Monograph:
Wei, H.L. and Billings, S.A. (2006) A comparative study on global wavelet and polynomial models for nonlinear regime-switching systems. Research Report. ACSE Research Report no. 940 . Automatic Control and Systems Engineering, University of Sheffield

---

A Comparative Study on Global Wavelet and Polynomial Models for Nonlinear Regime-Switching Systems

H. L. Wei and S. A. Billings

Email: {S.Billings@Shef.ac.uk, W.Hualiang@Shef.ac.uk}

Research Report No. 940

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

September 8, 2006

A Comparative Study on Global Wavelet and Polynomial NARX Models for Nonlinear Regime-Switching Systems

Hua-Liang Wei and Stephen A. Billings

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

S.Billings@Sheffield.ac.uk, W.Hualiang@Sheffield.ac.uk

September 8 , 2006

Abstract—A comparative study of wavelet and polynomial models for nonlinear regime-switching (RS) systems is carried out. Regime-switching systems, considered in this study, are a class of severely nonlinear systems, which exhibit abrupt changes or dramatic breaks in behavior, due to regime switching caused by associated events. Both wavelet and polynomial models are used to describe discontinuous dynamical systems, where it is assumed that no a priori information about the inherent model structure and the relative regime switches of the underlying dynamics is known, but only observed input-output data are available. An orthogonal least squares (OLS) algorithm interfered with by an error reduction ratio (ERR) index and regularised by an approximate minimum description length (AMDL) criterion, is used to construct parsimonious wavelet and polynomial models. The performance of the resultant wavelet models is compared with that of the relative polynomial models, by inspecting the predictive capability of the associated representations. It is shown from numerical results that wavelet models are superior to polynomial models, in respect of generalization properties, for describing severely nonlinear regime-switching systems.

Keywords—NARX models; Nonlinear system identification; Regime-switching systems; Wavelets.

1. Introduction

Nonlinear regime-switching (RS) systems, considered in this study, are a class of severely nonlinear dynamical systems, which exhibit abrupt changes or jumps in behavior, due to regime switching driven by associated events. Nonlinear regime-switching behaviour exists widely in both engineering and non-engineering processes. More often, both the inherent model structure and the relative regime switches of the underlying processes are totally unknown or very little is known about them, but observations for the system inputs and outputs are available. System identification techniques can thus be applied to obtain an equivalent input-output representation for the underlying systems. General modelling frameworks including the ARX and ARMAX models (Ljung 1987, Söderström and Stoica 1989), the NARMAX model (Leontaritis and Billings 1985a, 1985b, Chen and Billings 1989, Pearson 1999), neural networks (Billings and Chen 1998, Liu 2001) and neurofuzzy

networks (Harris et al. 2002), and other techniques (Cherkassky and Mulier 1998), can be used to construct such an equivalent input-output representation. In cases where the main objective of system modelling is focused on stability analysis and controller design, some specific model types, for example multiple models or multimode models (Sontag 1981, Billings and Voon 1987, Murry-Smith and Johansen 1997, Bemporad et al. 2000) may be more appropriate for regime-switching dynamical systems, but to obtain such a specific model some a priori information on the inherent model structure and the relative individual regime switches of the underlying systems may be required. These specific model types will not be the pivot of this study; on the contrary, global model types will be used to describe the input-output behaviour of given regime-switching systems, under an assumption that either the inherent model structure or the relative regime switches for the underlying processes are totally unknown.

One of the most commonly used approaches for modelling a structure-unknown nonlinear system is to construct a nonlinear model using some specific types of functions including polynomials, kernel functions, wavelets and other candidate basis functions. In practice, most types of functions can only be used to approximate certain nonlinear relationships effectively. In some cases, however, the nonlinear dynamics can not sufficiently be represented by a given class of functions because of the lack of good approximation properties. It is generally recognized that the basis functions used for representing general nonlinear functions should offer some flexibility in adapting to the complexity of the model structure so that the model can match, as closely as possible, the underlying dynamics. Wavelet techniques are one of the most popular and powerful tools for complex nonlinear signal processing. Compared with other basis functions, wavelets with localization in both the time and the frequency domains, possess several uniquely attractive properties and offer a flexible capability for approximating arbitrary functions.

This study introduces a new wavelet based modelling framework for nonlinear regime-switching systems, where polynomial models may lack good approximation properties. Wavelet models are constructed using wavelet basis functions selected from a prescribed dictionary. The dictionary may consist of a large number of candidate bases, but in many cases only a small number of significant bases need to be included in the wavelet model for a given nonlinear identification problem. An orthogonal least squares (OLS) algorithm interfered with by an error reduction ratio (ERR) index (Billings *et al*. 1989, Chen *et al*. 1989) and regularised by an approximate minimum description length (AMDL) criterion (Saito 1994, Antoniadis et al. 1997) , is used to select significant bases (model regressors). The performance of the resultant sparse wavelet models is compared with that of polynomial models, by inspecting the predictive capability of the associated representations. As will be seen, wavelet models are superior to polynomial models, in respect of generalization properties, for describing nonlinear regime-switching systems.

2. Regime-switching systems and the NARX model

Regime-switching systems, considered in this study, are a class of complex nonlinear dynamic systems, where possibly there exist discontinuities. Assume that the problem is defined in the space *S*, referring to as the problem space. Let $S_1, S_2, \cdots, S_p$ be a partition of *S*, with $\bigcup_{i=1}^{p} S_i = S$ and $S_i \bigcap S_j = \varnothing$ if $i \neq j$. Due to the effects of changes in either the internal state variables or exogenous input variables, the underlying system may present different local dynamics at each subspace $S_i \in S$, and often needs to be described using different local models. Such a system can be represented using the multiple model form below

$$y(t) = \begin{cases} f^1(y(t-1), \cdots, y(t-n_y), u(t-1), \cdots, u(t-n_u), \boldsymbol{\theta}_1), & \xi \in S_1 \\ f^2(y(t-1), \cdots, y(t-n_y), u(t-1), \cdots, u(t-n_u), \boldsymbol{\theta}_2), & \xi \in S_2 \\ \vdots \qquad\qquad\qquad \vdots \qquad\qquad\qquad \vdots & \\ f^p(y(t-1), \cdots, y(t-n_y), u(t-1), \cdots, u(t-n_u), \boldsymbol{\theta}_p), & \xi \in S_p \end{cases} \tag{1}$$

where $f^i(\cdot)$ (*i*=1,2, ..., *p*) are different linear or nonlinear functions, which are often unknown and which need to be identified from given observations of the input $u(t)$ and the output $y(t)$, n_u and n_y are the maximum input and output lags, $e(t)$ is the noise sequence, ξ is a vector formed by part or all of the lagged input and output variables $\{y(t-1), \cdots y(t-n_y), u(t-1), \cdots, u(t-n_u)\}$, $\boldsymbol{\theta}_i$ is the associated parameter vector of the *i*th local model.

If a priori information on the inherent local model structure and the individual regime switches of the underlying systems are available, the multiple model (1), may be identified directly from given input-output data. If, however, the underlying processes are totally structure-unknown in either the local model structure or the regime switches, global model types then need to be considered to describe the input-output behaviour of given regime-switching systems.

