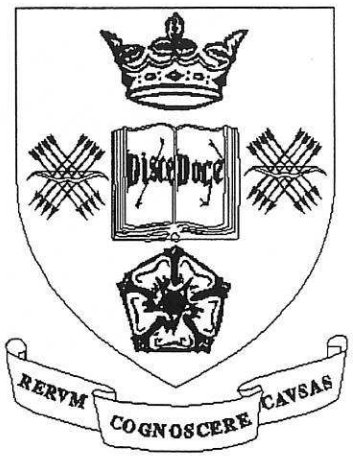
A Unified Wavelet-Based Modelling Framework for Nonlinear System Identification: the WANARX Model Structure

H. L. Wei, S. A. Billings

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



Research Report No. 839

June 2003

A Unified Wavelet-Based Modelling Framework for Nonlinear System Identification: the WANARX Model Structure

H.L. Wei , S.A. Billings
Department of Automatic Control and Systems Engineering, University of Sheffield
Mappin Street, Sheffield, S1 3JD, UK

A new unified modelling framework based on the superposition of additive submodels, functional components, and wavelet decompositions is proposed for nonlinear system identification. A nonlinear model, which is often represented using a multivariate nonlinear function, is initially decomposed into a number of functional components via the well known analysis of variance (ANOVA) expression, which can be viewed as a special form of the NARX(*Nonlinear AutoRegressive with eXogenous* inputs) model for representing dynamic input-output systems. By expanding each functional component using wavelet decompositions including the regular lattice frame decomposition, wavelet series and multiresolution wavelet decompositions, the multivariate nonlinear model can then be converted into a linear-in-the-parameters problem, which can be solved using least-squares type methods. An efficient model structure determination approach based upon a forward orthogonal least squares (OLS) algorithm, which involves a stepwise orthogonalization of the regressors and a forward selection of the relevant model terms based on the error reduction ration (ERR), is employed to solve the linear-in-the-parameters problem in the present study. The new modelling structure is referred to as a Wavelet-based ANOVA decomposition of the NARX model or simply WANARX model, and can be applied to represent high-order and high dimensional nonlinear systems.

Keywords: Nonlinear system identification; NARX and NARMAX models; wavelets; orthogonal least squares

1. Introduction

The main task in mathematical modelling is to construct a mapping, which connects the inputs and outputs and reflects the relationship between these with an acceptable accuracy. In experimental data based modelling, known as system identification, the key problem is to construct a suitable model, which involves the smallest number of input variables (lagged inputs and lagged outputs for dynamical systems) and the simplest model structure containing the smallest number of adjustable parameters. For high dimensional systems, however, parsimony and accuracy are difficult to achieve simultaneously. Therefore, trade-offs between model parsimony, accuracy, and validity have to be considered.

A key problem in modelling high dimensional nonlinear systems is to develop efficient model construction procedures that overcome the curse-of-dimensionality. One approach for representing continuous functions of several variables is to describe multivariate functions as a superposition of a number of continuous functions with fewer variables. This is the essence of Hilbert's 13^{th} problem, which was resolved by Kolmogorov where it was concluded that every continuous function of several variables can be represented by the superposition of functions with only two variables (see, Gorban 1998 and the references therein). The problem of representing continuous functions of several variables by continuous functions of a single variable has also been solved and this can be expressed using Kolmogorov's theorem which states that: every continuous function of n variables defined in the standard n-dimensional cube can be represented in the following form:

$$f(x_1, x_2, \cdots, x_n) = \sum_{q=1}^{2n+1} g_q \left(\sum_{p=1}^{n} h_q^{(p)}(x_p) \right) = \sum_{q=1}^{2n+1} g_q \left(h_q^{(1)}(x_1) + h_q^{(2)}(x_2) + \cdots + h_q^{(n)}(x_n) \right) \quad (1)$$

where $g_q(\cdot)$ are some continuous functions which depend on the function f and $h_q^{(p)}(\cdot)$ are some continuous functions which are independent the function f. This theorem guarantees that every continuous function can be approximated by the operations of addition, multiplication, and superposition of a number of continuous single variable functions with arbitrary accuracy. This theorem, however, does not provide a solution for how to choose the additive functional components. Therefore it is not easily applicable in real system modelling.

Several applicable approaches have been proposed to realize the idea of representing multivariate functions using a superposition of a number of functions with fewer variables. The projection pursuit regression algorithm (Friedman 1981), radial basis function networks (Chen et al 1990, 1992), and multi-layer perceptron (MPL) architecture (Haykin 1994) are among these representations for multivariate functions. The existing strategies that attempt to approximate general functions in high dimensions are based on additive functional submodels including the polynomial NARMAX (Nonlinear AutoRegressive Moving Average with eXogenous inputs) representation introduced by Billings and Leontaritis (1982, 1985), the multivariate adaptive regression splines (MARS) introduced by Friedman(1991), and the adaptive spline modelling of observational data (ASMOD) introduced by Kavli (1993). The functional components can be arbitrary functions with fewer arguments and with global or local properties. Kernel functions, splines, polynomials and other basis functions can all be chosen as functional components (Hastie and Tibshirani 1990).

A multivariate nonlinear function can often be decomposed into a number of functional components via the well known functional analysis of variance (ANOVA) expansions

$$f(x_1, x_2, \cdots, x_n) = f_0 + \sum_{i=1}^{n} f_i(x_i) + \sum_{1 \le i < j \le n} f_{ij}(x_i, x_j) + \sum_{1 \le i < j < k \le n} f_{ijk}(x_i, x_j, x_k) + \cdots$$
$$+ \sum_{1 \le i_1 < \cdots < i_m \le n} f_{i_1 i_2 \cdots i_m}(x_{i_1}, x_{i_2}, \cdots, x_{i_m}) + \cdots + f_{12\cdots n}(x_1, x_2, \cdots, x_n) \quad (2)$$

where the first functional component f_0 is a constant to indicate the intrinsic varying trend; f_i, f_{ij}, $\cdots$, are univariate, bivariate, etc., functional components. The univariate functional components $f_i(x_i)$ represent the independent contribution to the system output that arises from the action of the ith variable x_i alone; the bivariate functional components $f_{ij}(x_i, x_j)$ represent the interacting contribution to the system output from the input variables x_i and x_j, etc. As that will be seen later, the ANOVA expansion (2) can be viewed as a special form of the NARX (Nonlinear AutoRegressive with eXogenous inputs) model for dynamic input and output systems.

Among almost all the functions used for the purpose of approximation, none has had such an impact and spurred so much interest as *wavelets*. Wavelet decompositions outperform many other approximation schemes and offer a flexible capability for approximating arbitrary functions. Wavelet basis functions have the property of localization in both time and frequency. Due to this inherent property, wavelet approximations provide the foundation for representing arbitrary functions economically, using just a small number of basis functions.

Wavelet algorithms (Coca and Billings 2001) process data at different scales or resolutions, which make wavelet representations more adaptive compared with other basis functions. Therefore, wavelet decompositions can be used to represent each functional component in the model (2).

In this paper, a new model structure which combines wavelets and the additive functional ANOVA decomposition of the NARX model, called the Wavelet-based ANOVA decomposition of the NARX model or simply WANARX, is introduced as a basis for nonlinear system identification. The wavelet decompositions, which have excellent approximation properties, are used to express each functional component. By expanding each functional component into wavelet decompositions, the multivariate nonlinear function can then be converted into a linear-in-the-parameters problem, which can be solved using least-squares type of methods. A stepwise forward least squares (OLS) algorithm, along with an error reduction ratio (ERR) index is used to select the significant model terms from a large number of candidate terms. Emphasis is concentrated on wavelet series and multiresolutoin decompositions in this study from the point of view of practical data analysis.

The rest of the paper is organised as follows. The paper starts with a description of the well-known NARMAX model in Section 2, and the additive functional ANOVA decomposition of the NARX model is introduced. In Section 3, some introductory material on wavelet decompositions including wavelet frame decompositions, wavelet series and wavelet multiresolution decompositions, which establish the foundation for the WANARX model, are described. Section 4 shows how to expand a WANARX model using the wavelet decompositions. Section 5 addresses system variable selection and model term detection problems. In Section 6, some practical issues associated with the implementation of the WANARX model are discussed. Two examples, one a simulated system and one based on real data relating to the magnetosphere, are given in Section 7 to demonstrate the effectiveness and applicability of the WANARX modelling structure. Conclusions are given in Section 8.

2. Nonlinear input-output representations

In the past few decades, system identification and analysis methods for nonlinear systems have been extensively studied with many applications in approximation, prediction and control. Several nonlinear models have been proposed in the literature including the NARMAX model representation which was initially proposed by Billings and Leontaritis (Billings and Leontaritis 1982, Leontaritis and Billings 1985). The NARMAX model (Pearson 1999) takes the form of the following nonlinear difference equation:

$$y(t) = f(y(t-1),\cdots,y(t-n_y),u(t-1),\cdots,u(t-n_u),e(t-1),\cdots,e(t-n_e)) + e(t) \tag{3}$$

where f is an unknown nonlinear mapping, $u(t)$ and $y(t)$ are the sampled input and output sequences, n_u and n_y are the maximum input and output lags, respectively. The noise variable $e(t)$ with maximum lag n_e, is immeasurable but is assumed to be bounded and uncorrelated with the inputs. The model (3) relates the inputs and outputs and takes into account the combination effects of measurement noise, modelling errors and unmeasured disturbances represented by the noise variable $e(t)$.

The NARX model is a special case of the NARMAX model and is described as

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

One of the popular representations for the NARMAX model (3) is polynomial models, since any continuous function can be arbitrarily well approximated by a polynomial model (Schumaker 1981). Taking the case of SISO systems as an example and expanding model (3) by defining the function $f(\cdot)$ to be a polynomial of degree ℓ gives the representation

$$y(t) = \theta_0 + \sum_{i_1=1}^{n} f_{i_1}(x_{i_1}(t)) + \sum_{i_1=1}^{n}\sum_{i_2=i_1}^{n} f_{i_1 i_2}(x_{i_1}(t), x_{i_2}(t)) + \cdots + \sum_{i_1=1}^{n}\cdots\sum_{i_\ell=i_{\ell-1}}^{n} f_{i_1 i_2 \cdots i_\ell}(x_{i_1}(t), x_{i_2}(t), \cdots, x_{i_\ell}(t)) + e(t) \tag{5}$$

where $\theta_{i_1 i_2 \cdots i_m}$ are parameters, $n = n_y + n_u + n_e$ and

$$f_{i_1 i_2 \cdots i_m}(x_{i_1}(t), x_{i_2}(t), \cdots, x_{i_m}(t)) = \theta_{i_1 i_2 \cdots i_m} \prod_{k=1}^{m} x_{i_k}(t),\ 1 \le m \le \ell \tag{6}$$

$$x_k(t) = \begin{cases} y(t-k) & 1 \le k \le n_y \\ u(t-(k-n_y)) & n_y+1 \le k \le n_y+n_u \\ e(t-(k-n_y-n_u)) & n_y+n_u+1 \le k \le n_y+n_u+n_e \end{cases} \tag{7}$$

The degree of a multivariate polynomial is defined as the highest order among the terms, for example, the degree of the polynomial $h(x_1, x_2, x_3) = a_1 x_1^4 + a_2 x_2 x_3 + a_3 x_1^2 x_2 x_3^2$ is $\ell = 2+1+2=5$. Similarly, a NARMAX model with polynomial degree ℓ means that the order of each term in the model is not higher than ℓ.

As a general and natural representation for a wide class of linear and nonlinear systems, model (5) includes, as special cases, several model types, including the Volterra and Wiener representations, time-invariant and time-varying AR(X), NARX and ARMA(X) structures, output-affine and rational models, and the bilinear model (Pearson 1995). The ANOVA expansions (2) can also be viewed as a special case of the NARMAX model while representing dynamic input and output systems.

