)$ (d) $\Phi_{\varepsilon,(u^2\varepsilon)(e)}(\tau)$, (e) $\Phi_{\varepsilon,(u^2\varepsilon^2)}(\tau)$. Dashed line 95% confidence interval

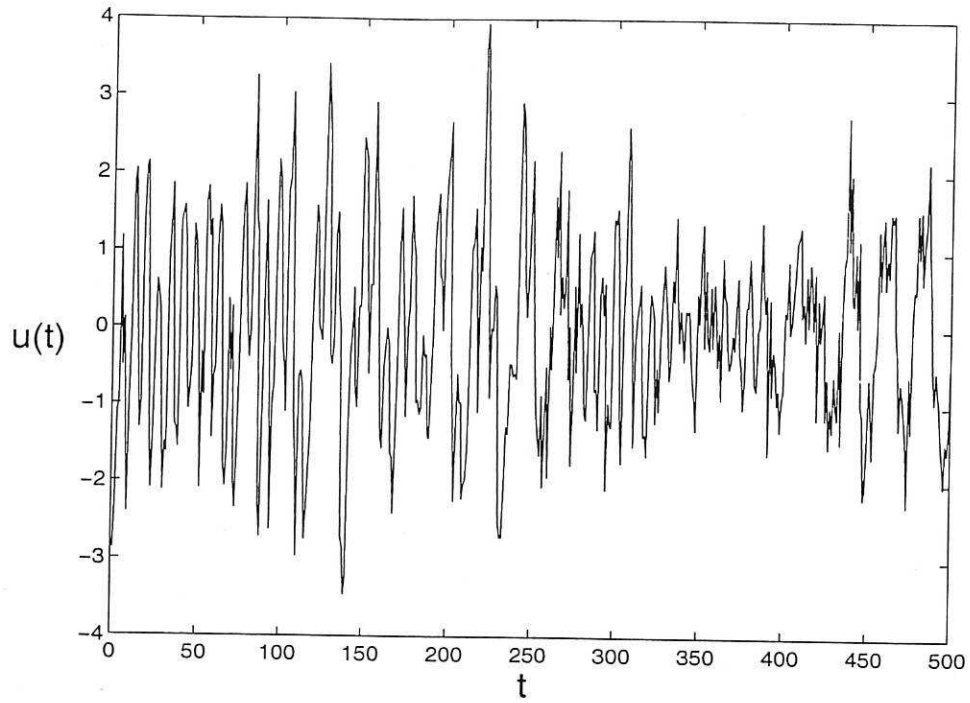