A wide class of input-output nonlinear dynamical systems can be represented by the NARX (N*onlinear* A*uto*R*egressive* with *e*X*ogenous* inputs) model of the form (Leontaritis and Billings 1985a, 1985b, Pearson 1999)

$$y(t) = f(y(t-1), \cdots, y(t-n_y), u(t-1), \cdots, u(t-n_u)) + e(t) \tag{2}$$

where the nonlinear mapping f is often unknown and needs to be identified from given observational data of the input $u(t)$ and the output $y(t)$, n_u, n_y and $e(t)$ are defined as in (1). The nonlinear mapping f can be constructed using a variety of local or global basis functions including polynomials, kernel functions, splines, radial basis functions, neural networks and wavelets. One of the most popular representations is the well-known Kolmogorov-Gabor polynomial model (Leontaritis and

Billings 1985a, 1985b), which takes the form below

$$y(t)=\theta_0+\sum_{i_1=1}^{d}\theta_{i_1}x_{i_1}(t)+\sum_{i_1=1}^{d}\sum_{i_2=i_1}^{d}\theta_{i_1i_2}x_{i_1}(t)x_{i_2}(t)+\cdots+\sum_{i_1=1}^{d}\cdots\sum_{i_\ell=i_{\ell-1}}^{d}\theta_{i_1i_2\cdots i_\ell}x_{i_1}(t)x_{i_2}(t)\cdots x_{i_\ell}(t)+e(t) \qquad (3)$$

where

$$x_k(t)=\begin{cases}y(t-k), & 1\le k\le n_y\\ u(t-k+n_y), & n_y+1\le k\le d=n_y+n_u\end{cases} \qquad (4)$$

The nonlinear degree of the polynomial model (3) is referred to be ℓ, which is determined by the highest order of all the candidate model terms. A more general representation for the multivariate nonlinear function f in the NARX model (1) is to decompose f into a number of functional components via the well-known functional analysis of variance (ANOVA) expansions (Friedman 1991, Chen 1993, Li *et al*. 2001, Wei and Billings 2004)

$$y(t)=\sum_{i=1}^{d}f_i(x_i(t))+\sum_{1\le i<j\le d}f_{i,j}(x_i(t),x_j(t))+\sum_{1\le i<j<k\le d}f_{i,j,k}(x_i(t),x_j(t),x_k(t))+\cdots$$

$$+\sum_{1\le i_1<\cdots<i_m\le d}f_{i_1,i_2,\cdots,i_m}(x_{i_1}(t),x_{i_2}(t),\cdots,x_{i_m}(t))+e(t) \qquad (5)$$

where $m\le d$, $i_m\in\{1,2,\cdots,d\}$ and the function $f_{i_1,\cdots,i_m}$ (j=1,2, …,d) does not contain terms that are included in functional components with an order smaller than *m*. Detailed discussions on the functional ANOVA expansion (5) can be found in Billings and Wei (2005).

Many types of functions can be employed to express the functional components $f_{i_1,\cdots,i_m}$ in model (5). In this study, however, wavelet decompositions will be used to approximate each of these functional components. Experience shows that the representation of up to second order of functional components in model (5), using wavelet decompositions, can often provide a satisfactory approximation for many high dimensional problems providing that the input variables are properly selected (Wei and Billings 2004, Wei et al. 2004a, Billings and Wei 2005).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.

It is known that wavelet decompositions are based on a mother wavelet prototype function, and temporal analysis is performed using some contracted, high-frequency versions of the same function. Data analysis can thus be implemented using the corresponding wavelet coefficients. The ANOVA expansion (5), where each functional component is approximated using wavelet decompositions, can thus be easily converted into a linear-in-the-parameters form

$$y(t)=\sum_{m=1}^{M}\theta_m\phi_m(t)+e(t) \qquad (6)$$

where $\mathbf{x}(t)=[x_1(t),x_2(t),\cdots,x_d(t)]^T$ is the 'input' (predictor) vector, $\phi_m(t)=\phi_m(\mathbf{x}(t))$ are the model regressors, θ_m are the model parameters, and *M* is the total number of candidate regressors.

The initial wavelet model (6) often involves a large number of candidate model terms. Experience suggests that most of the candidate model terms can be removed from the model, and that only a small number of significant model terms are needed to provide a satisfactory representation for most nonlinear dynamical systems. The orthogonal least square type algorithms (Billings *et al*. 1989, Chen *et al*. 1989) can be used to select significant model terms. The initial OLS-ERR type algorithms, however, cannot automatically determine the model size. To ameliorate the agility and enhance the capability of the OLS-ERR algorithm, an approximate minimum description length (AMDL) criterion (Saito 1994, Antoniadis et al. 1997), will be introduced to aid the determination of the associated model size, and this is described below.

3. The OLS-ERR algorithm

Consider the term selection problem for the linear-in-the-parameters model (6). Let $\{(\mathbf{x}(t),y(t)):\mathbf{x}\in\mathbf{R}^d,y\in\mathbf{R}\}_{t=1}^N$ be a given training data set and $\mathbf{y}=[y(1),\cdots,y(N)]^T$ be the vector of the output. Let $I=\{1,2,\cdots,M\}$, and denote by $\Omega=\{\phi_m:m\in I\}$ the dictionary of candidate model terms in an initially chosen candidate regression model similar to (6). The dictionary Ω can be used to form a variant vector dictionary $\mathcal{D}=\{\boldsymbol{\varphi}_m:m\in I\}$, where the *m*th candidate basis vector $\boldsymbol{\varphi}_m$ is formed by the *m*th candidate model term $\phi_m\in\Omega$, in the sense that $\boldsymbol{\varphi}_m=[\phi_m(\mathbf{x}(1)),\cdots,\phi_m(\mathbf{x}(N))]^T$. The model term selection problem is equivalent to finding, from *I*, a subset of indices, $I_n=\{i_m:m=1,2,\cdots,n,i_m\in I\}$ where $n\le M$, so that so that $\mathbf{y}$ can be approximated using a linear combination of $\boldsymbol{\alpha}_{i_1},\boldsymbol{\alpha}_{i_2},\cdots,\boldsymbol{\alpha}_{i_n}$.

3.1 The forward orthogonal search procedure

A non-centralised squared correlation coefficient will be used to measure the dependency between two associated random vectors. The non-centralised squared correlation coefficient between two vectors **x** and **y** of size *N* is defined as

$$C(\mathbf{x},\mathbf{y})=\frac{(\mathbf{x}^T\mathbf{y})^2}{\|\mathbf{x}\|^2\|\mathbf{y}\|^2}=\frac{(\mathbf{x}^T\mathbf{y})^2}{(\mathbf{x}^T\mathbf{x})(\mathbf{y}^T\mathbf{y})}=\frac{(\sum_{i=1}^N x_iy_i)^2}{\sum_{i=1}^N x_i^2\sum_{i=1}^N y_i^2} \tag{7}$$

It has been shown in Wei et al. (2004b) that the above squared correlation coefficient is closely related to the error reduction ratio (ERR) criterion (a very useful index in respect to the significance of model terms), defined in the standard orthogonal least squares (OLS) algorithm for model structure selection (Billings et al.1989, Chen et al. 1989).

The model structure selection procedure starts from equation (6). Let $\mathbf{r}_0=\mathbf{y}$, and

$$\ell_1 = \arg \max_{1 \le j \le M} \{C(\mathbf{y}, \boldsymbol{\varphi}_j)\} \tag{8}$$

where the function $C(\cdot,\cdot)$ is the correlation coefficient defined by (7). The first significant basis can thus be selected as $\boldsymbol{\alpha}_1 = \boldsymbol{\varphi}_{\ell_1}$, and the first associated orthogonal basis can be chosen as $\mathbf{q}_1 = \boldsymbol{\varphi}_{\ell_1}$. The model residual, related to the first step search, is given as

$$\mathbf{r}_1 = \mathbf{r}_0 - \frac{\mathbf{y}^T \mathbf{q}_1}{\mathbf{q}_1^T \mathbf{q}_1} \mathbf{q}_1 \tag{9}$$