Now, consider the NARX model (4) and assume that the nonlinear mapping f in the model (4) can be decomposed into a number of functional components as the ANOVA expansion (2), then the NARX model (4) can be expressed as

$$\begin{aligned} y(t) &= f(x_1(t), x_2(t), \cdots, x_n(t)) + e(t) \\ &= f_0 + F_1(x(t)) + F_2(x(t)) + \cdots + F_m(x(t)) + \cdots + F_n(x(t)) + e(t) \end{aligned} \tag{8}$$

where $x(t) = [x_1(t), x_2(t), \cdots, x_n(t)]^T$ and

$$x_k(t) = \begin{cases} y(t-k), & 1 \le k \le n_y \\ u(t-k+n_y), & n_y+1 \le k \le n = n_y+n_u \end{cases} \tag{8a}$$

$$F_1(x(t)) = \sum_{i=1}^{n} f_i(x_i(t)) \tag{8c}$$

$$F_2(x(t)) = \sum_{i=1}^{n}\sum_{j=i+1}^{n} f_{ij}(x_i(t), x_j(t)) \tag{8d}$$

$$F_m(x(t)) = \sum_{1\le i_1<i_2<\cdots<i_m\le n} f_{i_1 i_2\cdots i_m}(x_{i_1}(t), x_{i_2}(t), \cdots, x_{i_m}(t)), \quad 2<m<n, \tag{8e}$$

$$F_n(x(t)) = f_{12\cdots n}(x_1(t), x_2(t), \cdots, x_n(t)) \tag{8f}$$

This can be referred to as the ANOVA decomposition of the NARX model. Although the ANOVA expansion (2) or the NARX model (8) involves up to 2^n different functional components, experience shows that a truncated representation containing the components up to the bivariate functional terms is often sufficient

$$y(t) = f_0 + \sum_{p=1}^{n} f_p(x_p(t)) + \sum_{p=1}^{n}\sum_{q=p+1}^{n} f_{pq}(x_p(t), x_q(t)) + e(t) \tag{9}$$

This can often provide a satisfactory description of $y(t)$ for many high dimensional problems providing that the input variables are properly selected. The presence of only low order functional components does not necessarily imply that the high order variable interactions are not significant, nor does it mean the nature of the nonlinearity of the system is less severe. An exhaustive search for all the possible submodel structures of (2) is demanding and can be prohibitive because of the curse-of-dimensionality. A truncated representation is advantageous and practical if the higher order terms can be ignored. In practice, the constant term f_0 can often be omitted since it can be combined into other functional components.

In practice, many types of functions, such as kernel functions, splines, polynomials and other basis functions can be chosen to express the functional components in model (8). In the present study, however, wavelet decompositions, which are discussed in the next section, will be chosen to describe the functional components in the additive models (8) and (9), and this will be referred to as the wavelet-based the ANOVA decomposition of the NARX model or simply the WANARX model.

3. Wavelet transform and wavelet decompositions

Wavelet analysis is based on a wavelet prototype function, called the *analysing wavelet*, *mother wavelet*, or simply *wavelet*. Temporal analysis is performed using a contracted, high-frequency version of the same function. Because the signal or function to be studied can be represented in terms of wavelet decompositions, data operations can also be performed using the corresponding wavelet coefficients.

3.1 The continuous wavelet transform

From wavelet theory, the continuous wavelet transform (CWT) of a given function $f \in L^2(R)$ with respect to the *mother wavelet* φ is defined as (Chui 1992, Daubechies 1992).

$$(W_\varphi f)(a,b) = \int_{-\infty}^{\infty} f(x)\varphi^*_{(a,b)}(x)dx \tag{10}$$

where $\varphi^*_{(a,b)}(x)$ indicates the complex conjugate of the function $\varphi_{(a,b)}(x)$, which is obtained by dilating and translating the mother wavelet $\varphi(x)$ as follows

$$\varphi_{(a,b)}(x) = a^{-\frac{1}{2}}\varphi\left(\frac{x-b}{a}\right), \quad a \in R^+, b \in R \tag{11}$$

The CWT (10) is invertible subject to a mild restriction imposed on the wavelet φ

$$C_\varphi = \int_0^\infty \frac{|\hat{\varphi}(\omega)|^2}{\omega} d\omega < \infty \tag{12}$$

in the sense that

$$f(x) = \frac{1}{C_\varphi} \int_0^\infty \frac{da}{a^2} \int_{-\infty}^\infty [(W_\varphi f)(a,b)] \varphi_{(a,b)}(x) db \tag{13}$$

where $\hat{\varphi}$ is the Fourier transform of the function φ. Equation (10) states that the continuous wavelet transform $(W_\varphi f)(a,b)$ is the correlation of $f(x)$ with a scaling (dilation) a and a shift (translation) b. The inverse transform (13) guarantees that the function $f(x)$ can be reconstructed from the CWT and it can be interpreted in at least two different ways. On the one hand, this shows how to reconstruct the function f from the wavelet transform and, on the other hand, the inverse transform gives a recipe showing how to write any arbitrary f as a superposition of the wavelet functions $\varphi_{(a,b)}(x)$.

The wavelet transform (10) and (13) can be immediately extended to the multidimensional, say *d*-dimensional case, by taking direct products of d univariate wavelets (Zhang 1992). For a given function $f \in L^2(R^d)$ $(d \geq 2)$, the continuous wavelet transform can be expressed as

$$(W_\Psi f)(a,b) = \int_{R^d} f(x) \Psi^*_{a,b}(x) dx \tag{14}$$

$$f(x) = \frac{1}{C_\Psi} \int_{R^d} \int_{R^{+d}} \frac{da}{\sigma^2} \int_{-\infty}^\infty [(W_\Psi f)(a,b)] \Psi_{a,b}(x) db \tag{15}$$

where $\sigma = \prod_{i=1}^d a_i$, $C_\Psi = (C_{\varphi_0})^d$ and

$$\Psi_{a,b}(x) = \sigma^{-1/2} \Psi\left(\frac{x_1 - b_1}{a_1}, \frac{x_2 - b_2}{a_2}, \cdots, \frac{x_d - b_d}{a_d} \right) \tag{16}$$

$$\Psi(x_1, x_2, \cdots, x_d) = \prod_{i=1}^d \varphi_0(x_i) \tag{17}$$

$R^{+d} = R^+ \times R^+ \times \cdots \times R^+$ and φ_0 is a univariate wavelet satisfying the admissibility condition (10). Other forms of multi-dimensional wavelet transforms are also used in practice.

3.2 Wavelet frame decomposition

Let φ be a *d*-dimensional wavelet function and $f \in L^2(R^d)$. Assume that there exists a denumerable family derived from φ

$$\Omega = \left\{ \varphi_{(a_j,b_j)} : \varphi_{(a_j,b_j)}(x) = (\det(Q_j))^{1/2} \varphi(A_j(x - b_j)) \right\} \tag{18}$$

where $b_j \in R^d$ is a translation vector , $a_j = [a_j^{(1)}, a_j^{(2)}, \cdots, a_j^{(d)}]^T \in R^{+d}$ is a dilation vector, and $Q_j = diag[(a_j^{(1)})^{-1}, (a_j^{(2)})^{-1}, \cdots, (a_j^{(d)})^{-1}]$. Rearrange the elements of Ω so that

$$\Omega = \{\phi_n : n \in \Gamma\} \tag{19}$$

where the index set Γ might be finite or infinite. The wavelet function φ is said to generate a *frame* $\{\phi_n\}_{n\in\Gamma}$ in $L^2(R^d)$ if there exist two positive constants A and B satisfying $0 < A \le B < \infty$, such that for any function $f \in L^2(R^d)$

$$A\|f\|^2 \le \sum_{n\in\Gamma} |< f, \phi_n >|^2 \le B\|f\|^2 \tag{20}$$

The frame is said to be tight if $A = B$ (Chui 1992, Daubechies 1992). The symbol '$< \cdot, \cdot >$' in (20) denotes the inner product of two functions in $L^2(R^d)$.

Under the condition that φ generates a frame, it is assured that any function $f \in L^2(R^d)$ can be recovered from the inverse wavelet transform (15) in the sense that

$$f(x) = \sum_{j\in\Gamma} c_j \phi_j(x) \tag{21}$$

or

$$f(x) = \sum_{j\in\Gamma} c_j Q_j^{1/2} \varphi(A_j(x - b_j)) = \sum_{j\in\Gamma} w_j \varphi(Q_j(x - b_j)) \tag{22}$$

This is called the *wavelet frame decomposition*, which can be approximated by a neural network structure and it is therefore often referred to as a *wavelet network* (Zhang 1992).

3.3 Wavelet series

In practical applications the CWT is often discretised in both the scaling and dilation parameters for computational efficiency. Based on this discretization, wavelet decompositions can be obtained to provide an alternative basis function representation. Take the univariate wavelet as an example. The most popular approach to discetise the CWT is to restrict the dilation and translation parameters to a dyadic lattice as $a = 2^{-j}$ and $b = k2^{-j}$ with $j, k \in Z$. Other non-dyadic ways of discretization are also available.

Let $\varphi_{j,k}(x) = 2^{j/2}\varphi(2^j x - k)$ be an orthogonal family with respect to $j, k \in Z$, then the wavelet transform of a function $f \in L^2(R)$ can be expressed as

$$c_{j,k} = (W_\varphi f)(2^{-j}, k2^{-j}) =< f, \varphi_{j,k} >, \quad j, k \in Z, \tag{23}$$

Hence the discrete wavelet transform (23) can be viewed as a discretised version of the CWT (14). Any $f \in L^2(R)$ can be uniquely described as

$$f(x)=\sum_{j=-\infty}^{\infty}\sum_{k=-\infty}^{\infty}c_{j,k}\varphi_{j,k}(x) \tag{24}$$

where the convergence of the series in (24) is in $L^2(R)$, namely

$$\lim_{J_1,J_2,K_1,K_2\to\infty}\left\|f(x)-\sum_{j=-J_1}^{J_2}\sum_{k=-K_1}^{K_2}c_{j,k}\varphi_{j,k}(x)\right\|=0 \tag{25}$$

In general, however, it is not necessary to require $\{\varphi_{j,k}\}$ to be an orthogonal basis of $L^2(R)$ in the sense that

$$<\varphi_{j,k},\varphi_{\ell,m}>=\delta_{j,\ell}\cdot\delta_{k,m}\ ,\ \ j,k,\ell,m\in Z \tag{26}$$

In fact, the following two conditions are sufficient to guarantee a wavelet φ to form a wavelet series (Chui 1992):

(i) *The function family* $\{\varphi_{j,k}\}_{j,k\in Z}$ *is a Riesz basis of* $L^2(R)$, *in the sense that the linear span of* $\varphi_{j,k}$ *is dense in* $L^2(R)$, *and there exist positive constants A and B, with* $0<A\le B<\infty$, *such that*

$$A\sum_{j=-\infty}^{\infty}\sum_{k=-\infty}^{\infty}\left|c_{j,k}\right|^2\le\left\|\sum_{j=-\infty}^{\infty}\sum_{k=-\infty}^{\infty}c_{j,k}\varphi_{j,k}\right\|_2^2\le B\sum_{j=-\infty}^{\infty}\sum_{k=-\infty}^{\infty}\left|c_{j,k}\right|^2 \tag{27}$$

for all doubly bi-infinite square-summable sequences $\{c_{j,k}\}$.