Figure 7: **Example 2**: Input data $u(t)$ - Accelerometer (A) measurements

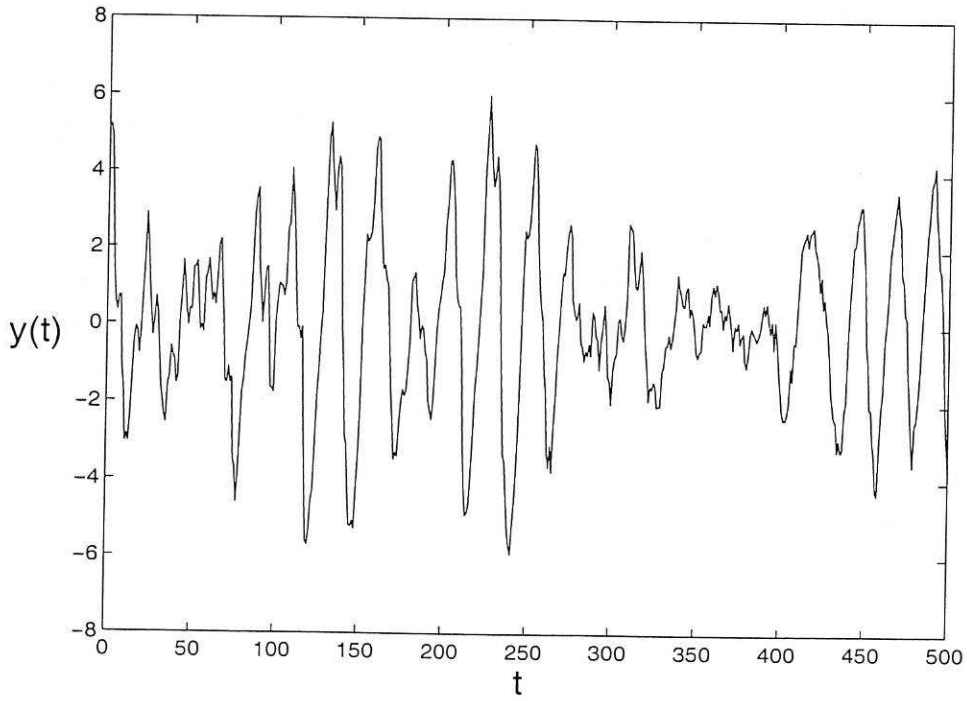


Figure 8: **Example 2**: Output