In general, the *m*th significant model term can be chosen as follows. Assume that at the (*m*-1)th step, a subset $\mathcal{D}_{m-1}$, consisting of (*m*-1) significant bases, $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \cdots, \boldsymbol{\alpha}_{m-1}$, has been determined, and the (*m*-1) selected bases have been transformed into a new group of orthogonal bases $\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_{m-1}$ via some orthogonal transformation. Let

$$\mathbf{q}_j^{(m)} = \boldsymbol{\varphi}_j - \sum_{k=1}^{m-1} \frac{\boldsymbol{\varphi}_j^T \mathbf{q}_k}{\mathbf{q}_k^T \mathbf{q}_k} \mathbf{q}_k \tag{10}$$

$$\ell_m = \arg \max_{j \ne \ell_k, 1 \le k \le m-1} \{C(\mathbf{y}, \mathbf{q}_j^{(m)})\} \tag{11}$$

where $\boldsymbol{\varphi}_j \in \mathcal{D} - \mathcal{D}_{m-1}$, and $\mathbf{r}_{m-1}$ is the residual vector obtained in the (*m*-1)th step. The *m*th significant basis can then be chosen as $\boldsymbol{\alpha}_m = \boldsymbol{\varphi}_{\ell_m}$ and the *m*th associated orthogonal basis can be chosen as $\mathbf{q}_m = \mathbf{q}_{\ell_m}^{(m)}$. The residual vector $\mathbf{r}_m$ at the *m*th step is given by

$$\mathbf{r}_m = \mathbf{r}_{m-1} - \frac{\mathbf{y}^T \mathbf{q}_m}{\mathbf{q}_m^T \mathbf{q}_m} \mathbf{q}_m \tag{12}$$

Subsequent significant bases can be selected in the same way step by step. From (12), the vectors $\mathbf{r}_m$ and $\mathbf{q}_m$ are orthogonal, thus

$$\| \mathbf{r}_m \|^2 = \| \mathbf{r}_{m-1} \|^2 - \frac{(\mathbf{y}^T \mathbf{q}_m)^2}{\mathbf{q}_m^T \mathbf{q}_m} \tag{13}$$

By respectively summing (12) and (13) for *m* from 1 to *n*, yields

$$\mathbf{y} = \sum_{m=1}^{n} \frac{\mathbf{y}^T \mathbf{q}_m}{\mathbf{q}_m^T \mathbf{q}_m} \mathbf{q}_m + \mathbf{r}_n \tag{14}$$

$$\| \mathbf{r}_n \|^2 = \| \mathbf{y} \|^2 - \sum_{m=1}^{n} \frac{(\mathbf{y}^T \mathbf{q}_m)^2}{\mathbf{q}_m^T \mathbf{q}_m} \tag{15}$$

The model residual $\mathbf{r}_n$ will be used to form a criterion for model selection, and the search procedure will be terminated when the norm $\| \mathbf{r}_n \|^2$ satisfies some specified conditions. Note that the quantity $\mathrm{ERR}_m = C(\mathbf{y}, \mathbf{q}_m)$ is just equal to the *m*th error reduction ratio (Billings et al. 1989, Chen et al. 1989),

brought by including the mth basis vector $\boldsymbol{\alpha}_m = \boldsymbol{\varphi}_{\ell_m}$ into the model, and that $\sum_{m=1}^{n} C(\mathbf{y}, \mathbf{q}_m)$ is the increment or total percentage that the desired output variance can be explained by $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \cdots, \boldsymbol{\alpha}_n$.

Note that some tricks can be used to avoid selecting strongly correlated model terms. Assume that at the (m-1)th step, a subset $\mathcal{D}_{m-1}$,consisting of m-1 significant bases, $\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_{m-1}$, has been determined. Also assume that $\boldsymbol{\varphi}_j \in \mathcal{D} - \mathcal{D}_{m-1}$ is strongly correlated with some bases in $\mathcal{D}_{m-1}$, that is, $\boldsymbol{\varphi}_j$ is a linear combination of $\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_{m-1}$. Thus, $(\mathbf{q}_j^{(m)})^T \mathbf{q}_j^{(m)} = 0$. In the implementation of the algorithm, the candidate basis $\boldsymbol{\varphi}_j$ will be automatically discarded if $(\mathbf{q}_j^{(m)})^T \mathbf{q}_j^{(m)} < \delta$, where δ is a positive number that is sufficiently small. In this way, any severe mullticolinearity or ill-conditioning can be avoided.

3.2 Model size determination

The determination of model size is critical in dynamical modelling because neither an over-fitting nor an under-fitting model is desirable. For problems in the real world, however, the true model size is generally unknown and needs to be estimated from the data. Model selection criteria are often established on the basis of estimates of prediction errors, by inspecting how the identified model performs on future (never used) data sets.

In the present study, an approximate minimum description length (AMDL) criterion developed by Saito (1994) and Antoniadis et al. (1997), on the basis of the Rissanen's MDL criterion (Rissanen 1983), will be used to determine the model size. For the case of single regression model, AMDL is defined as

$$\mathrm{AMDL}(n) = 0.5 \log_2[\mathrm{MSE}(n)] + \frac{1.5 n \log_2 N}{N} = 0.5 \log_2\left(\frac{\| \mathbf{r}_n \|^2}{N} \right) + \frac{1.5 n \log_2 N}{N} \tag{16}$$

where MSE is the mean-square-error from the associated model, N is the length of the associated training data set, n is the number of model terms, and $\mathbf{r}_n$ is the associated model residual. A similar strategy has been employed in Wei et al. (2006), where a Bayesian information criterion (BIC) criterion was used.

3.3 Parameter estimation

It is easy to verify that the relationship between the selected original bases $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \cdots, \boldsymbol{\alpha}_n$, and the associated orthogonal bases $\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_n$, is given by

$$\mathbf{A}_n = \mathbf{Q}_n \mathbf{R}_n \tag{17}$$

where $\mathbf{A}_n = [\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_n]$, $\mathbf{Q}_n$ is an $N \times n$ matrix with orthogonal columns $\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_n$, and $\mathbf{R}_n$ is an $n \times n$ unit upper triangular matrix whose entries $u_{ij} (1 \le i \le j \le n)$ are calculated during the orthogonalization procedure. The unknown parameter vector, denoted by $\boldsymbol{\theta}_n = [\theta_1, \theta_2, \cdots, \theta_n]^T$, for the

model with respect to the original bases, can be calculated from the triangular equation $\mathbf{R}_n\boldsymbol{\theta}_n = \mathbf{g}_n$ with $\mathbf{g}_n = [g_1, g_2, \cdots, g_n]^T$, where $g_k = (\mathbf{y}^T\mathbf{q}_k)/(\mathbf{q}_k^T\mathbf{q}_k)$ for k=1,2, …, *n*.

4. Numerical examples

This section presents three examples to demonstrate that wavelet models, produced by the forward orthogonal regression (OLS-ERR) algorithm, can be used to effectively describe severely nonlinear regime-switching systems where possibly there exist discontinuities. As will be seen, resultant wavelet models are superior to polynomial models, in respect of generalization properties, for nonlinear regime-switching systems considered in the examples.

Note that in the polynomial model identification procedure the original observational data were directly used to construct the model. In the wavelet modelling procedure, however, the original observed data, if not in [0,1], were initially normalized to [0,1] via a transform $x(t) = (\tilde{x}(t) - a)/(b - a)$, where $\tilde{x}(t)$ indicate the initial observations, and *a* and *b* represent the prescribed boundary for the associated observations. The identification procedure was therefore performed using normalized values *x*(*t*). The outputs of an identified model were then recovered to the original measurement space by taking the associated inverse transform.

The *model predicted output* (MPO) were used to measure the model performance of the identified models. For an identified model $y(t) = \hat{f}(\mathbf{x}(t))$, the model predicted output is defined as $\hat{y}(t) = \hat{f}(\hat{\mathbf{x}}(t))$, implying that $\hat{y}(t)$ is produced from the identified model iteratively, where $\mathbf{x}(t) = [x_1(t), x_2(t), \cdots, x_d(t)]^T$ is defined by (3) and (5), and $\hat{\mathbf{x}}(t) = [\hat{x}_1(t), \cdots, \hat{x}_{n_y}(t), x_{n_y+1}(t) \cdots, x_d(t)]^T$ is the predicted value of $\mathbf{x}(t)$.