(ii) *There exists at least one function* $\widetilde{\varphi}\in L^2(R)$, *such that the family* $\widetilde{\varphi}_{j,k}(x)=2^{j/2}\widetilde{\varphi}(2^j x-k)$ *is a Riesz basis of* $L^2(R)$ *and is dual to* $\{\varphi_{j,k}\}_{j,k\in Z}$, *and the* pair $(\varphi,\widetilde{\varphi})$ *satisfies the bi-orthogonal property in the sense that*

$$<\varphi_{j,k},\widetilde{\varphi}_{\ell,m}>=\delta_{j,\ell}\cdot\delta_{k,m}\ ,\ \ j,k,\ell,m\in Z \tag{28}$$

If $\{\varphi_{j,k}\}$ is an orthogonal basis in $L^2(R)$, then it is clear that (17) holds with $\widetilde{\varphi}_{j,k}=\varphi_{j,k}$. Theoretically, if the dual pair $(\varphi,\widetilde{\varphi})$ exists and the above conditions (i) and (ii) hold, then every $f\in L^2(R)$ can be uniquely written as

$$f(x)=\sum_{j,k=-\infty}^{\infty}<f,\widetilde{\varphi}_{j,k}>\varphi_{j,k}(x) \tag{29}$$

and this is called a *wavelet series*. In comparison with the CWT, the wavelet series is more computationally efficient. But this is obtained at the expense of increased restrictions on the choice of the basic wavelet φ. The wavelet series (29) can be extended to *d*-dimensional case by taking tensor products of one-dimensional wavelets or by choosing the radial types of wavelets and this will be discussed later.

3.4 Mulitresolution wavelet decompositions

It is known that for solving identification problems based on the regression representation it is useful to have a basis of orthogonal (semi-orthogonal or bi-orthogonal) functions whose support can be made as small as required and which provides a universal approximation to any $L^2(R)$ function with arbitrary desired accuracy.

One of the original objectives of wavelet theory was to construct orthogonal (semi-orthogonal) basis in $L^2(R)$. The principles for constructing orthogonal wavelets are as follows:

(*i*) The family $\{\varphi_{j,k}(x) = 2^{j/2}\varphi(2^j x - k), j,k \in Z\}$ constitutes an orthogonal basis for the space $L^2(R)$;

(*ii*) There exists a function ϕ, called a *scaling function* related to the mother wavelet φ, such that the elements of the family $\{\phi(t-k)\}_{k\in Z}$ are mutually orthogonal;

(*iii*) For $\forall j \in Z$, the family $\{\phi_{j,k}(x) = 2^{j/2}\phi(2^j x - k), k \in Z\}$ constitute an orthogonal basis for $L^2(R)$;

(*iv*) The basic function φ and the scaling function ϕ are related by some deterministic equations.

To satisfy the above aims, an orthogonal wavelet system can be constructed using *multiresolution analysis* (MRA)(Mallat 1989, Chui 1992). Let W_j ($j \in Z$) denote some wavelet subspaces, which are defined as the closure of the linear span of the wavelet functions $\{\varphi_{j,k}\}_{k\in Z}$, namely

$$W_j = \overline{span}\{\varphi_{j,k}, k \in Z\} \tag{30}$$

which satisfy

$$W_i \cap W_j = \{\varnothing\}, \text{ for any } i \neq j \tag{31}$$

where the over-bar denotes closure. It follows that $L^2(R)$ can be decomposed as a direct sum of the spaces W_j:

$$L^2(R) = \cdots \oplus W_{-1} \oplus W_0 \oplus W_1 \oplus \cdots \tag{32}$$

in the sense that every function $f \in L^2(R)$ has a unique decomposition

$$f(x) = \cdots + g_{-1}(x) + g_0(x) + g_1(x) + \cdots = \sum_{j\in Z} g_j(x) \tag{33}$$

The circles around the plus signs in (32) indicate "orthogonal sums". The decomposition of (32) is usually called an *orthogonal decomposition* of $L^2(R)$.

For each $j \in Z$, consider the closed subspaces of $L^2(R)$

$$V_j = \cdots \oplus W_{j-2} \oplus W_{j-1}, \; j \in Z \tag{34}$$

which have the following properties:

(*i*) $\cdots \subset V_{-1} \subset V_0 \subset V_1 \subset \cdots$,

(*ii*) $(\overline{\bigcup_{j\in Z} V_j}) = L^2(R)$ (the over-bar here indicates closure),

(*iii*) $\bigcap_{j\in Z} V_j = \{\varnothing\}$,

(*iv*) $V_{j+1} = V_j \oplus W_j, \; \forall j \in Z$,

(*v*) $f(x) \in V_j \Leftrightarrow f(2x) \in V_{j+1}, \; \forall j \in Z$,

(*vi*) $f(x) \in V_j \Leftrightarrow f(x-2^j k) \in V_j,\ \forall j,k \in Z$,

(*vii*) $\{\phi(t-k)\}_{k\in Z}$ is an orthogonal basis for V_0.

It is clear that every function $f \in L^2(R)$ can be approximated as closely as desirable by the projections $P_j f$ in V_j. Another important intrinsic property of these spaces is that more and more variations of $P_j f$ are removed as $j \to -\infty$. In fact, these variations are peeled off, level by level in decreasing order of the rate of variations (frequency bands) and stored in the complementary W_j, shown as in property (*iv*).

Assume that the wavelet φ and the corresponding scaling function ϕ constitute an orthogonal wavelet system, then any function $f \in L^2(R)$ can be expressed as the following *multiresolution wavelet decomposition*

$$f(x) = \sum_k \alpha_{j_0,k}\phi_{j_0,k}(x) + \sum_{j\geq j_0}\sum_k \beta_{j,k}\varphi_{j,k}(x) \tag{35}$$

where the wavelet coefficients $\alpha_{j_0,k}$ and $\beta_{j,k}$ can be calculated in theoretical by the innner products:

$$\alpha_{j_0,k} =< f,\phi_{j_0,k} >= \int f(x)\varphi^*_{j_0,k}(x)dx \tag{36}$$

$$\beta_{j,k} =< f,\varphi_{j,k} >= \int f(x)\varphi^*_{j,k}(x)dx \tag{37}$$

and j_0 is an arbitrary integer representing the lowest resolution or scaling level. Notice from (32) that if $j_0 \to -\infty$, the approximation representation (35) becomes the wavelet decomposition (29). In addition, based on (34) and the properties of MRA, any function $f \in L^2(R)$ can be arbitrarily closely approximated in V_J for some sufficiently large integer *J*. That is, for any ε >0, there exists a sufficiently large integer *J*, such that

$$\left\| f(x) - \sum_k < f,\phi_{J,k} > \phi_{J,k}(x) \right\| < \varepsilon \tag{38}$$

This means that in wavelet series representation, the wavelet bases can be replaced by orthogonal scaling functions with a large resolution scale.

Using the concept of *tensor products*, the multiresolution decomposition (35) can be immediately generalised to the muti-dimensional case, where a multiresolution wavelet decomposition can be defined by taking the *tensor product* of the one-dimensional scaling and wavelet functions (Mallat 1989). Let $f \in L^2(R^d)$, then $f(x)$ can be represented by the *multiresolution wavelet decomposition* as

$$f(x_1,\cdots,x_d) = \sum_k \alpha_{j_0,k}\Phi_{j_0,k}(x_1,\cdots,x_d) + \sum_{j\geq j_0}\sum_k\sum_{l=1}^{2^d-1}\beta^{(l)}_{j,k}\Psi^{(l)}_{j,k}(x_1,\cdots,x_d) \tag{39}$$

where $k = (k_1,k_2,\cdots,k_d) \in Z^d$ and

$$\Phi_{j_0,k}(x_1,\cdots,x_d) = 2^{j_0 d/2}\prod_{i=1}^{d}\phi(2^{j_0}x_i - k_i) \tag{40}$$

$$\Psi_{j,k}^{(l)}(x_1,\cdots,x_d) = 2^{jd/2}\prod_{i=1}^{d}\eta^{(i)}(2^j x_i - k_i) \tag{41}$$

with $\eta^{(i)} = \phi$ or φ (scalar scaling function or the mother wavelet) but at least one $\eta^{(i)} = \varphi$. In the two-dimensional case, the multiresolution approximation can be generated, for example, in terms of the dilation and translation of the two-dimensional scaling and wavelet functions

$$\begin{cases} \Phi_{j,k_1,k_2}(x_1,x_2) = \phi_{j,k_1}(x_1)\phi_{j,k_2}(x_2) \\ \Psi^{(1)}_{j,k_1,k_2}(x_1,x_2) = \phi_{j,k_1}(x_1)\varphi_{j,k_2}(x_2) \\ \Psi^{(2)}_{j,k_1,k_2}(x_1,x_2) = \varphi_{j,k_1}(x_1)\phi_{j,k_2}(x_2) \\ \Psi^{(3)}_{j,k_1,k_2}(x_1,x_2) = \varphi_{j,k_1}(x_1)\varphi_{j,k_2}(x_2) \end{cases} \tag{42}$$

4. Expanding the WANARX model using wavelet decompositions

The wavelet decompositions including the wavelet frame decomposition, wavelet series and wavelet multiresolution decompositions discussed in Section 3 can be adapted to express each functional component in the NARX model (8). Notice that it is impossible in practice to count infinite frame terms or wavelet bases in a wavelet decomposition. Therefore, the infinite decompositions are always truncated at appropriate dilations (resolutions) and translations.

4.1 Expanding the functional components using wavelet frame decomposition

Each functional component in the NARX model (8) can be expressed using the truncated wavelet frame decomposition (22), take the functional component $f_{i_1 i_2 \cdots i_m}(x_{i_1}(t), x_{i_2}(t), \cdots, x_{i_m}(t))$ as an example, this can be expanded as

$$\begin{aligned} & f_{i_1 i_2 \cdots i_m}(x_{i_1}(t), x_{i_2}(t), \cdots, x_{i_m}(t)) \\ & \quad = \sum_{i=1}^{I_{i_1 i_2 \cdots i_m}} w_i^{(i_1 i_2 \cdots i_m)} \Psi^{(i_1 i_2 \cdots i_m)}(Q_i^{(i_1 i_2 \cdots i_m)}(x^{(i_1 i_2 \cdots i_m)}(t) - b_i^{(i_1 i_2 \cdots i_m)})) \end{aligned} \tag{43}$$

where

$$x^{(i_1 i_2 \cdots i_m)}(t) = [x_{i_1}^{(i_1 i_2 \cdots i_m)}(t), x_{i_2}^{(i_1 i_2 \cdots i_m)}(t), \cdots, x_{i_3}^{(i_1 i_2 \cdots i_m)}(t)]^T$$

$$b_i^{(i_1 i_2 \cdots i_m)} = [b_i^{(i_1 i_2 \cdots i_m)(1)}, b_i^{(i_1 i_2 \cdots i_m)(2)}, \cdots, b_i^{(i_1 i_2 \cdots i_m)(m)}]^T,$$

$$a_i^{(i_1 i_2 \cdots i_m)} = [a_i^{(i_1 i_2 \cdots i_m)(1)}, a_i^{(i_1 i_2 \cdots i_m)(2)}, \cdots, a_i^{(i_1 i_2 \cdots i_m)(m)}]^T,$$

$$Q_i^{(i_1 i_2 \cdots i_m)} = diag[(a_i^{(i_1 i_2 \cdots i_m)(1)})^{-1}, (a_i^{(i_1 i_2 \cdots i_m)(2)})^{-1}, \cdots, (a_i^{(i_1 i_2 \cdots i_m)(m)})^{-1}],$$

$I_{i_1 i_2 \cdots i_m}$ is the number of wavelets in the wavelet library composed of all the wavelets under consideration. $\Psi^{(i_1 i_2 \cdots i_m)}$ indicates that different types of wavelets can be employed simultaneously for approximating different functional components. This might enable the wavelet decomposition to be more flexible than traditional wavelet networks.