Data $y(t)$-Accelerometer (B) Measurements

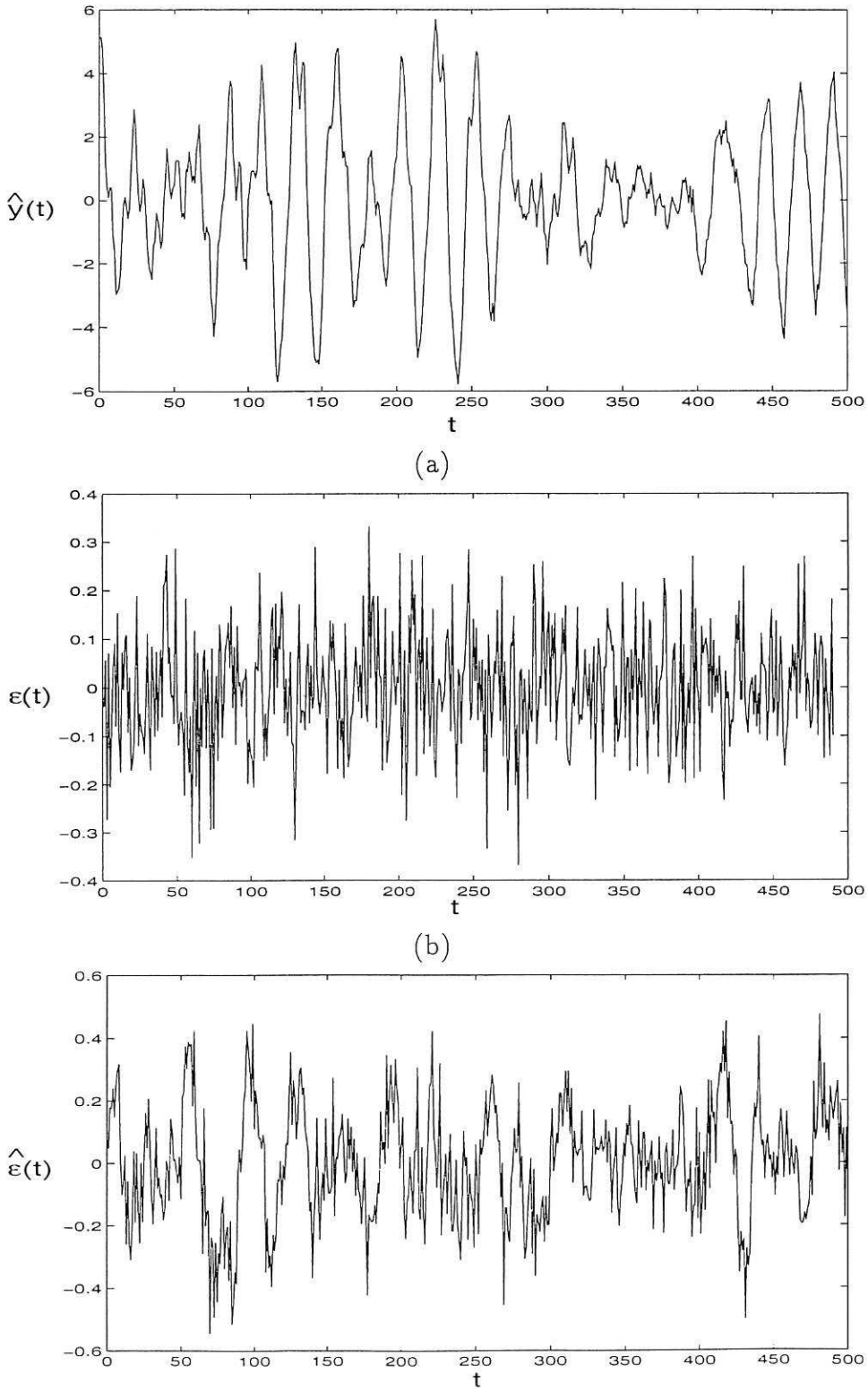
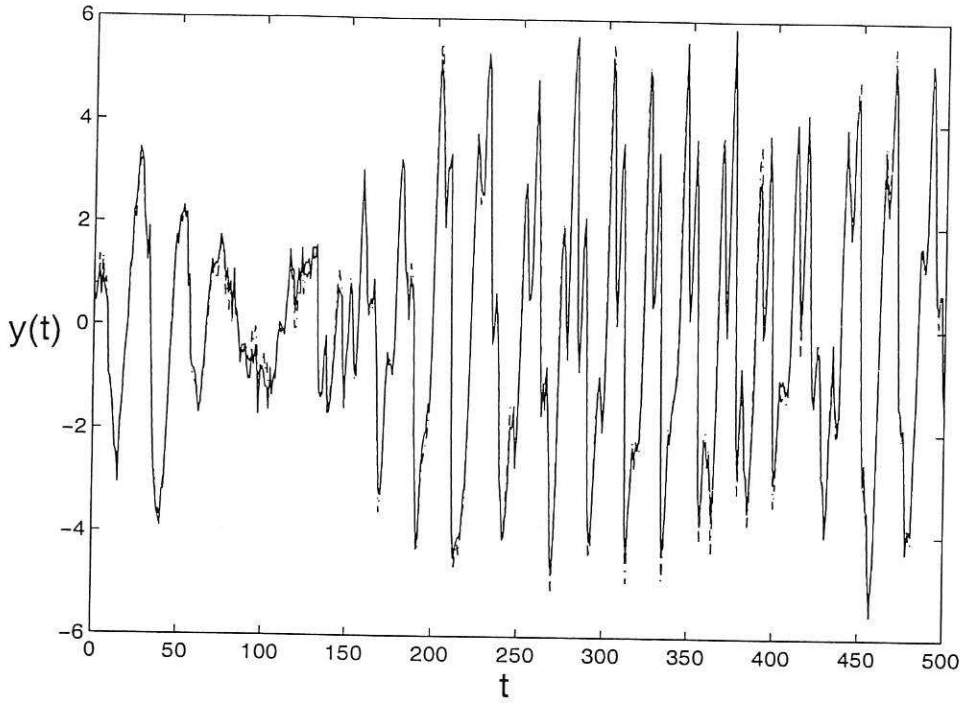
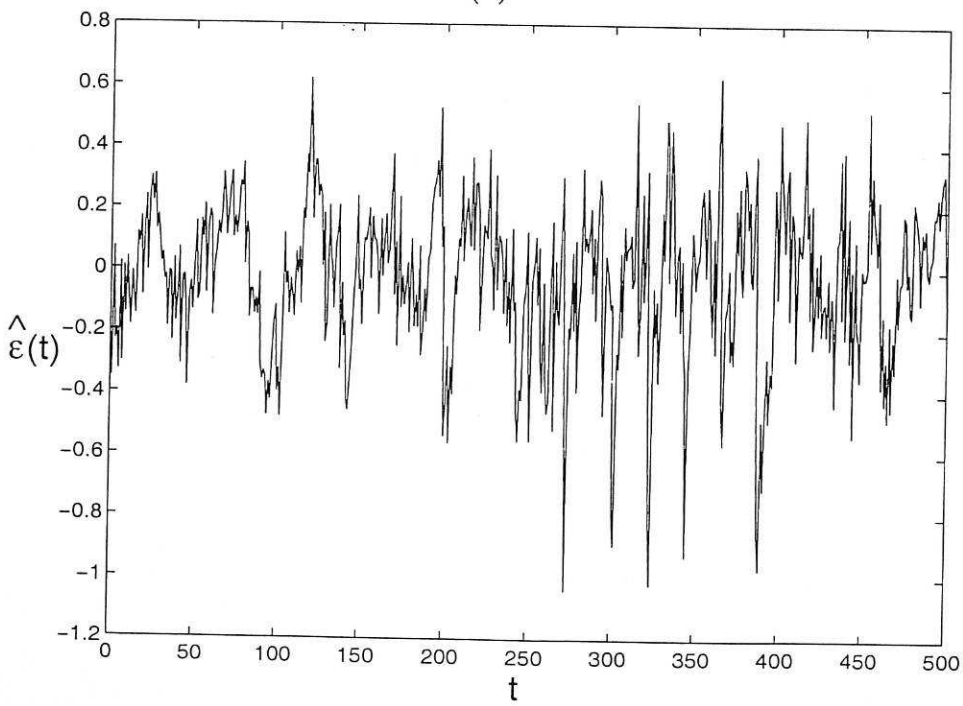