4.1 Example 1.

Consider a two-input piecewise finite impulse response (FIR) model below

$$f(x_1(t), x_2(t)) = \begin{cases} u_1(t-1) + u_2(t-1), & (x_1, x_2) \in S_1 \\ u_1(t-1) - 2u_2(t-1), & (x_1, x_2) \in S_2 \\ 3u_1(t-1) + 2u_2(t-1), & (x_1, x_2) \in S_3 \end{cases}, \tag{18}$$

where $x_1(t) = u_1(t-1)$, $x_2(t) = u_2(t-1)$, and the three regimes (subspaces) S_1, S_2 and S_3 are defined as below:

$S_1 = \{(x_1, x_2) : 0 \le x_1 \le x_2 \le 1, x_2 \le 0.5\}$,

$S_2 = \{(x_1, x_2) : 0 \le x_1 \le 1, 0.5 < x_2, 1 \le x_1 + x_2\}$,

$S_3 = \{(x_1, x_2) : x_1 \le 0.5, x_1 < x_2, x_1 + x_2 < 1\}$,

Clearly, $S_1 \bigcup S_2 \bigcup S_3 = [0,1]\times[0,1]$, see Figure 1 for a clearer visualisation. One thousand data points were generated from (18) by setting $y(t) = f(x_1, x_2) + e$, where the two input variables $u_1(t)$ and $u_2(t)$ were uniformly distributed in [0,1], and *e* was a Gaussian noise with zero mean and standard derivation 0.1. The distribution of the 1000 input data points in the three subspaces S_1, S_2 and S_3 is shown in Figure 1, and the first return map formed by the 1000 data points is shown in Figure 2. The predictor vector (the 'input' vector) was chosen to be $\mathbf{x}(t) = [x_1(t), x_2(t)]^T$, and the initial polynomial and wavelet model was respectively chosen to be

$$y(t) = \hat{f}(\mathbf{x}(t)) = \theta_0 + \sum_{i_1=1}^{2} \theta_{i_1} x_{i_1}(t) + \sum_{i_1=1}^{2} \sum_{i_2=i_1}^{2} \theta_{i_1 i_2} x_{i_1}(t) x_{i_2}(t) + \cdots + \sum_{i_1=1}^{2} \cdots \sum_{i_5=i_4}^{2} \theta_{i_1 \cdots i_5} x_{i_1}(t) \cdots x_{i_5}(t) \tag{19}$$

and

$$y(t) = \hat{f}(\mathbf{x}(t)) = \sum_{j=0}^{4} \left\{ \sum_{k_1 \in B_j} \sum_{k_2 \in B_j} c_{j;k_1,k_2} \psi_{j;k_1,k_2}(x_1(t), x_2(t)) \right\} \tag{20}$$

where $B_j = \{k : -3 \le k \le 2 + 2^j\}$, $c_{j;k_1,k_2}$ are coefficients, $\psi_{j;k_1,k_2}(x_1, x_2) = 2^j \psi(2^j x_1 - k_1, 2^j x_2 - k_2)$ are the dilated and translated versions of the 2-D truncated Mexican hat wavelet, $\psi(x_1, x_2)$, that has been proposed in Billings and Wei (2005) and is defined as below:

$$\psi(\mathbf{x}) = \begin{cases} (2 - \| \mathbf{x} \|^2) e^{-\frac{1}{2}\|\mathbf{x}\|^2}, & \mathbf{x} \in [-3,3]\times[-3,3] \\ 0, & \text{otherwise} \end{cases} \tag{21}$$

The total number of candidate model terms (basis functions) involved in the initial polynomial model (19) and the initial wavelet model (20) was 21 and 893, respectively. Based on the 1000 data points, the OLS-ERR algorithm was applied to select significant model terms for both the polynomial and the wavelet model identification. The AMDL criterion, shown in Figure 3, suggested that the number of model terms for the polynomial and wavelet models was 17 and 36 respectively.

To inspect and compare the performance of the identified 17-term polynomial model and 36-term wavelet model, both models were simulated by choosing the two input variables $u_1(t)$ and $u_2(t)$ as below:

- Both $u_1(t)$ and $u_2(t)$ were uncorrelated random sequences, with 500 data points, uniformly distributed in [0,1], but note that the test data was different from the data used for model estimation.
- $u_1(t) = 0.5 + 0.4\sin(\pi t / 50)$ and $u_2(t) = 0.5 + 0.4\sin(\pi t / 20 + \pi / 3)$ for t=1,2,…,500. The 500 input data points in the three subspaces S_1, S_2 and S_3 is shown in Figure 4.

Model predicted outputs, from both the identified polynomial and wavelet models, were compared with that produced from the true noise-free model (18). The predicted results, corresponding to the

above two test cases, are shown in Figures 5 and 6, respectively, where only a fraction of the data points is displayed for a closer visualisation. For the first test case, the mean-square-error (MSE), for the model predicted outputs from the identified polynomial and wavelet models, was calculated to be 0.2768 and 0.0923, respectively, over all the 500 test data points. For the second test case, the MSE was calculated to be 0.3719 and 0.1073, respectively, over the test data set. Clearly, the identified wavelet model is significantly superior to the polynomial model for the regime-switching system described by (18).

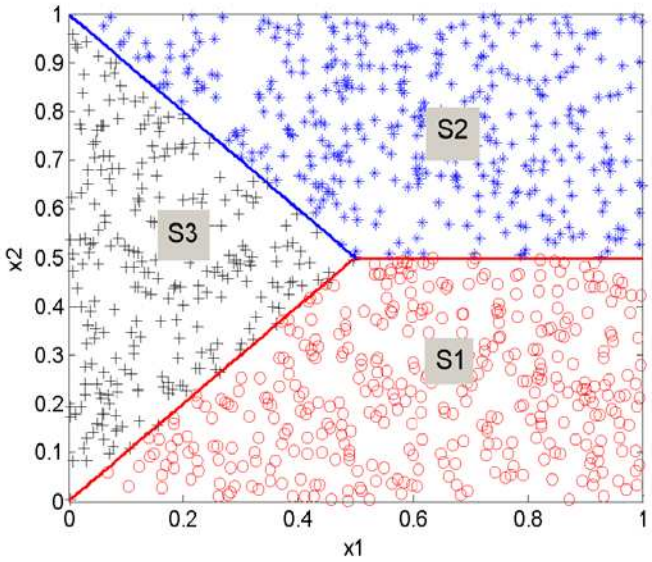

Figure 1. The three subspaces S_1, S_2 and S_3, and the distribution of the 1000 input data points used for model estimation described in Example 1.

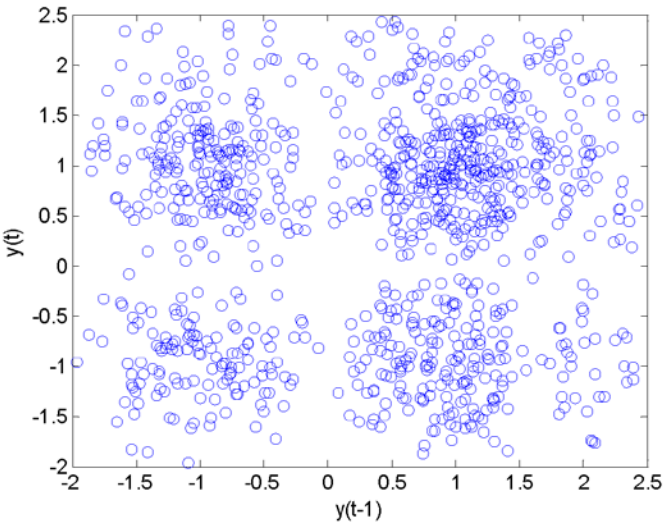

Figure 2. The first return map formed by the 1000 training data points used for model estimation described in Example 1.