Inserting (43) into (8), yields

$$\hat{y}(t) = \hat{f}_0 + \sum_{i_1=1}^{n}\sum_{k=1}^{I_{i_1}} w_k^{(i_1)}\Psi^{(i_1)}(Q_k^{(i_1)}(x^{(i_1)}(t) - b_k^{(i_1)}))$$
$$+ \sum_{1\le i_1<i_2\le n}\sum_{k=1}^{I_{i_1 i_2}} w_k^{(i_1 i_2)}\Psi^{(i_1 i_2)}(Q_k^{(i_1 i_2)}(x^{(i_1 i_2)}(t) - b_k^{(i_1 i_2)})) + \cdots$$
$$+ \sum_{k=1}^{I_{12\cdots n}} w_k^{(12\cdots n)}\Psi^{(12\cdots n)}(Q_k^{(12\cdots n)}(x^{(12\cdots n)}(t) - b_k^{(12\cdots n)})) \tag{44}$$

This will be referred to as a *super wavelet network*. The values of the decomposition parameters $w_k^{(i_1 i_2\cdots i_m)}$, $b_k^{(i_1 i_2\cdots i_m)}$ and $a_k^{(i_1 i_2\cdots i_m)}$ can be obtained by minimizing a criterion function, say

$$\min\ V = \frac{1}{2}\sum_{t=1}^{T} e^2(t) = \frac{1}{2}\sum_{t=1}^{T}[y(t) - \hat{y}(t)]^2 \tag{45}$$

where $y(t)$ is the measurement at time t, and T is the data length. To minimise the function V, gradient descent type methods are required and thus the gradients of unknown parameters should be calculated first. Taking part of the univariate and bivariate functional components in (45) as an example, the partial differentials of V with respect to the parameters are

$$\frac{\partial V}{\partial w_k^{(i_1)}} = -\sum_{t=1}^{T} e(t)\Psi^{(i_1)}(\tilde{x}^{(i_1)}), \tag{46a}$$

$$\frac{\partial V}{\partial b_k^{(i_1)(1)}} = -\sum_{t=1}^{T} e(t) w_k^{(i_1)}\left(-\frac{1}{a_k^{(i_1)(1)}}\right)\frac{\partial\Psi^{(i_1)}(\tilde{x}^{(i_1)})}{\partial\tilde{x}^{(i_1)}}, \tag{46b}$$

$$\frac{\partial V}{\partial a_k^{(i_1)(1)}} = -\sum_{t=1}^{T} e(t) w_k^{(i_1)}\left(-\frac{1}{a_k^{(i_1)(1)}}\right)\tilde{x}^{(i_1)}\frac{\partial\Psi^{(i_1)}(\tilde{x}^{(i_1)})}{\partial\tilde{x}^{(i_1)}}, \tag{46c}$$

k=1,2,..., I_{i_1}; i_1=1,2,...,n;

$$\frac{\partial V}{\partial w_k^{(i_1 i_2)}} = -\sum_{t=1}^{T} e(t)\Psi^{(i_1 i_2)}(\tilde{x}^{(i_1 i_2)}), \tag{46d}$$

$$\frac{\partial V}{\partial b_k^{(i_1 i_2)(p)}} = -\sum_{t=1}^{T} e(t) w_k^{(i_1 i_2)}\left(-\frac{1}{a_k^{(i_1 i_2)(p)}}\right)\frac{\partial\Psi^{(i_1 i_2)}(\tilde{x}^{(i_1 i_2)})}{\partial\tilde{x}_{i_p}^{(i_1 i_2)}}, \tag{46e}$$

$$\frac{\partial V}{\partial a_k^{(i_1 i_2)(p)}} = -\sum_{t=1}^{T} e(t) w_k^{(i_1 i_2)}\left(-\frac{1}{a_k^{(i_1 i_2)(p)}}\right)\tilde{x}_p^{(i_1 i_2)}\frac{\partial\Psi^{i_1 i_2)}(\tilde{x}^{(i_1 i_2)})}{\partial\tilde{x}_{i_p}^{(i_1 i_2)}}, \tag{46f}$$

p=1,2; k=1,2,..., $I_{i_1 i_2}$; $1 \le i_1 < i_2 \le n$

$$\frac{\partial V}{\partial w_k^{(i_1 i_2 i_3)}} = -\sum_{t=1}^{T} e(t)\Psi^{(i_1 i_2 i_3)}(\tilde{x}^{(i_1 i_2 i_3)}), \tag{46g}$$

... ... (etc.)

where

$$\widetilde{x}^{(i_1)} = \frac{x^{(i_1)}(t) - b_k^{(i_1)(1)}}{a_k^{(i_1)(1)}},$$

$$\widetilde{x}^{(i_1 i_2)} = [\widetilde{x}_{i_1}^{(i_1 i_2)}, \widetilde{x}_{i_2}^{(i_1 i_2)}]^T = \left[\frac{x_{i_1}^{(i_1 i_2)}(t) - b_k^{(i_1 i_2)(1)}}{a_k^{(i_1 i_2)(1)}}, \frac{x_{i_2}^{(i_1 i_2)}(t) - b_k^{(i_1 i_2)(2)}}{a_k^{(i_1 i_2)(2)}} \right]^T.$$

Once the gradients have been obtained, Gauss-Newton type optimisation methods including steepest decent and stochastic gradient methods can be used to obtain the unknown parameters.

Note that the wavelets used in the adaptive wavelet decomposition (wavelet networks) should be explicitly expressible and differentiable. This restricts the choice of basic wavelet functions used for wavelet networks to a special class. The Morlet wavelet $\varphi(x) = \exp(j\omega_0^T x - \|x\|^2/2)$, Gaussian wavelet $\varphi(x) = x_1 x_2 \cdots x_d \exp(-\|x\|^2/2)$, and Marr (Mexican hat) wavelet $\varphi(x) = (d - x_1 x_2 \cdots x_d)\exp(-\|x\|^2/2)$ are among the examples which are often used in wavelet networks. The symbol $\|\cdot\|$ here denotes the Euclidian norm in $L^2(R^d)$.

Notice that a radial wavelet is often considered and the family (18) is often restricted in a regular grid, that is, the translation and dilation parameters a_j and b_j in (18) are designed to form a double indexed regular lattice

$$\{(a_j, b_k) = (\alpha^{-j}, k\beta\alpha^{-j}) : j \in Z, k \in Z^d\} \tag{47}$$

where the scalar parameters α and β are the discretization step with typical values $\alpha = 2$ and $\beta = 1$. Expanding each functional component in (8) using a radial wavelet frame

$$\Omega_R = \{\Psi_{j,k} : \Psi_{j,k}(x) = \alpha^{jd/2}\Psi(\alpha^j x - k\beta), j \in Z, k \in Z^d\} \tag{48}$$

The NARX structure (8) can then be converted into a linear-in-the-parameters problem, which can be solved using regression analysis techniques, and this will be referred to as a *super wavelet network on fixed grid*, or, *super fixed grid wavelet network*. If only the last functional component $f_{12\cdots n}(x_1(t), x_2(t), \cdots, x_n(t))$ in (8) is considered and expanded using a radial wavelet frame with the form of (48), then this decomposition can be treated as a standard linear regression problem with the dilated and translated wavelets as the regressors. This will be referred to as a *fixed grid wavelet network*, which is a special case of the *super fixed grid wavelet networks* considered here.

4.2 Expanding the functional components using wavelet series

Consider the functional component $f_{i_1 i_2 \cdots i_m}(x_{i_1}(t), x_{i_2}(t), \cdots, x_{i_m}(t))$ in the NARX expansion (8). This functional component can be approximated using the truncated wavelet series (29)

$$f_{i_1 i_2 \cdots i_m}(x_{i_1}(t), x_{i_2}(t), \cdots, x_{i_m}(t))$$

$$= \sum_{j=j_m}^{J} \sum_{k_1 \cdots k_m} c_{j;k_1,k_2,\cdots k_m}^{(i_1 i_2 \cdots i_m)} B_m(2^j x_{i_1}(t) - k_1, 2^j x_{i_2}(t) - k_2, \cdots, 2^j x_{i_m}(t) - k_m) \tag{49}$$

where $k = [k_1, k_2 \cdots, k_m]^T \in Z^m$ is an m-dimensional index, $B_m(x)$ is an m-dimensional wavelet or scaling function and can be decomposed as the direct product of m one-dimensional functions

$$B_m(x) = B_m(x_1, x_2, \cdots, x_m) = \prod_{i=1}^{m} \psi(x_i) \tag{50}$$

where $\psi(\cdot)$ is a scalar wavelet or scaling function. Now (49) can be expressed as

$$f_{i_1 i_2 \cdots i_m}(x_{i_1}(t), x_{i_2}(t), \cdots, x_{i_m}(t)) = \sum_{j=j_m}^{J} \sum_{k_1, k_2, \cdots, k_m} c_{j;k_1,k_2,\cdots,k_m}^{(i_1 i_2 \cdots i_m)} \prod_{p=1}^{m} \psi(2^j x_{i_p} - k_p) \tag{51}$$

Inserting (51) into (10), yields

$$y(t) = \hat{f}(x_1(t), x_2(t), \cdots, x_n(t)) = \hat{f}_0 + \hat{F}_1(x(t)) + \hat{F}_2(x(t)) + \cdots + \hat{F}_n(x(t)) + e(t) \tag{52}$$

where $\hat{f}_0$ is a constant and

$$\hat{F}_1(x(t)) = \sum_{p=1}^{n} \sum_{j=j_1}^{J_1} \sum_{k} c_{j;k}^{(p)} \psi(2^j x_p(t) - k_p) \tag{52a}$$

$$\hat{F}_2(x(t)) = \sum_{1 \le p < q \le n} \sum_{j=j_2}^{J_2} \sum_{k_1, k_2} c_{j;k_1,k_2}^{(pq)} \psi(2^j x_p(t) - k_1) \psi(2^j x_q(t) - k_2) \tag{52b}$$

$$\hat{F}_m(x(t)) = \sum_{1 \le i_1 < \cdots i_m \le n} \sum_{j=j_m}^{J_m} \sum_{k_1, \cdots, k_m} c_{j;k_1,\cdots,k_m}^{(i_1 i_2 \cdots i_m)} \prod_{p=1}^{m} \psi(2^j x_{i_p}(t) - k_p) \tag{52c}$$

$$\hat{F}_n(x(t)) = \sum_{j=j_n}^{J_n} \sum_{k_1, k_2, \cdots, k_n} c_{j;k_1,k_2,\cdots,k_n}^{(12\cdots n)} \prod_{p=1}^{n} \psi(2^j x_p(t) - k_p) \tag{52d}$$

Generally, the initial resolution j_m $(m = 1,2,\cdots,n)$ can be chosen to be the same $j_1 = j_2 = \cdots = j_n = j_0$, similarly for the maximum resolution, that is, $J_1 = J_2 = \cdots = J_n = J$.

Assume that M wavelet bases (mother wavelet or scaling functions) are required to expand the NARX model (8), and for convenience of representation also assume that the M wavelet bases are ordered according to a single index m, that is, $W = \{\psi_m\}_{m=1}^{M}$, then (52) can be expressed as a linear-in-the-parameters form as below:

$$y(t) = \sum_{m=1}^{M} \theta_m \psi_m(t) + e(t) \tag{53}$$

which can be solved using linear regression techniques. Note that, the regressor family $W = \{\psi_m\}_{m=1}^{M}$ might be redundant, since in practice it is usually true that the sampled data only form a sparse distribution in the input space. Consequently, the regression problem is often ill-posed and therefore some approaches should be employed to resolve this problem. It has been proven that the forward orthogonal least squares (OLS) method is

an effective approach to solve this ill-posed problem (Billings et al. 1988, 1989, Korenberg et al. 1988, Chen et al. 1989). The regressor selection problem will be discussed in the next section.