Figure 9: **Example 2**: Model responses: (a) Model Predicted Output $\hat{y}(t)$, (b) One Step Ahead Prediction Error $\varepsilon(t)$ and (c) Model Prediction Error $\hat{\varepsilon}(t)$ over the estimation interval



(a)



(b)

Figure 10: **Example 2**: (a) Measured Output (cont.) superimposed on the Model Predicted Output (dash-dot) and (b) the Model Prediction Error over the test interval

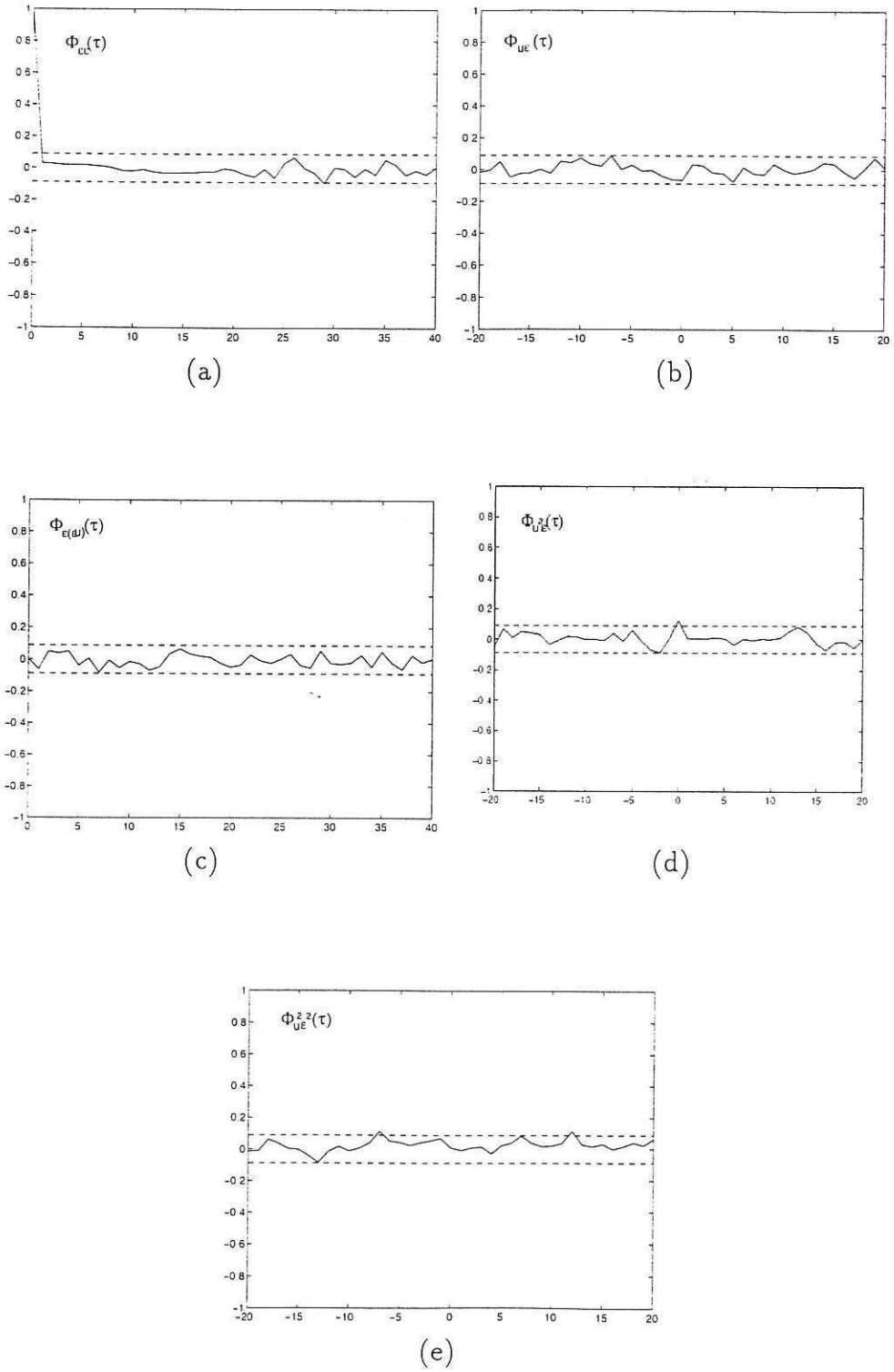


Figure 11: **Example 2**: Correlation Tests - (a) $\Phi_{\varepsilon,\varepsilon}(\tau)$, (b) $\Phi_{u,\varepsilon}(\tau)$, (c) $\Phi_{\varepsilon,(\varepsilon u)}(\tau)$, (d) $\Phi_{\varepsilon,(u^2\varepsilon)}(\tau)$ (e) $\Phi_{\varepsilon,(u^2\varepsilon^2)}(\tau)$. Dashed line 95% confidence interval

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. on Automatic Control*, 19:716–723.

Billings, S. A. (1980). Identification of nonlinear systems - a survey. *IEE Proceedings Part D*, 127(6):272–285.

Billings, S. A. and Chen, S. (1989). Extended model set, global data and threshold model identification of severely nonlinear systems. *Int. J. Control*, 50(5):1897–1923.

Billings, S. A., Chen, S., and Kronenberg, M. J. (1988). Identification of MIMO non-linear systems using a forward-regression orthogonal estimator. *Int. J. Control*, 49:2157–2189.

Billings, S. A. and Coca, D. (1999). Discrete wavelet models for identification and qualitative analysis of chaotic systems. *Int. J. of Bifurcation and Chaos*, 9(7):1263–1284.