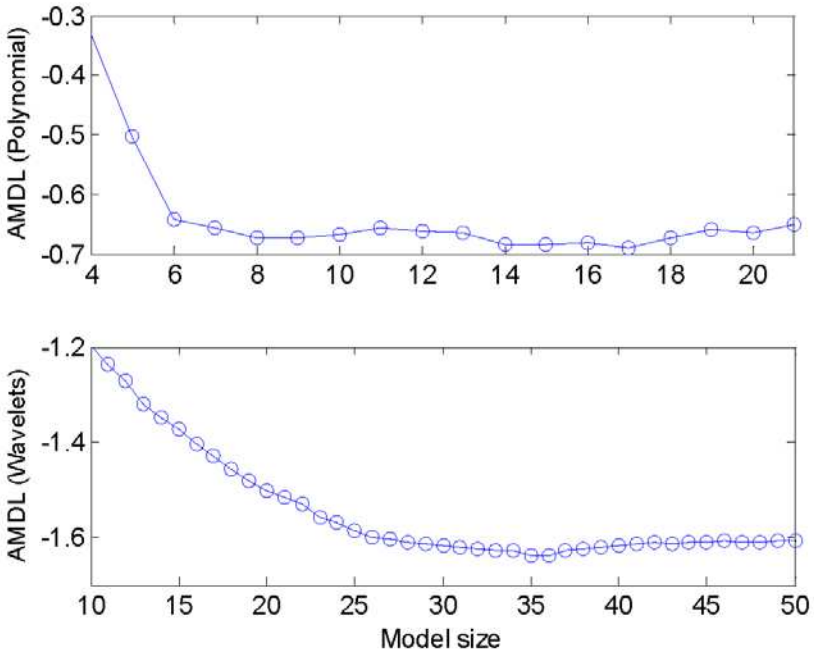

Figure 3. AMDL index versus the number of model terms for the polynomial models (the plot at the top) and the wavelet models (the plot at the bottom) identification described in Example 1.

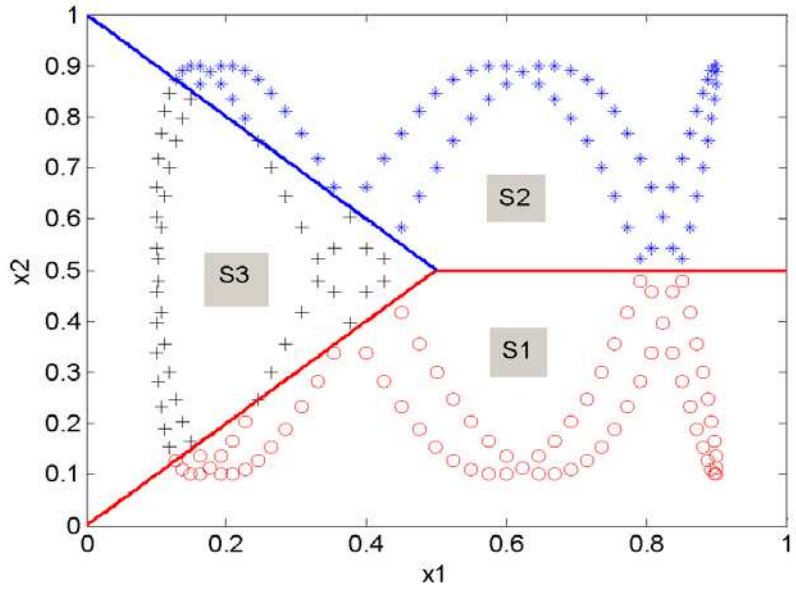

Figure 4. The distribution of the 500 input data points (in the sine wave input case) used for model test in Example 1.

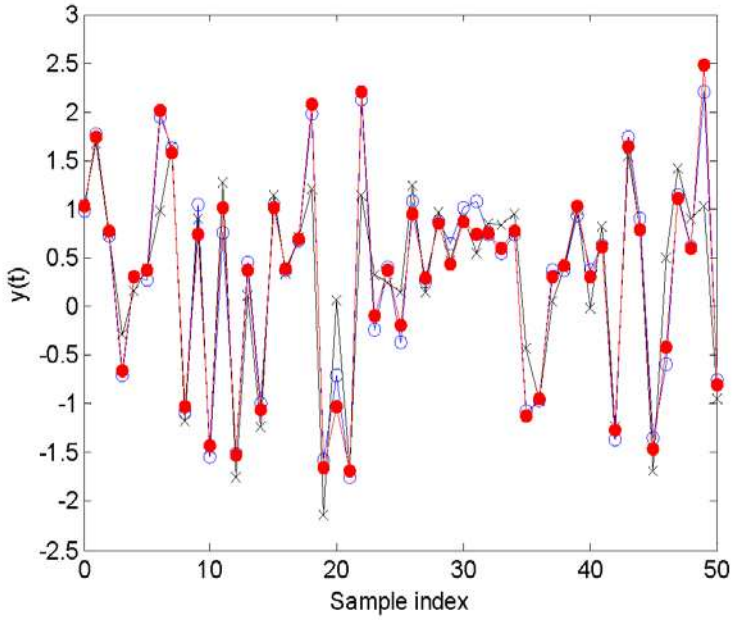

Figure 5. A comparison of the model predicted outputs produced by the identified polynomial and wavelet models, with the true values produced by the noise-free model (18), driven by the input in the first test case described in Example 1. Circles present the true values, dots present the model predicted output from the wavelet model, and crosses present the model predicted output from the polynomial model.

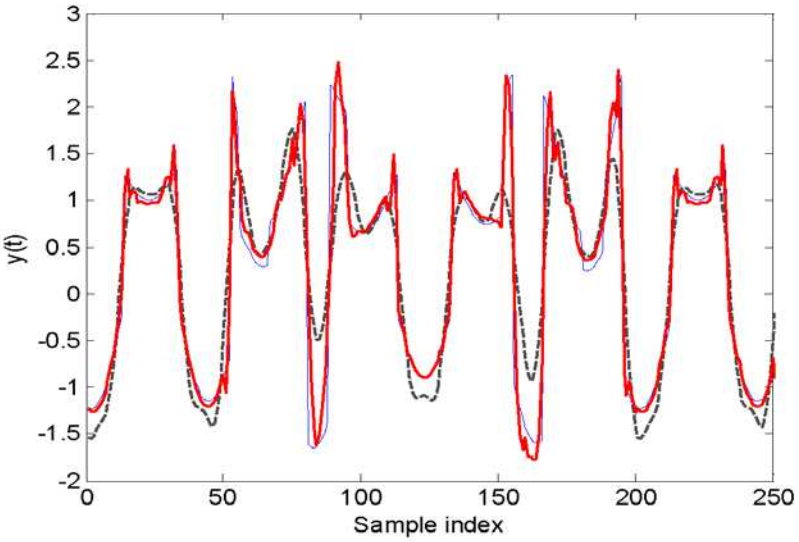

Figure 6. A comparison of the model predicted outputs produced by the identified polynomial and wavelet models, with the true values produced by the noise-free model (18), driven by the input in the second test case described in Example 1. The thin solid line presents the true values, the thick solid line resents the model predicted output from the wavelet model, and thick dashed line represents the model predicted output from the polynomial model.

4.2 Example 2

A nonlinear system was described by the following model

$$x(t)=\begin{cases}\dfrac{x(t-1)[x(t-1)-0.5x(t-2)]}{10+[x^2(t-1)+x^2(t-2)]^{1/2}}+u(t-1)+\xi(t), & \text{if } x(t-1)x(t-2)\ge 0,\\ \dfrac{x(t-2)[x(t-1)+0.5x(t-2)]}{10+[x^2(t-1)+x^2(t-2)]^{1/2}}+u(t-1)+\xi(t), & \text{if } x(t-1)x(t-2)<0,\end{cases} \tag{22a}$$

$$y(t)=x(t)+\eta(t) \tag{22b}$$

where $\xi(t)$ and $\eta(t)$ were Gaussian white noise with zero mean and standard deviation σ_{ξ}=0.1 and σ_{η}=0.5, respectively. By choosing the input $u(t)$ as a random sequence that was uniformly distributed in [-10,10], model (22) was simulated and 500 input-output data points were collected after the system has settled down.