From (29), every wavelet φ, orthogonal or not, generates a wavelet series representation of any $f \in L^2(R)$ as long as the dual $\tilde{\varphi}$ of φ exists. From (38), the wavelet bases in the wavelet series representation (24) can be replaced by orthogonal scaling functions with a large resolution. This provides more freedom in the choice of basis functions in the wavelet series decomposition.

4.3 Expanding the functional components using multiresolution wavelets models

Take the two-dimensional additive model (9) as an example. Expanding each functional component in the model (9) into the truncated multiresolution wavelet decompositions (35) or (39) following Wei et al. (2003a)

$$f_p(x_p(t)) = \sum_k \alpha_{j_1,k}^{(p)} \phi_{j_1,k}(x_p(t)) + \sum_{j \geq j_1} \sum_k \beta_{j,k}^{(p)} \varphi_{j,k}(x_p(t)), \quad p = 1,2,\cdots,n, \tag{54}$$

$$\begin{aligned} f_{pq}(x_p(t), x_q(t)) = & \sum_{k_1} \sum_{k_2} \alpha_{j_2;k_1,k_2}^{(pq)(1)} \phi_{j_2,k_1}(x_p(t)) \phi_{j_2,k_2}(x_q(t)) \\ & + \sum_{j \geq j_2} \sum_{k_1} \sum_{k_2} \beta_{j;k_1,k_2}^{(pq)(1)} \phi_{j,k_1}(x_p(t)) \varphi_{j,k_2}(x_q(t)) \\ & + \sum_{j \geq j_2} \sum_{k_1} \sum_{k_2} \beta_{j;k_1,k_2}^{(pq)(2)} \varphi_{j,k_1}(x_p(t)) \phi_{j,k_2}(x_q(t)) \\ & + \sum_{j \geq j_2} \sum_{k_1} \sum_{k_2} \beta_{j;k_1,k_2}^{(pq)(3)} \varphi_{j,k_1}(x_p(t)) \varphi_{j,k_2}(x_q(t)), \; 1 \leq p < q \leq n. \end{aligned} \tag{55}$$

Inserting Eqs (54) and (55) into (9) yields

$$\begin{aligned} y(t) &= \hat{f}(x_1, x_2, \cdots, x_n) + e(t) \\ &= c_0 + \sum_{p=1}^{n} \sum_k \alpha_{j_1,k}^{(p)} \phi_{j_1,k}(x_p(t)) + \sum_{p=1}^{n} \sum_{j \geq j_1} \sum_k \beta_{j,k}^{(p)} \varphi_{j,k}(x_p(t)) \\ &+ \sum_{1 \leq p < q \leq n} \sum_{k_1} \sum_{k_2} \alpha_{j_2;k_1,k_2}^{(pq)(1)} \phi_{j_2,k_1}(x_p(t)) \phi_{j_2,k_2}(x_q(t)) \\ &+ \sum_{1 \leq p < q \leq n} \sum_{j \geq j_2} \sum_{k_1} \sum_{k_2} \beta_{j;k_1,k_2}^{(pq)(1)} \varphi_{j,k_1}(x_p(t)) \phi_{j,k_2}(x_q(t)) \\ &+ \sum_{1 \leq p < q \leq n} \sum_{j \geq j_2} \sum_{k_1} \sum_{k_2} \beta_{j;k_1,k_2}^{(pq)(2)} \varphi_{j,k_1}(x_p(t)) \phi_{j,k_2}(x_q(t)) \\ &+ \sum_{1 \leq p < q \leq n} \sum_{j \geq j_2} \sum_{k_1} \sum_{k_2} \beta_{j;k_1,k_2}^{(pq)(3)} \varphi_{j,k_1}(x_p(t)) \varphi_{j,k_2}(x_q(t)) + e(t) \end{aligned} \tag{56}$$

which can be rearranged and converted into a linear-in-the-parameters problem in the form of (53) with respect to the wavelet coefficients $\alpha_{j_1,k}^{(p)}$, $\beta_{j_1,k}^{(p)}$ (p=1,2,...,n), and $\alpha_{j_2;k_1,k_2}^{(pq)(1)}$, $\beta_{j;k_1,k_2}^{(pq)(i)}$ ($1 \leq p < q \leq n$, i=1,2,3). This can be solved using least squares type algorithms, which will be discussed in the next section.

Although many functions can be chosen as scaling and wavelet functions, most of these are not suitable in system identification applications, especially in the case of multidimensional and multiresolution expansions because of the *curse-of-dimensionality*. An implementation, which has been tested with very good results, involves B-spline and B-wavelet functions in multiresolution wavelet decompositions (Billings and Coca 1999, Liu et al 2000, Coca and Billings 2001,Wei and Billings 2002). B-spline wavelets were originally introduced by

Chui and Wang (1992) to define a class of semi-orthogonal wavelets. The reasons that make this implementation particularly suitable in system identification are summarized below:

- B-spline wavelets are piecewise polynomial functions, efficient algorithms for computing these functions and their derivatives are available.
- B-spline wavelets have local support and provide near-optimal time-frequency localization.
- B-spline wavelets outperform other wavelet decompositions in terms of approximation rate .This means that few resolution levels are required to approximate a function in order to achieve a given accuracy. Since each extra level doubles the amount of computations, the choice of wavelet is clearly important. This supports the key parsimony principle in system identification.
- B-spline wavelets $\psi^{[m]}(x)$ are symmetric for even order m and anti-symmetric for odd order m, that is, $\psi^{[m]}(x) = (-1)^m \psi^{[m]}(2m-1-x)$, where [0, $2m$-1] is the support of the B-spline wavelets $\psi^{[m]}(x)$. In application to signal analysis, it is very important for wavelet functions to possess the property of symmetry and anti-symmetry. This is essential to avoid distortion in the reconstruction of compressed data (Chui 1992).

The definition of B-spline wavelets is given in Appendix A. For more details about the properties of B-spline wavelets, see the work of Chui and Wang (Chui 1992, Chui and Wang 1992).

4.4 Hybrid decomposition models

It has been shown in subsections 4.1, 4.2 and 4.3 that each functional component in the NARX model (8) can be expressed using wavelet networks, wavelet series or multiresolution wavelet decompositions. Usually, all the functional components in the NARX model (8) are expanded using the same decomposition, for example, the super wavelet network where all the functional components in the NARX model (8) are expressed using the radial wavelet network with the same type of radial mother wavelet, or, the multiresolution wavelet model (56) where all the functional components are expressed using wavelet multiresolution decompositions with the same type of mother wavelet and scaling function. However, it should be pointed out that it is not necessary to require all the functional components be expressed using the same type of decomposition with the same mother wavelet. In practice, different types of decompositions or different types of mother wavelets can be used simultaneously in a WANARX model, for example,

- Expand all the first-order (unvariate) functional components using wavelet multiresolution decompositions based on a certain type of wavelet and scaling function, say the Haar wavelet (first-order B-spline wavelet) and scaling function, and expand all the second-order(bivariate) functional components using wavelet multiresolution decompositions based on another type of wavelet and scaling function, say the 4th-order B-spline wavelet and scaling function.
- Expand all the first-order (unvariate) functional components using wavelet multiresolution decompositions and expand all the second-order (bivariate) functional components using wavelet series.

The idea of using hybrid decomposition models is to sufficiently utilize the local properties of different types of basic wavelets or scaling functions simultaneously, and to remedy the weakness of one wavelet and/or scaling

function with another. A hybrid decomposition model is often advantageous over a single decomposition model which use only a single type of mother wavelet.

4.4.1 A*daptive wavelet decompositions versus wavelet series and multiresolution wavelet decompositions*

As noted in the section 4.1, the wavelets used in adaptive wavelet decompositions (wavelet networks) should be explicitly expressible and differentiable. The gradients of the criterion function V, and thus the gradients for each of the wavelet functions should be calculated beforehand, and then Gauss-Newton type of optimisation methods such as steepest decent and stochastic gradient methods can be used to optimize the unknown parameters. Gauss-Newton optimisation methods are often in some sense initial-condition dependent. When the number of parameters is large, the convergence rate will be very slow and a great number of iterations are required. In addition, these methods are apt to converge to local minimum. In general, therefore, the adaptive wavelet decomposition may not be suitable for high dimensional problems.

Using wavelet series or multiresolution wavelet decompositions, the WANARX model (8) can be converted into a linear-in-the-parameters problem with respect to the corresponding wavelet coefficients. Notice, however, that the number of potential terms in the model might be very large, but a lot of the candidate terms may be redundant and should be removed from the model. The well known forward orthogonal least squares (OLS) algorithm (Billings et al. 1988, 1989, Korenberg et al. 1988, Chen et al. 1989), combined with the error reduction ratio (ERR) index, which measures the significance of each candidate model term, can be used to solve linear-in-the-parameters problems involving a great number of candidate terms which might possess severe redundancy.

4.4.2 *Radial wavelet networks versus compactly supported wavelet multiresolutoin decompositions*

Both radial wavelet networks (Zhang 1997) and multiresolution wavelet decomposition models (Billings and Coca 1999, Liu et al 2000, Coca and Billings 2001, Wei and Billings 2002) provide powerful representations for nonlinear systems. The model based on the radial wavelet frame (48), or the fixed grid wavelet network, resembles in effect the well known radial basis function (RBF) networks in structure with the Gaussian or thin-spline functions replaced by radial wavelets, which can generate single scaling wavelet frames. The main advantage of the decomposition based on the radial wavelet frame (48) is that the radial construction often leads to a smaller number of candidate regressors (model terms) compared with the multiresolution wavelet decompositions where the compactly supported tensor product wavelets are used. Comparing the multiresolution wavelet models with the radial wavelet networks in detail, the following differences are worth noting:

i) The compactly supported wavelet basis functions, for example, the B-spline wavelet and scaling functions considered in this study, define a hierarchical multiresolution structure with fixed and regular dilation-translation sampling. Thus the location and scale of each basis function is known beforehand (see sections 6.2 and 6.3 for details). In radial wavelet networks, however, the basis functions have to be defined by means of a separate approach, for example, to check the value of each radial wavelet with respect to all the process sampling points.

ii) In the compactly supported wavelet multiresolutoin model, it is not required that every regressor (model term) include all the process variables as in a radial wavelet network. This allows more flexibility in selecting the correct model structure and avoids model over-fitting.

iii) B-spline wavelets are compactly supported. Thus, at a given resolution scale, the number of B-spline wavelets is deterministic. In fact, at each resolution level only the B-spline wavelets which cover the data domain need to be considered. This means that a limited number of B-spline wavelets need to be considered in the truncated multiresolution wavelet model and these are determined by the lowest and the highest resolution scales. Although almost all radial wavelet functions are nearly compactly supported, they only vanish rapidly as the independent variables of these functions are far from the centre. In practice, radial wavelets are usually truncated so that the wavelet support overlaps with the data domain. However, the truncation of the wavelet support might deteriorate the natural approximation property of wavelets.

5. System variable selection and model term (wavelet regressor) determination

Variable and term selection are generic problems in nonlinear system identification. Once the significant variables have been selected, the model terms can be determined using a term selection algorithm operating over the selected variables, a parsimonious model structure can then be identified from the candidate model set, and finally the parameters can be estimated based on this model structure.