Billings, S. A. and Voon, W. S. F. (1983). Structure detection and model validity tests in the identification of nonlinear systems. *Proceedings of the Institution of Electrical Engineers. Part D*, 130:193–199.

Billings, S. A. and Voon, W. S. F. (1986a). Correlation based model validity tests for nonlinear models. *Int. J. Control*, 44:235–244.

Billings, S. A. and Voon, W. S. F. (1986b). A prediction-error and stepwise regression estimation algorithm for non-linear systems. *Int. J. Control*, 44(3):803–822.

Billings, S. A. and Zhu, Q. M. (1994a). Model Validation Tests for Multivariable Nonlinear Models including Neural Networks. *Int. J. Control*, 62(4):749–766.

Billings, S. A. and Zhu, Q. M. (1994b). Nonlinear Model Validation using Correlation Tests. *Int. J. Control*, 60(6):1107–1120.

Chen, S. and Billings, S. A. (1988). Recursive maximum likelihood identification of a nonlinear output-affine model. *Int. J. Control*, 48:1605–1629.

Chen, S. and Billings, S. A. (1989). Representations of non-linear systems: the NARMAX model. *Int. J. Control*, 49(3):1013–1032.

Chen, S., Billings, S. A., Cowan, C. F. N., and Grant, P. M. (1990). Practical identification of narmax models using radial basis functions. *Int. J. Control*, 52:1327–1350.

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

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

Chui, C. K. and Wang, J. Z. (1992). On compactly supported spline wavelets and a duality principle. *IEEE Trans. Amer. Math. Soc.*, 330:903–915.