The predictor vector was chosen to be $\mathbf{x}(t)=[x_1(t),x_2(t),x_3(t)]^T=[y(t-1),y(t-2),u(t-1)]^T$. The initial polynomial model was chosen to be

$$y(t)=\hat{f}(\mathbf{x}(t))=\theta_0+\sum_{i_1=1}^{3}\theta_{i_1}x_{i_1}(t)+\sum_{i_1=1}^{3}\sum_{i_2=i_1}^{3}\theta_{i_1i_2}x_{i_1}(t)x_{i_2}(t)+\cdots+\sum_{i_1=1}^{3}\cdots\sum_{i_4=i_3}^{3}\theta_{i_1\cdots i_4}x_{i_1}(t)\cdots x_{i_4}(t) \tag{23}$$

A total of 35 candidate model terms was involved in this polynomial model. The initial wavelet model was chosen to be

$$y(t)=\hat{f}(\mathbf{x}(t))=\sum_{p=1}^{3}f_p(x_p(t))+\sum_{p=1}^{2}\sum_{q=2}^{3}f_{pq}(x_p(t),x_q(t)) \tag{24}$$

where the univariate functions f_p were of the form below

$$f_p(x_p(t))=\sum_{j=0}^{5}\sum_{k\in A_j}c_{j;k}\phi_{j;k}(x_p(t)) \tag{25}$$

where $A_j=\{k:-4\le k\le 3+2^j\}$, $c_{j;k}$ are coefficients, $\phi_{j;k}(x)=2^{j/2}\phi(2^jx-k)$ are the dilated and translated versions of the 1-D truncated Mexican hat wavelet, $\phi(x)$, defined as below (Billings and Wei 2005):

$$\phi(x)=\begin{cases}(1-x^2)e^{-x^2/2} & x\in[-4,4]\\ 0 & \text{otherwise}\end{cases} \tag{26}$$

The bivariate functions f_{pq} were approximated using the 2-D truncated Mexican hat wavelets similar to (20). A total of 3012 candidate wavelet basis functions were involved in the initial wavelet model (24). Note that original observed data were initially normalized to [0,1], via the transform

$x_i(t) = (\tilde{x}_i(t) - a_i)/(b_i - a_i)$, where $\tilde{x}_i(t)$ indicates the initial observations, and $a_i = -15$, $b_i = 15$ for i=1,2 (corresponding to the output variables), and $a_i = -10$, $b_i = 10$ for i=3 (corresponding to the input variable). The identification procedure was performed using the normalized values of $x_i(t)$. The outputs of an identified model were then recovered to the original measurement space by taking the associated inverse transform.

Based on the 500 estimation data points, the OLS-ERR algorithm was applied to select significant model terms for both the polynomial and the wavelet model identification. The AMDL criterion, shown in Figure 7, suggested that the number of model terms for the polynomial and wavelet models was 8 and 18 respectively.

To inspect and compare the performance of the identified 8-term polynomial model and the 18-term wavelet model, both models were simulated by choosing the input as a random sequence, with 500 data points, that was uniformly distributed in [-10,10]. Model predicted outputs, from both the polynomial and wavelet modes, were compared with that produced by the true model (22), and these are shown in Figure 8, where again only part data points were displayed. The MSE was calculated to be 0.7588 and 0.6870, respectively, with respect to the model predicted outputs from the identified polynomial and wavelet models, over the all 500 test data points. Clearly, the identified wavelet model is apparently superior to the polynomial model for the regime-switching system described by (22).

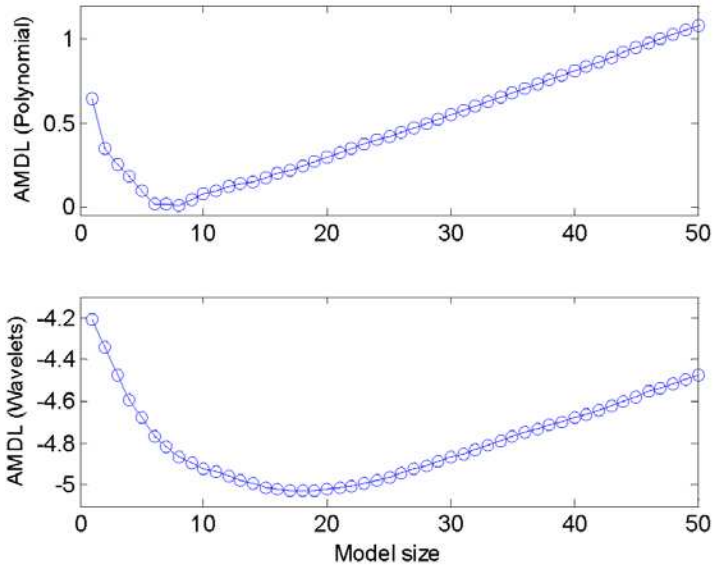

Figure 7. AMDL index versus the number of model terms for polynomial model (the plot at the top) and wavelet model (the plot at the bottom) identification described in Example 2. The AMDL criterion for the wavelet model was calculated in the normalised space.

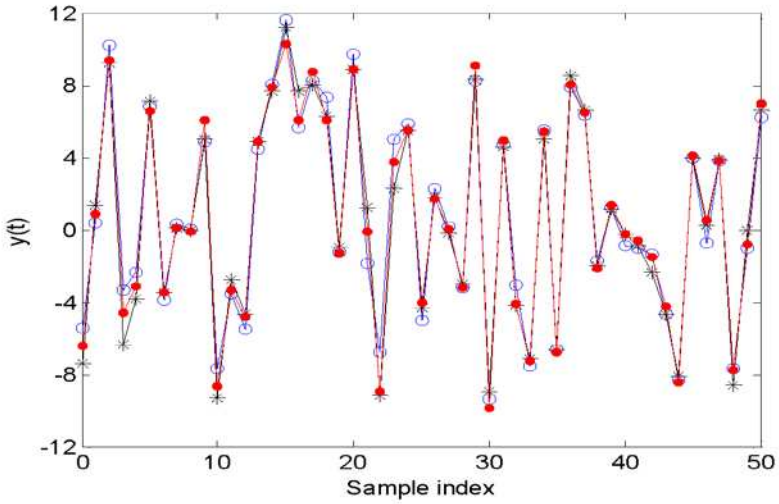

Figure 8. A comparison of the model predicted outputs produced by the identified polynomial and wavelet models, with the true values produced by model (22), described in Example 2. Circles present the true values, dots present the model predicted output from the wavelet model, and stars present the model predicted output from the polynomial model.