5.1 System variable selection

The first problem encountered in WANARX modelling is how to determine which variables should be included in the model. It is often the case in practice that some of the variables x_1, x_2, $\cdots$, x_n are redundant and only a subset of these variables is significant. Inclusion of redundant variables might result in a much more complex model since the number of model terms increases dramatically with the number of variables. Furthermore, including redundant variables might lead to a large number of free parameters in the model, and as a consequence the model may become oversensitive to training data and is likely to exhibit poor generalisation properties. Therefore, it is important to determine which variables should be included in the model.

The purpose of variable selection is to pre-select a subset consisting of the significant variables or to eliminate redundant variables from all the candidate variables of a system under study prior to model term detection. It is required that the selected significant variables alone should sufficiently represent the system. Based on these observations, a new effective variable selection algorithm (Wei et al. 2003b), has been proposed and can be used to select significant variables prior to fitting a WANARX model.

5.2 Model term determination

As explained in Section 4, the truncated regular wavelet frame, wavelet series and multiresolution wavelet decompositions can be converted into a linear-in-the-parameters form

$$y(t) = \sum_{m=1}^{M} \theta_m p_m(t) + e(t) \tag{57}$$

where $p_m(t)$ (m=1,2,...,M) are regressors (model terms) produced by the dilated and translated versions of mother wavelets or scaling functions, which are in the dictionary considered. Generally, not all the model terms make an equal contribution to the system output and terms, which make little contribution can be omitted. A

parsimonious representation, which contains only the significant terms, can often be obtained without the loss of representational accuracy by eliminating the redundant terms. Define

$$P^{(m)} = \{p_{i_k} : 1 \le i_k \le M;\ \ k = 1,2,\cdots,m\},\ \ m=1,2,\ldots,M, \tag{58}$$

The model term selection procedure is in fact an iterative process which searches through a nested term set in the sense that

$$P^{(1)} \subset P^{(2)} \subset \cdots \subset P^{(m)} \subset \cdots \tag{59}$$

This makes both the complexity and the accuracy of the representation based on these term sets to increase until a suitable term set is found, i.e., there exists an integer M_0 (generally $M_0 << M$), such that the model

$$y(t) = \sum_{k=1}^{M_0} \theta_{i_k} p_{i_k}(t) + e(t) \tag{60}$$

provides a satisfactory representation over the range considered for the measured input-output data.

A fast and efficient model structure determination approach has been implemented using the forward orthogonal least squares (OLS) algorithm and the error reduction ratio (ERR) criterion, which was originally introduced to determine which terms should be included in a model(Billings et al. 1988, 1989, Korenberg et al. 1988, Chen et al. 1989). This approach has been extensively studied and widely applied in nonlinear system identification (see, for example, Chen et al. 1991, Wang and Mendel 1992, Zhu and Billings 1996, Zhang 1997, Hong and Harris 2001). The forward OLS algorithm involves a stepwise orthogonalization of the regressors and a forward selection of the relevant terms in (57) based on the error reduction ration (ERR) (Billings et al. 1988, 1989).

6. Some practical issues associated with implementation

Emphasis is concentrated on wavelet series and multiresolution decompositions, and it is assumed that some compactly supported wavelets or/and scaling functions are considered in these decompositions. Some practical issues including data normalization, highest resolution level determination, translation parameter selection and wavelet dictionary determination are considered.

6.1 Data pre-processing

The original observational data $\tilde{x}(t) = [\tilde{x}_1(t), \tilde{x}_2(t), \cdots, \tilde{x}_n(t)]^T$ are often normalized into a standard domain, for example the unit hypercube $[0,1]^n$, for the convenience of problem description. This is especially true when a compactly supported wavelet and/or a scaling function are chosen in the wavelet series (29) and (38), and the multiresolution decomposition (39). Taking the univariate Haar wavelet (the first-order B-spline wavelet) as an example, it is much easier to select the starting resolution level and the range of the shift parameters if the sample data has been normalized to [0, 1].

Assume that the initial observations $\widetilde{x} \in R^n$ fall into the finite hypercube $[a_1, b_1] \times [a_2, b_2] \times \cdots [a_n, b_n]$, $\widetilde{x}(t)$ can be normalized into the unit hypercube $[0,1]^n$ by means of the following simple linear transform $x_i(t) = (\widetilde{x}_i(t) - a_i)/(b_i - a_i)$, $i = 1,2,\cdots,n$.

By another transform, $x_i(t) = [2\widetilde{x}_i(t) - (b_i + a_i)]/(\widetilde{x}_{i\max} - \widetilde{x}_{i\min})$, $i = 1,2,\cdots,n$, the original data $\widetilde{x}$ can be normalized into the standard hypercube $[-1,1]^n$, where $\widetilde{x}_{i\max} = \max_t \{x_i(t)\}_{t=1}^N$, $\widetilde{x}_{i\min} = \max_t \{x_i(t)\}_{t=1}^N$.

The modelling can then be performed in the standard hypercube $[0,1]^n$ or $[-1,1]^n$, and the model output can then be recovered to the original system operating domain by taking the inverse transform which converts x back into $\widetilde{x}$.

6.2 Determination of the highest multiresolution level

In theory, the wavelet series (29) and the multiresolution wavelet decomposition (35) are infinite expansions. In practice, however, it is impossible to include infinite terms in these wavelet decompositions. Therefore, the infinite decompositions are always truncated at appropriate dilations (resolutions) and translations.

Consider the one-dimensional multiresolution wavelet decomposition (35) and assume that the function $f(x)$ is defined in [0, 1] and x is an independent variable which is uniformly distributed in $[0,1]$, that is, x itself can be considered as "time", then the basis functions (dilated and translated versions of the wavelet and scaling function) in the multiresolution wavelet decomposition (35) are mutually orthogonal and the decomposition is unique. Assume also that the Haar wavelet (the first-order B-spline wavelet) and scaling function are used in the decomposition, then a truncated decomposition with the initial resolution scale j_0 and the highest resolution scale $j_{\max}=J$ can be expressed as

$$f(x) = \sum_{k=0}^{2^{j_0}-1} \alpha_{j_0,k} \phi_{j_0,k}(x) + \sum_{j=j_0}^{J} \sum_{k=0}^{2^j-1} \beta_{j,k} \varphi_{j,k}(x) \tag{61}$$

Clearly, the higher the upper resolution scale level J, the more accurate the approximation is. A recommended approach for selecting the highest scale J is to utilize the features of the sampled signal, for example, the natural frequency of the signal to be approximated. Assume that the maximum natural frequency of the sampled signals is $f_{\max}$, the highest scale can be empirically chosen as $j_{\max} = [\log_2(Mf_{\max})]$, where M is a positive number, say between 2^4 and 2^6, and $[\cdot]$ denotes taking the integer value of the corresponding number (Wei and Billings 2002).

In practical identification problems, however, the orthogonality of the mutiresolution wavelet decomposition might be lost, since most observational data fail to satisfy the uniform distribution assumption. Also in dynamical system modelling, the variable x in (61) is usually dependent on time t, and $x(t)$ often represents lagged outputs $y(t-p)$($p = 1,2,\cdots,n_y$) or lagged inputs $u(t-q)$ ($q = 1,2,\cdots,n_u$), which are usually sparse in the normalized interval [0, 1]. The empirical rule $j_{\max} = [\log_2(Mf_{\max})]$ for selecting the highest resolution scale can however still be used.

6.3 Shift parameter selection

For a compactly supported wavelet, the shift parameter k is determined by the corresponding resolution scale j. For example, at a given scale j, the shift parameter k in the Haar wavelet multiresolution decomposition (61) is chosen as $k = 0,1,\cdots,2^{j-1}$ (j=0,1,...). Generally, for a compactly supported wavelet $\varphi(x)$ with an integer support $S_\varphi = [0, K_s]$, where K_s is integer, the support for the dilated and translated wavelet $\varphi_{j,k}(x) = 2^{j/2}(2^j x - k)$ is $[2^{-j}k,\ 2^{-j}(K_s + k)]$, therefore, the shift parameter k at a resolution scale j should be taken as $-(K_s - 1) \le k \le 2^{j-1} - 1$. This is also true for a compactly supported scaling function $\phi(x)$.

6.4 Wavelet dictionary determination

Taking the truncated wavelet series (52) and the truncated mutiresolution wavelet decomposition (56) as an example. The elements of the wavelet dictionary are defined as the wavelet bases (dilated and translated versions of wavelets and scaling functions) involved in the decompositions. The number of all the dilated and translated versions of wavelets and/or scaling functions is defined as the length of the wavelet dictionary. The model terms in the approximation expressions are produced by some of the elements of the wavelet dictionary. Clearly, once the mother wavelets and/or scaling functions have been chosen, the wavelet dictionary is determined by the resolution scale parameter j and the shift parameter k. For compactly supported wavelets and scaling functions, the wavelet dictionary depends upon the initial resolution scale $j_{\min}$ and the highest resolution scale $j_{\max}$. Therefore, it is important to choose appropriate values for the initial resolution scale $j_{\min}$ and the highest resolution scale $j_{\max}$, since these values determine the degree of the complexity of the wavelet dictionary whatever types of wavelets are used. Theoretically, for a given initial resolution scale $j_{\min}$, the higher the upper resolution scale level $j_{\max}$, the more accurate the approximation is, however this may result in a more complex wavelet dictionary and thus a more complex decomposition, since too much resolution might result in a severely redundant wavelet dictionary or an over-fitted model.

In practice, for dynamical systems identification, the variable x in the wavelet function $\varphi_{j,k}(x)$ and the scaling function $\phi(x)$ is usually the lagged system inputs or/and outputs, and the observations of $x(t)$ are often sparsely distributed and therefore the problem can be ill-posed. This can produce a wavelet dictionary and the candidate model terms (regressors) that are redundant. However, the redundancy problem can be solved and the significant terms can be detected using a term detection algorithm.

7. Examples

In this section, two examples are provided to illustrate the application of the WANARX modelling structure. The input-output data used for identification in the first example are simulated from a nonlinear system with a known model; it is assumed, however, that no a priori information is available. The second example involves a

real system and the measurements taken from satellite data, correspond to the solar wind parameter VB_s (input) and the D_{st} index (output) for this terrestrial magnetospheric dynamic system.

7.1 Example 1—a nonlinear system disturbed by noise

Consider the following model

$$y(t)=\frac{0.5+y(t-1)}{1+y^2(t-1)}-\frac{2y(t-2)u(t-1)}{1+u^2(t-1)}+u(t-1)+\xi(t) \tag{62}$$

where $u(t)$ is an impulse sequence with random amplitude $A(t)$ and random duration $\Delta(t)$, $5\le A(t)\le 19$, $1\le \Delta(t)\le 40$; $\xi(t)$ is a noise sequence obeying a normal distribution with a standard derivation σ_ξ^2=0.0025. A data set consisted of 1000 input-output samples, which are illustrated in Figure 1, was generated by simulating the system. The data set was divided into two parts: the first 500 samples (from 1 to 500) were used for identification and the second part (from 501 to 1000) was used for testing.

The aim of the identification was to fit a WANARX model to describe the relationship between the input and output. The first step is to determine the significant variables which can sufficiently describe the relationship between the input and output. The variable selection algorithm of Wei et al. (2003b) was applied and the three significant variables: {$y(t$-1), $y(t$-2), $u(t$-1)} were selected. A one-dimensional WANARX model was therefore selected for this system

$$\begin{aligned} y(t) &= f(y(t-1),y(t-2),u(t-1))+e(t) \\ &= f_1(y(t-1))+f_2(y(t-2))+f_3(u(t-1))+e(t) \end{aligned} \tag{63}$$

Expanding each $f_i(\cdot)$ using the multiresolution wavelet decomposition (35), gives

$$f_i(x_i(t))=\sum_{k\in K^0}\alpha_{0,k}^{(i)}\phi_{0,k}(x_i(t))+\sum_{j=0}^{4}\sum_{k\in K_j}\beta_{j,k}^{(i)}\varphi_{j,k}(x_i(t)),\quad i=1,2,3, \tag{64}$$

where $x_1(t)=y(t-1)$, $x_2(t)=y(t-2)$, $x_3(t)=u(t-1)$; the 4th-order B-spline wavelet and scaling function were used in this decompostion, thus $K^0=\{-3,-2,-1,0\}$ and $K_j=\{-6,-5,\cdots,-1,\ 0,1,\cdots,2^j-1\}$ for j=0,1,2,3,4.