Coca, D. and Billings, S. A. (1997). Continuous-time system identification for linear and nonlinear systems using wavelet decompositions. *Int. J. of Bifurcation and Chaos*, 7(1):87–96.

Daubechies, I. (1993). Wavelet transforms and orthonormal wavelet bases. *Proceedings of Symposia in Applied Mathematics*, 47.

de Boor, C. (1978). *A practical guide to spline.* Springer Verlag.

DeVore, R. A., Jawerth, B., and Popov, V. (1992). Compression of wavelet decompositions. *American Journal of Mathematics*, 114:737–785.

Donoho, D. L. (1993a). Nonlinear wavelet methods for recovery of signals, densities and spectra from indirect and noisy data. In Daubechies, I., editor, *Proceedings of the Symposia in Apllied Mathematics*, pages 159–205. American Mathematical Society.

Donoho, D. L. (1993b). Wavelet shrinkage and W.V.D: A 10-minute tour. In Meyer, Y. and S.Roques, editors, *Progress in Wavelet Analysis and Applications*, pages 109–128. Editions Frontieres, Gif-sur-Yvette, France.

Eykhoff, P. (1974). *System identification.* Wiley, New York.

Friedman, J. H. and Stuetzle, W. (1981). Projection pursuit regression. *J. Amer. Stat. Assoc.*, 76:817–823.

Haber, R. and Unbehauen, H. (1990). Structure identification for nonlinear dynamic systems-A survey on input/output approaches. *Automatica*, 26(4):651–677.

Juditski, A., Hjalmarsson, H., Beneviste, B., Deylon, H., Ljung, L., Sjöberg, J., and Zhang, Q. (1995). Nonlinear black-box models in system identification: Mathematical foundations. *Automatica*, 31(12):1725–1750.

Kavli, T. (1993). ASMOD - An algorithm for adaptive spline modelling of observational data. *Int. J. Control*, 58(4):947–967.

Kavli, T. and Weyer, E. (1994). ASMOD (Adaptive Modelling of Observational Data) some theoretical and experimental results. *Proceedings of the IEEE Coll. on Advances in Neural Networks for Control & Systems, Berlin.*

Leontaritis, I. J. and Billings, S. A. (1985a). Input-output parametric models for non-linear systems - Part II: Stochastic non-linear systems. *Int. J. Control*, 41:329–344.

Leontaritis, I. J. and Billings, S. A. (1985b). Input-output parametric models for non-linear systems- Part I: Deterministic non-linear systems. *Int. J. Control*, 41:303–328.

Ljung, L. (1987). *System identification: Theory for the user.* Wiley, New York.

Mallat, S. G. (1989). A theory of multiresolution signal decomposition; The wavelet representation. *IEEE Pattern Anal. and Machine Intelligence*, 11:674–693.

Mallat, S. G. and Zhang, Z. (1993). Matching pursuit with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415.

Mees, A. I. (1993). Parsimonious dynamical reconstruction. *Int. J. Bifurcation and Chaos*, 3(3):669–675.

Pearson, R. K. and Ogunnaike, B. A. (1997). Nonlinear process identification. In Henson, M. A. and Seborg, D. E., editors, *Nonlinear Process Control*, chapter 2. Prentice Hall, Englewood Cliffs.

Sjöberg, J., Zhang, Q., Ljung, L., Beneviste, B., Deylon, H., Glorennec, H., Hjalmarsson, H., and Juditski, A. (1995). Nonlinear black-box modelling in system identification: A unified overview. *Automatica*, 31(12):1691–1724.

Söderström, T. and Stoica, P. (1989). *System identification.* Prentice Hall International, London.

Sureshbabu, N. and Farrell, J. A. (1999). Wavelet-based system identification for nonlinear control. *IEEE Transactions on Automatic Control*, 44(2):412–417.

Sweldens, W. and Piessens, R. (1994). Asymptotic error expansion of wavelet approximations of smooth functions II. *Numerische Mathematick (preprint).*

Zhang, Q. (1997). Using wavelet network in nonparametric estimation. *IEEE Trans. on Neural Network*, 8(2):227–236.

Zhang, Q. and Beneviste, A. (1992). Wavelet networks. *IEEE Trans. on Neural Networks*, 3:889–898.