4.3 Example 3

A nonlinear system model was given by

$$x(t)=\begin{cases} -a_0+a_1x(t-1)-a_2x(t-2)-u(t-1), & \zeta\in S_1 \\ b_0+b_1x(t-1)-b_2x(t-2)+u(t-1), & \zeta\in S_2 \\ b_0-b_1x(t-1)-b_2x(t-2)+u(t-1), & \zeta\in S_3 \end{cases}, \tag{27a}$$

$$y(t)=x(t)+e(t) \tag{27b}$$

where $a_0=10$, $a_1=0.25$, $a_2=0.75$, $b_0=20$, $b_1=0.6$, $b_2=0.8$, $e(t)$ was a noise signal, $\zeta=[x(t-1),x(t-2)]$, and the three subspaces S_1, S_2 and S_3 are defined as below:

$$S_1=\{(x_1,x_2): x_2\ge 0.25x_1^2\},\ S_2=\{(x_1,x_2): x_2<0.25x_1^2, x_1\le 0\},\ S_3=\{(x_1,x_2): x_2<0.25x_1^2, x_1>0\},$$

Clearly the union of S_1, S_2 and S_3 is the whole plane in the 2-D space. By setting $e(t)$ to be Gaussian white noise with zero mean and standard deviation σ_e =5, and by choosing the input $u(t)$ as a random sequence that was uniformly distributed in [-5,5], model (27) was simulated and 2500 input-output data points were collected after the system has settled down. The first 500 data points was used for model estimation and the remaining 2000 data points were used for model validation. The first return maps produced by the 2000 noisy test data and the associated noisy-free data are shown in Figures 9 and 10, respectively.

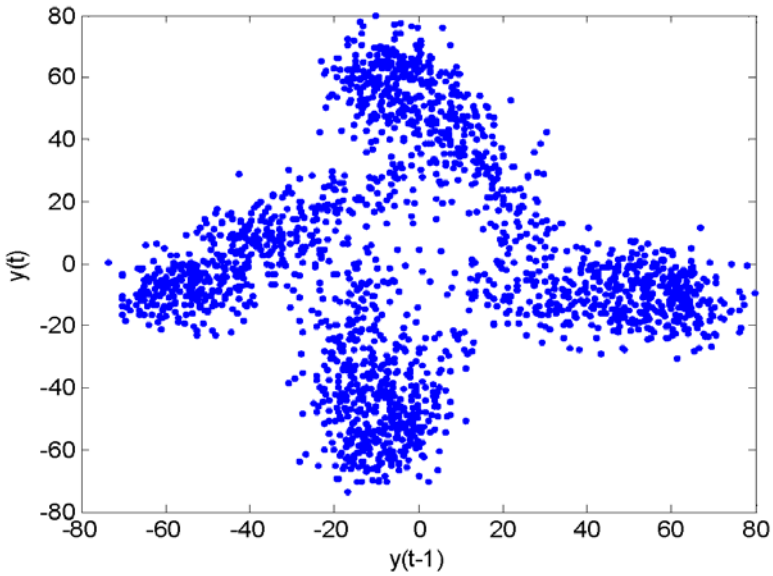

Figure 9. The first return map formed by the 2000 noisy training data points generated by model (27) in Example 3.

-

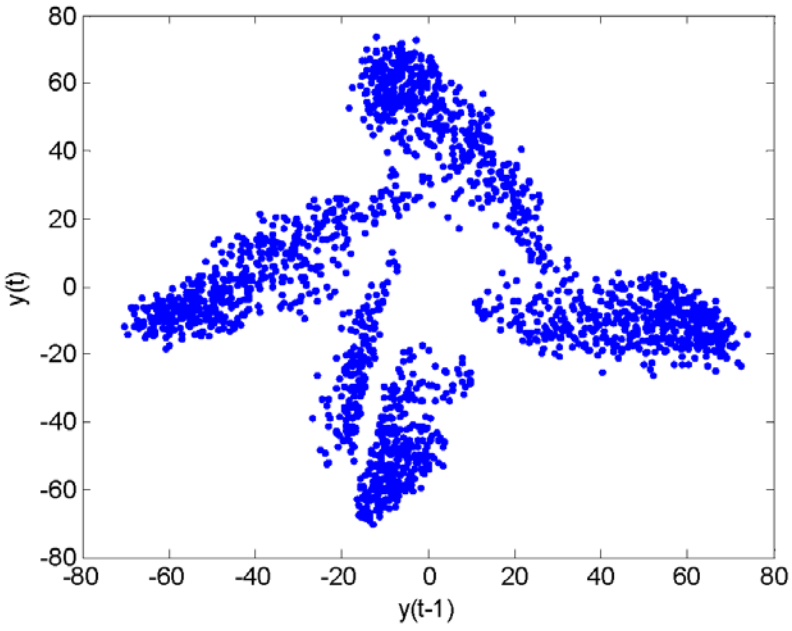

Figure 10. The first return map form by the associated noisy-free data (2000 points) generated by model (27) in Example 3.

Note that the output of the model (27a) was sensitive to the model parameters. For example, a small disturbance in the parameter b_2 may lead to a dramatic difference in the model response under the same input. Figure 11 presents this phenomenon, where the difference $\varepsilon(t) = y(t; b_2 + \delta) - y(t; b_2)$, with b_2=0.8 and $\delta = 10^{-4}$, was displayed. Clearly, with the process going on, a small discrepancy in the parameter b_2 produces significant difference in the model response. This means that it might be difficult to identify a model that can produce accurate long term predictions.

The predictor vector, the initial polynomial model, and the initial wavelet model were chosen exactly the same as in Example 2. The original observed data were initially normalized to [0,1], by setting $a_i = -80$ and $b_i = 80$ for i=1,2 (corresponding to the output), and $a_i = -5$ and $b_i = -5$ for i=3 (corresponding to the input). The identification procedure was performed using normalized values of $x_i(t)$. The outputs of an identified model were then recovered to the original measurement space by taking the associated inverse transform. Based on the 500 data points, the OLS-ERR algorithm was applied to select significant model terms for both the polynomial and the wavelet model identification. The AMDL criterion suggested that the number of model terms for the polynomial and wavelet models was 12 and 21 respectively.

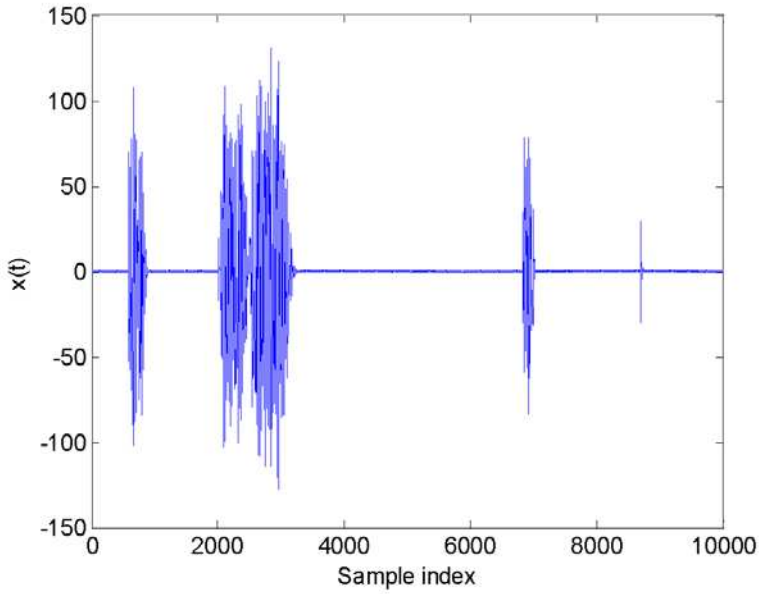

Figure 10. The difference $\varepsilon(t) = y(t; b_2 + \delta) - y(t; b_2)$ produced by the noise-free model (27a), described in Example 3, where $\delta = 10^{-4}$.

It was known that the original model given by (27) is sensitive to the associated model parameters, it was thus difficult to obtain accurate long term predictions. This means that a simple comparison of long term predictions may not be a good measurement of model quality. As an alternative, the performance of the identified 12-term polynomial model and the 21-term wavelet model was inspected by calculating the associated first return maps. The first return maps, produced by the model predicted outputs, with respect to the identified polynomial and wavelet models, are presented in Figures 12 and 13, respectively. Clearly, the first return map produced by the identified wavelet model, compared with that produced by the polynomial model, provides a much closer representation for the first return map formed by the original measurements. This means that the identified wavelet model is superior to the associated polynomial model for the nonlinear regime-switching system described by (27).