Although 195 basis functions (model terms) are involved in the one-dimensional WANARX model, only 13 of these were selected to be significant using the forward OLS algorithm. The final model contained only 13 terms (basis functions), which are listed in Table 1. A comparison of the model predicted outputs and the measurements, along with the model prediction errors over the test set, are shown in Figure 2. The model predicted output (MPO) of an identified NARX model is defined as

$$\hat{y}_{mpo}(t)=f(\hat{y}_{mpo}(t-1),\cdots,\hat{y}_{mpo}(t-n_y),u(t-1),\cdots,u(t-n_u)) \tag{65}$$

The model predicted outputs are recursively estimated and are used to calculate the model prediction errors

$$\hat{e}(t) = y(t) - \hat{y}_{mpo}(t) \tag{66}$$

where $y(t)$ (t=1,2,...,N) are the system measurements.

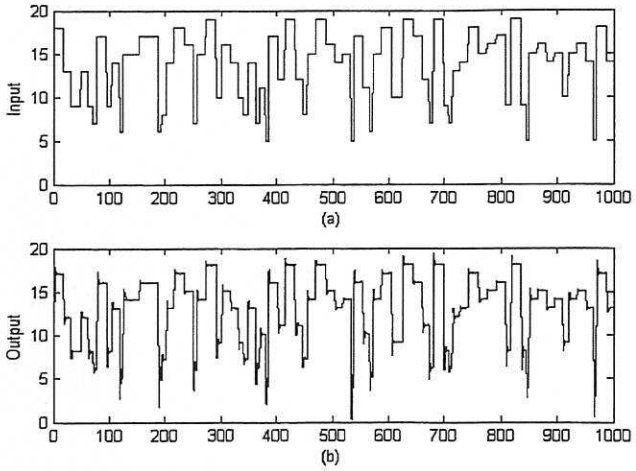


Figure 1 The system input and output for Example 1. (a) Input; (b) Output.

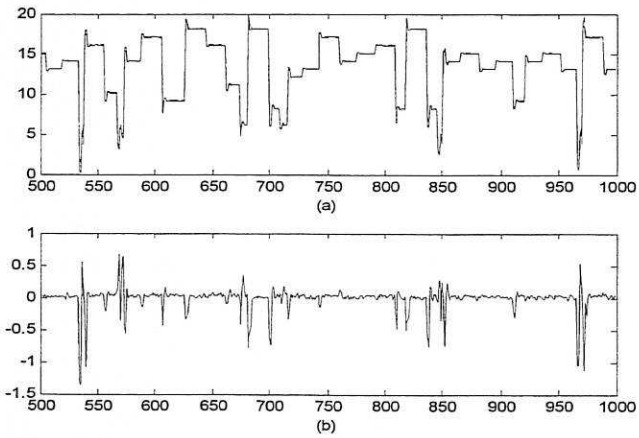


Figure 2 The model predicted output (MPO) and the model prediction errors for Example 1.
(a) Comparison of model predicted outputs and the measurements; (b) Model prediction errors.
(In (a), the solid line denotes the measurements, and the dashed line denotes the model predicted outputs.)

Table 1 The basis functions, parameters and the corresponding error reduction ratios for Example 1.

Search steps	Model terms	Parameters	ERRs $\times 100\%$
1	$\phi_{0,-1}(u(t-1))$	1.15884E+000	94.55335
2	$\phi_{0,0}(u(t-1))$	1.48091E+000	3.29547
3	$\varphi_{0,-1}(y(t-2))$	-5.26563E-001	2.10644
4	$\varphi_{0,-2}(y(t-2))$	1.33708e-001	0.01691
5	$\varphi_{0,-4}(u(t-1))$	5.43449E+000	0.00432
6	$\varphi_{1,-3}(y(t-2))$	1.93749E-002	0.00730
7	$\varphi_{1,-3}(y(t-1))$	-6.62867E-002	0.00205
8	$\varphi_{1,-1}(y(t-2))$	-2.83083E-002	0.00151
9	$\varphi_{3,-1}(y(t-1))$	6.52100E-003	0.00077
10	$\varphi_{2,-4}(y(t-2))$	6.14276E+000	0.00047
11	$\varphi_{1,-4}(u(t-1))$	1.80645E-002	0.00049
12	$\varphi_{3,-3}(u(t-1))$	1.87050E-002	0.00069
13	$\varphi_{3,-3}(y(t-2))$	1.71538E-001	0.00049
Note:	$\phi_{j,k}(x)=2^{j/2}\phi(2^j x-k)$ — the 4th-order B-splne functions; $\varphi_{j,k}(x)=2^{j/2}\varphi(2^j x-k)$ — the 4th-order B-splne wavelets.		

7.2 Example 2—a terrestrial magnetosphere dynamic system

The sun is a source of a continuous flow of charged particles, ions and electrons called the solar wind. The terrestial magnetic field shields the Earth from the solar wind, and forms a cavity in the solar wind flow that is called the terrestrial magnetosphere. The magnetopause is a boundary of the cavity, and its position on the day side (sunward side) of the magnetosphere can be determined as the surface where there is a balance between the dynamic pressure of the solar wind outside the magnetosphere and the pressure of the terrestrial magnetic field inside. A complex current system exists in the magnetosphere to support the complex structure of the magnetosphere and the magnetopause. Changes in the solar wind velocity, density or magnetic field lead to changes in the shape of the magnetopause and variations in the magnetospheric current system. In addition if the solar wind magnetic field has a component directed towards the south a reconnection between the terrestrial magnetic field and the solar wind magnetic field is initiated. Such a reconnection results in a very drastic modification to the magnetospheric current system and this phenomenon is referred to as magnetic storms. During a magnetic storm, which can last for hours, the magnetic field on the Earth's surface will change as a result of the variations of the magnetospheric current system. Changes in the magnetic field induce considerable currents in long conductors on the terrestrial surface such as power lines and pipe-lines. Unpredicted currents in power lines can lead to the blackouts of huge areas, the Ontario Blackout is just one recent example. Other

undesirable effects include increased radiation to crew and passengers on long flights, and effects on communications and radio-wave propagation. Forecasting geomagnetic storms is therefore highly desirable and can aid the prevention of such effects. The D_{st} index is used to measure the disturbance of the geomagnetic field in the magnetic storm. Numerous studies of correlations between the solar wind parameters and magnetospheric disturbances show that the product of the solar wind velocity V and the southward component of the magnetic field, quantified by B_s, represents the input that can be considered as the input to the magnetosphere. Denote the multiplied input by VB_s.

Figure 3 shows 1000 data points of measurement of the solar wind parameter VB_s (input) and the D_{st} index (output) with a sample period T=1hour. The purpose here is to identify a nonlinear model to represent the input-output relationship between VB_s (input) and D_{st}. The effects of other inputs on the system will be neglected in the present study. A variable selection algorithm of Wei et al. (2003b) was applied and nine significant variables, {y(t-1), y(t-2), y(t-3), y(t-4), y(t-5), y(t-6), y(t-7), u(t-1), u(t-2)} were selected. These nine variables are used to form a hybrid WANARX model for the data set

$$\begin{aligned} y(t) &= f(y(t-1), y(t-2), \cdots, y(t-7), u(t-1), u(t-2)) + e(t) \\ &= a_0 + \sum_{i=1}^{9} a_i x_i(t) + \sum_{i=1}^{9}\sum_{j=i}^{9} b_{ij} x_i(t) x_j(t) \\ &\quad + \sum_{i=1}^{9} f_i(x_i(t)) + \sum_{i=1}^{8}\sum_{j=i+1}^{9} f_{ij}(x_i(t), x_j(t)) + e(t) \end{aligned} \tag{67}$$

where $x_i(t) = y(t-i)$ for i=1,2,...,7 and $x_i(t) = y(t-i+7)$ for i=8,9, f_i and f_{ij} are unknown univariate and bivariate functions which can be approximated by one- and two-dimensional wavelet decompositions. In this example, both the input and output data points were initially normalized and the modelling procedure was performed on the standard hypercube $[0,1]^n$, where n=9. The first 500 input-output data points were used for model identification and the remaining 500 data points were used for testing. By expanding each f_i and f_{ij} using the wavelet series decompositions (52)(the 4th-order B-spline scaling functions were used in each decomposition), model (67) can be converted into a linear-in-the-parameters problem and the unknown parameters can be estimated using the forward OLS algorithm. The final identified model, which involved 16 regressors selected from 891 candidate terms, was of the form

$$y(t) = \theta_1 y(t-1) + \sum_{i=2}^{16} \theta_i B_i(t) \tag{68}$$

where $B_i(t)$ (i=2,3, ...,16) are wavelet regressors formed by the 4th-order B-spline scaling functions, and θ_i (i=1,2,...,16) are the parameters. The terms, parameters and corresponding ERR values are listed in Table 2. Notice again that each variable in the model (67) and (68) was initially normalized to $[0,1]$, and the model outputs were recovered to the original system operating domain by taking inverse transforms.

In practice the one-step-ahead (one-hour-ahead) predictions for the D_{st} index are not useful, since it is difficult during a few minutes to collect all data from both satellite measurements and ground based magnetometers and to feed them into the model (68) to obtain predictions. On the other hand, forecasting the D_{st} index several months ahead of the real measurements is not required. To be practically useful, the predictions should be made

on some time scale which is intermediate between the two extreme cases. A 12-hour-ahead prediction based on (68) is considered here. The comparisons between the 12-step-ahead predictions, the model predicted outputs and the measurements are shown in Figure 4. As expected the model predicted outputs are not as good as the 12-step-ahead predictions, but the model predicted outputs provide good long term predictions and give confidence in the identified model. The discrepancy between the model predicted outputs and the measured values of the D_{st} index are believed to be the result of other inputs which affect the system output but were not included in the current model.