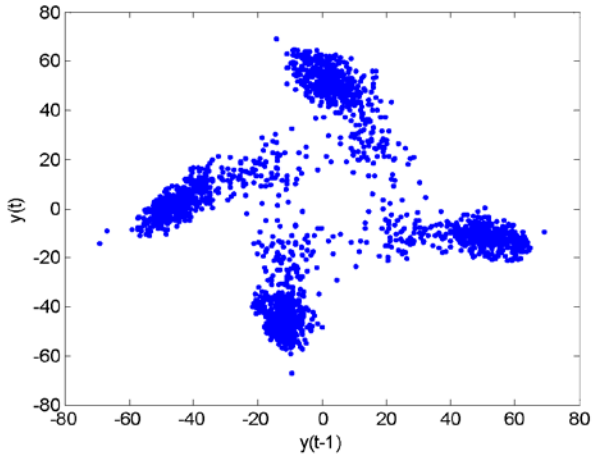

Figure 12. The first return map produced by the model predicted output of 2000 data points from the identified polynomial model, driven by a random sequence that was uniformly distributed in [-5,5].

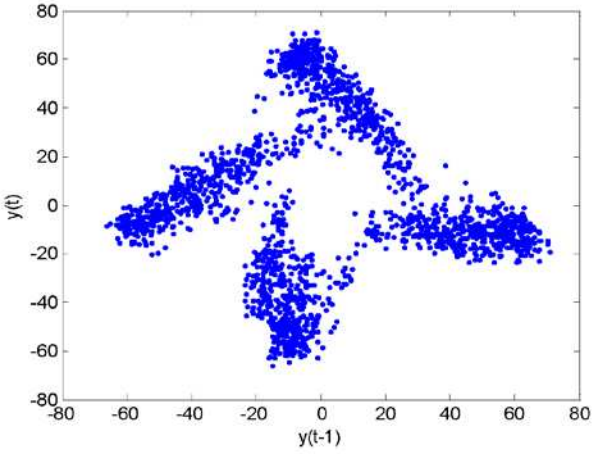

Figure 13. The first return map produced by the model predicted output of 2000 data points from the identified wavelet model, driven by a random sequence that was uniformly distributed in [-5,5].

5. Conclusions

A new wavelet based modelling approach has been introduced for nonlinear regime-switching system identification, where it was assumed that the inherent model structure and relative regime switches of the underlying systems are totally unknown. For this type of severely nonlinear systems, where a priori knowledge on the model structure is unavailable, identifying a global input-output model, from available data, may be a good initial step towards further understanding of the underlying dynamics. It has been demonstrated that wavelets, with local properties in both the time and the frequency domains, are powerful for constructing global nonlinear models for regime-switching systems. The performance of the resultant wavelet models, as has been shown, is much superior to that of the associated polynomial models for dealing with the identification problems described in the examples.

At first sight, a wavelet model may involve a large number of candidate model terms for high dimensional problems. In most cases, however, most of the candidate model terms are redundant and only a small number of significant model terms are necessary to include in the final model to describe the nonlinear dynamics with a good accuracy. More fortunately, an efficient model structure and term selection algorithms including the forward orthogonal regression (OLS-ERR) algorithm, coupled with the AMDL criterion (or other efficient criteria), can be employed to determine which model terms and how many model terms should be included in the model and finally a parsimonious model can be obtained.

Wavelet bases, in dynamical wavelet modelling, involve two key parameters: the scales (the dilation parameter) and the positions (the translation parameter). While it is believed that the higher the scales the more accurate the wavelet model is, experience from numerous case studies has shown that the dilation parameter need not to be chosen very high. As a rule of thumb, setting the scale parameter to values up to 2 or 3 can often work well.

Acknowledgements

The authors gratefully acknowledge that this work was supported by EPSRC(UK).

References

A. Antoniadis, I. Gijbels, G. Gregoire, “Model selection using wavelet decomposition and applications”, *Biometrika*, 84(4), pp. 751-763, 1997.

A. Bemporad, G. Ferrari-Trecate, and M. Morari, “Observability and Controllability of piecewise affine and hybrid systems”, *IEEE Trans. Auto. Control*, 45(10), pp. 1864-1876, 2000.

S. A. Billings and S. Chen, “The determination of multivariable nonlinear models for dynamic systems using neural networks”, In C.T. Leondes (Ed.), *Neural Network Systems Techniques and Applications*, San Diego: Academic Press, pp. 231-278, 1998.

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

S. A. Billings and W. S. F. Voon, "Piecewise linear identification of non-linear systems", *Int. J. Control*, 46(1), pp.215-235, 1987.

S. A. Billings and H. L. Wei, "A new class of wavelet networks for nonlinear system identification", *IEEE Trans. Neural Networks*, 16(4), pp. 862-874, 2005.

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

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

Z. H. Chen, "Fitting multivariate regression functions by interaction spline models", *J. R. Stat. Soc. Ser. B-Methodol.*, 55(2), pp. 473-491, 1993.

V. Cherkassky and F. Mulier, *Learning from Data*, New York: John Wiley & Sons, 1998.

J. H. Friedman, "Multivariate adaptive regression splines", *Ann. Stat.*, 19(1), pp. 1-67, 1991.

C. J. Harris, X. Hong, and Q. Gan, *Adaptive Modelling, Estimation and Fusion from Data*: *A Neurofuzzy Approach*, Berlin: Springer-Verlag, 2002.

I. J. Leontaritis and S. A. Billings, "Input-output parametric models for non-linear systems—part 1: deterministic non-linear systems", *Int J. Control*, 41(2), 303-328, 1985a.

I. J. Leontaritis and S. A. Billings, "Input-output parametric models for non-linear systems—part 2: Stochastic Non-Linear Systems", *Int J. Control*, 41(2), 329-344, 1985b.

L. Ljung, *System Identification : Theory for the User*, Englewood Cliffs : Prentice-Hall, 1987.

G. Li, C. Rosenthal, and H. Rabitz, "High dimensional model representation", *J. Physical Chemistry*, 105(33), pp. 7765-7777, 2001.

G. P. Liu, *Nonlinear Identification and Control: A Neural Network Approach*, Berlin: Springer-Verlag,
2001.

R. Murry-Smith and T. A. Johansen (ed.), Multiple model approaches to modelling and control, London : Taylor & Francis, 1997.

R. K. Pearson, *Discrete-Time Dynamic Models*, Oxford: Oxford University Press, 1999.

J. Rissanen, "A universal prior for integers and estimation by minimum description length", *Ann. Stat.*, 11, pp. 416-431, 1983.

N. Saito, "Simultaneous noise suppression and signal compression using a library of orthonormal bases and the minimum description length criterion", In *Wavelet in Geophysics*, Foufoula-Georgiou, E. and Kumar, P., Eds, New York: Academic, pp. 299-324, 1994.

E. D. Sontag, "Nonlinear regulation: the piecewise linear approaches", *IEEE Trans. Auto. Control*, 26(2), pp. 346-358, 1981.

T. Söderström and P. Stoica, *System Identification*, New York : Prentice Hall, 1989.

H. L. Wei and S. A. Billings, “A unified wavelet-based modelling framework for nonlinear system identification: the WANARX model structure”, *Int. J. Control*, 77(4), pp. 351-366, 2004.

H. L. Wei, S. A. Billings, and M. A. Balikhin, “Prediction of the *Dst* index using multiresolution wavelet models”, *J. Geophys. Res*, 109(A7), A07212, doi:10.1029/2003JA010332, 2004a.

H. L. Wei, S. A. Billings, and J. Liu, “Term and variable selection for nonlinear system identification”, *Int. J. Control*, 77(1), pp. 86-110, 2004b.

H. L. Wei, S. A. Billings, and M. A. Balikhin, “Wavelet based nonparametric NARX models for nonlinear input-output system identification”, Accepted by *Int. J. Syst. Sci.*, 2006.