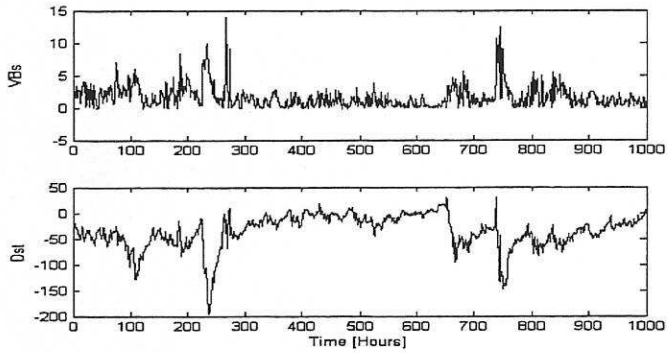


Figure 3 The input and output data of the terrestrial magnetospheric dynamic system in Example 2

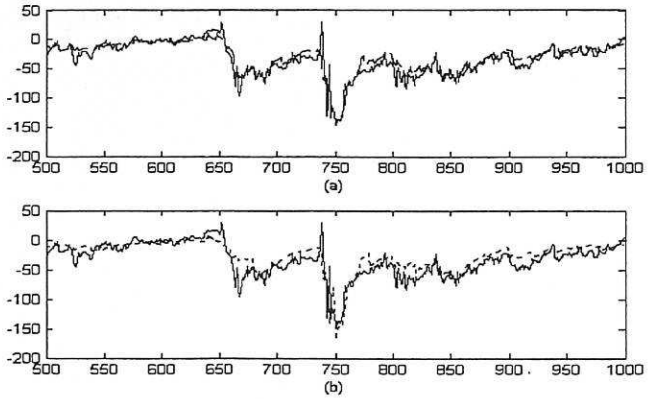


Figure 4 Comparisons of the six-step-ahead predictions, model predicted outputs and the measurement for the solar wind Dst index in Example 2. (a) 12-step-ahead predictions; (b) Model predicted outputs. (Solid—measurements; Dashed—12-step-ahead predicted outputs; Dotted—model predicted outputs)

Table 2 The selected model terms, estimated parameters and the corresponding ERR values for the system in Example 2

Number	$B_i(t)$	θ_i	$ERR_i \times 100\%$
1	$y(t-1)$	6.10269e-001	95.65172
2	$\phi_{0,-1}(y(t-1))\phi_{0,-2}(u(t-1))$	6.39257e-001	2.06315
3	$\phi_{5,17}(u(t-1))$	2.17571e-003	1.02247
4	$\phi_{0,-3}(y(t-5))\phi_{0,0}(y(t-6))$	-4.09044e+001	0.41470
5	$\phi_{0,0}(y(t-7))\phi_{0,0}(u(t-2))$	7.36766e+001	0.09880
6	$\phi_{5,19}(u(t-1))$	4.01684e-002	0.02400
7	$\phi_{0,0}(u(t-1))\phi_{0,0}(u(t-2))$	-4.50903e+001	0.00962
8	$\phi_{5,18}(u(t-1))$	-5.89649e-002	0.00300
9	$\phi_{5,16}(u(t-1))$	-4.60957e-002	0.00368
10	$\phi_{5,13}(u(t-1))$	-4.82462e-002	0.00308
11	$\phi_{0,0}(y(t-2))\phi_{0,0}(u(t-2))$	-5.93993e+001	0.00746
12	$\phi_{5,16}(y(t-7))$	6.68900e-003	0.00343
13	$\phi_{0,0}(y(t-2))\phi_{0,-3}(y(t-3))$	5.40887e+000	0.00327
14	$\phi_{5,14}(u(t-1))$	1.51620e-002	0.00328
15	$\phi_{0,0}(y(t-3))\phi_{0,-2}(y(t-4))$	-6.02775e+000	0.00223
16	$\phi_{0,0}(y(t-2))\phi_{0,-2}(y(t-4))$	2.87946e+000	0.00345
Note:	$\phi_{j,k}(x) = 2^{j/2}\phi(2^j x - k)$ — the 4th-order B-spline scaling functions		

8. Conclusions

A unified wavelet-based NARX model structure, which incorporates wavelet networks, wavelet series and wavelet mutiresolution decompositions, has been introduced for nonlinear input-output system identification. The new WANARX model structure allows high-order nonlinear systems to be expressed as a sum of additive low-dimensional submodels. This in some sense partially alleviates the difficulty of the *curse-of-dimensionality* for high-order nonlinear system modelling.

Every functional component in each of the additive submodels can be decomposed using a wavelet frame decomposition, wavelet series or wavelet multiresolution decomposition. Emphasis in the present study focused on wavelet series and wavelet multiresolution decompositions, and a semi-orthogonal multiresolution wavelet decomposition (wavelet series) structure based on B-spline wavelets was recommended as a powerful approximation approach for a wide range of nonlinear systems. By expanding each functional component in the WANARX model using multiresolution wavelet decompositions, the model identification and parameter estimation problem can be converted into a linear-in-the-parameters problem, and an iterative model structure detection procedure coupled with the forward OLS algorithm and the ERR criteria can be used to efficiently select the significant model terms (wavelet regressors) and estimate the parameters simultaneously.

The new modelling approach based on the WANARX model structure has been successfully applied in nonlinear system identification and two examples were provided to demonstrate the applicability and effectiveness of this new modelling approach.

Acknowledgment

The authors gratefully acknowledge that part of this work was supported by EPSRC. We are grateful to Dr M. Balikhin for providing the magnetosphere data.

Appendix A

The B-spline wavelets

B-splines are piece-wise polynomial functions with good local properties, and were originally introduced by Chui and Wang (1992) as wavelet and scaling functions in multiroselution expansions.

The B-spline function of m th order is defined by the following recursive formula:

$$N_m(x) = \frac{x}{m-1} N_{m-1}(x) + \frac{m-x}{m-1} N_{m-1}(x-1), \; m \geq 2 \tag{A1}$$

with

$$N_1(x) = \chi_{[0,1)}(x) = \begin{cases} 1 & if \; x \in [0,1) \\ 0 & otherwise \end{cases} \tag{A2}$$

Setting N_m as the scaling function, that is, $\phi(x) = N_m(x)$, then both the wavelet and the scaling function can be expressed in terms of the scaling function $N_m(x)$ as follows

$$\phi(x) = \sum_{k=0}^{m} c_k N_m(2x-k) \tag{A3}$$

$$\varphi(x) = \sum_{k=0}^{3m-2} d_k N_m(2x-k) \tag{A4}$$

with the coefficients given by

$$c_k = \frac{1}{2^{m-1}} \binom{m}{k} \tag{A5}$$

$$d_k = \frac{(-1)^k}{2^{m-1}} \sum_{j=0}^{m} \binom{m}{j} N_{2m}(k-j+1), \; k = 0,1,\cdots,3m-2 \tag{A6}$$

Clearly, the support of the m th order B-spline wavelet and the associated scaling function are

$$\begin{cases} \text{supp}\,\phi = \text{supp} N_m = [0,m] \\ \text{supp}\,\varphi = [0,2m-1] \end{cases} \tag{A7}$$

Both the B-spline wavelets and the associated scaling functions are symmetric or anti-symmetric within the supports. The most commonly used B-spline wavelets are the linear $(m=2)$ and cubic $(m=4)$ cases, both of which can be expressed explicitly.

References

Billings, S.A. and Leontaritis,I.J.(1982), Parameter estimation techniques for nonlinear systems, *The 6th IFAC Symposium on Identification and Systems Parameter Estimation*, Washington, pp 427-432.

Billings, S.A., Korenberg, M. and Chen, S.(1988), Identification of nonlinear output-affine systems using an orthogonal least-squares algorithm, *International Journal of Systems Science*, **19(8)**,1559-1568.

Billings,S.A., Chen,S. and Korenberg,M.J.(1989), Identification of MIMO non-linear systems suing a forward regression orthogonal estimator, *International Journal of Control*, **49(6)**,2157-2189.

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

Chen,S., Billings,S.A., and Luo,W.(1989), Orthogonal least squares methods and their application to non-linear system identification, *International Journal of Control*, **50(5)**,1873-1896.

Chen,S.,Billings, S.A.,Cowan, C.F.N., and Grant, P.W.(1990), Nonlinear system identification using radial basis functions, *International Journal of Systems Science.*, **21(12)**, 2513-2539.

Chen, S, Cowan, C.F.N., Grant, P.M. (1991), Orthogonal least-squares learning algorithm for radial basis function networks, *IEEE Trans Neural Networks*, 2 (2), 302-309.

Chen,S.,Billings, S. A. and Grant, P. W.(1992), Recursive hybrid algorithm for nonlinear system identification using radial basis function network, *International Journal of Control.*, **55(5)**,1051-1070.

Chui,C. K.,1992, *An Introduction to Wavelets.* Boston; London : Academic Press.

Chui, C. K. and Wang, J. H.(1992), On compactly supported spline wavelets and a duality principle, *Trans. of the American Mathematical Society*, **330(2)**, 903-915.

Coca, D. and Billings, S.A.(2001), Non-linear system identification using wavelet multiresolution models, *International Journal of Control*, **74(18)**,1718-1736.

Daubechies,I.(1992), *Ten lectures on wavelets.* Philaelphia, Pennsylvania : Society for Industrial and Applied Mathematics.

Friedman,J.H. and Stuetzle, W.(1981), Projection pursuit regression, *Journal of the American Statistical Association*, **76(376)**, 817-823.

Friedman,J. H.(1991), Multivariate adaptive regression splines, *The Annals of Statistics*, **19(1)**, 1-67.

Gorban, A.N.(1998), Approximation of continuous functions of several variables by an arbitrary nonlinear continuous function of one variable, linear functions, and their supperpositions, *Applied Mathematics Letters*, **11(3)**, 45-49.

Haykin,S.(1994), *Neural networks: a comprehensive foundation.* New York : Macmillan; Oxford : Maxwell Macmillan International.

Hastie,T.J. and Tibshirani, R.J.(1990), *Generalized additive models* (London; Glasgow; Weinheim; New York; Tokyo; Melbourne; Madras: Chapman & Hall).

Hong, X. and Harris, C. J.(2001), Nonlinear model structure detection using optimum experimental design and orthogonal least squares, *IEEE Transactions On Neural Networks*, **12(2)**, 435-439.

Kavli, T. (1993), ASMOD—An algorithm for adaptive spline modelling of observational data, *International Journal of Control*, **58(4)**,947-967.

Korenberg, M., Billings, S.A., Liu, Y. P. and McIlroy P.J.(1988), Orthogonal parameter estimation algorithm for non-linear stochastic systems, *International Journal of Control*, **48(1)**,193-210.

Leontaritis,I.J. and Billings, S.A.(1985), Input-output parametric models for non-linear systems, (part I: deterministic non-linear systems; part II: stochastic non-linear systems), Int. Journal of Control, **41(2)**,303-344.

Liu, G.P., Billings, S.A. and Kadirkamnathan,V.(2000), Nonilear system identification using wavelet networks, *International Journal of Systems Science*, **31(12)**, 1531-1541.

Mallat,S.G.(1989),A theory for multiresolution signal decomposition: the wavelet representation, *IEEE Trans. On Pattern analysis and machine intelligence*, **11(7)**,674-693.

Pearson, R. K.(1995),Nonlinear input/output modelling, *Journal of Process Control*, **5(4)**, 197-211.

Pearson, R.K.(1999), *Descrete-time dynamic models*, New York; Oxford: Oxford University Press.

Schumaker, L.L.(1981), *Spline Functions: Basic theory*. New York: John Wiley & Sons.

Wang, L.X. and Mendel, J.M.(1992), Fuzzy basis functions, universal approximations, and orthogonal least squares learning, *IEEE Trans Neural Networks*, **3(5)**,807-814.

Wei, H.L., and Billings, S.A.(2002), Identification of time-varying systems using multi-resolution wavelet models, *International Journal of Systems Science*,**33(15)**,1217-1228.

Wei, H.L., Billings, S.A., and Balikhin, M.A.(2003a), Wavelet-based nonparametric models for nonlinear input-output system identification, Research Report No 831, Department of Automatic Control and Systems Engineering, the University of Sheffield, UK (submitted to *International Journal of Systems Science*).

Wei, H.L., Billings, S.A., and Liu J. (2003b), Term and variable selection for nonlinear system identification, Research Report No 837, Department of Automatic Control and Systems Engineering, the University of Sheffield, UK

Zhang, Q., and Benveniste,A.(1992), Wavelet networks, *IEEE Trans. Neural Networks*, **3(6)**, 889-898.

Zhang,Q. (1997),Using wavelet network in nonparametric estimation, *IEEE Trans. Neural Networks*, **8(2)**, 227-236.

Zhu, Q.M. and Billings, S.A.(1996), Fast orthogonal identification of nonlinear stochastic models and radial basis function neural networks, *International Journal of Control*, **64(5)**,871-886.