Robust Neurofuzzy Rule Base Knowledge Extraction and Estimation using Subspace Decomposition Combined with Regularization and D-optimality

Xia Hong, *Senior Member, IEEE,* Chris J. Harris and Sheng Chen, *Senior Member, IEEE*

***Abstract*— A new robust neurofuzzy model construction algorithm has been introduced for the modelling of *a priori* unknown dynamical systems from observed finite data sets in the form of a set of fuzzy rules. Based on a Takagi and Sugeno (T-S) inference mechanism a one to one mapping between a fuzzy rule base and a model matrix feature subspace is established. This link enables rule based knowledge to be extracted from matrix subspace to enhance model transparency. In order to achieve maximized model robustness and sparsity, a new robust extended Gram-Schmidt method has been introduced via two effective and complementary approaches of regularization and D-optimality experimental design. Model rule bases are decomposed into orthogonal subspaces, so as to enhance model transparency with the capability of interpreting the derived rule base energy level. A locally regularized orthogonal least squares algorithm, combined with a D-optimality used for subspace based rule selection, has been extended for fuzzy rule regularization and subspace based information extraction. By using a weighting for the D-optimality cost function, the entire model construction procedure becomes automatic. Numerical examples are included to demonstrate the effectiveness of the proposed new algorithm.**

***Index Terms*— Neurofuzzy networks, orthogonal decomposition, subspace, regularization, optimal experimental design.**

I. Introduction

Associative memory networks (such as B-spline networks, RBF's, support vector machines (SVM)) have been extensively developed [1], [2], [3], [4]. Most conventional neural networks lead only to 'black box' model representation, yet a neurofuzzy network has an inherent model transparency that helps users to understand the system behaviours, oversee critical system operating regions, and/or extract physical laws or relationships that underpin the system. Based on the fuzzy rules inference and model representation of Takagi and Sugeno [5], a neurofuzzy model can be functionally expressed as an operating point dependent fuzzy model with a local linear description that lends itself directly to conventional estimation and control synthesis [1], [6], [7]. The model output is decomposed into a convex combination of the outputs of individual rules, and the basis function can be interpreted as a fuzzy membership function of individual rules. This property is critically desirable for problems requiring insight into the underlying phenomenology, i.e. internal system behavior interpretability and/or knowledge (rule) representation of the underlying process.

The problem of *the curse of dimensionality* [8] has been a main obstacle in non-linear modelling using associative memory networks or fuzzy logic. Networks or knowledge representations that suffer from the curse of dimensionality include all lattice based networks such as Fuzzy Logic (FL), Radial Basis Function (RBF), Karneva distributed memory maps, and all neurofuzzy networks (e.g. adaptive network based fuzzy inference system (ANFIS) [9], Takagi and Sugeno model [5], etc.). This problem also mitigates against model transparency for high dimensional systems since they generate massive rule sets, or require too many parameters, making it impossible for a human to comprehend the resultant rule set. Consequently the major purpose of neurofuzzy model construction algorithms is to select a parsimonious model structure that resolves the bias/variance dilemma (for finite training data), has a smooth prediction surface (e.g. parameter control via regularization), produces good generalization (for unseen data), and with an interpretable representation -often in the form of (fuzzy) rules. For general linear in the parameter systems, an orthogonal least squares (OLS) algorithm based on Gram-Schmidt orthogonal decomposition can be used to determine the models significant elements and associated parameter estimates, and the overall model structure [10]. Regularization techniques have been incorporated into the orthogonal least squares (OLS) algorithm to produce a regularized orthogonal least squares (ROLS) algorithm that reduces the variance of parameter estimates [11], [12]. To produce a model with good generalization capabilities, model selection criteria such as the Akaike information criterion (AIC) [13] are usually incorporated into the procedure to determinate the model construction process. Yet the use of AIC or other information based criteria, if used in forward regression, only affects the stopping point of the model selection, but does not penalize regressors that might cause poor model performance, e.g. too large parameter variance or ill-posedness of the regression matrix, if this is selected. This is due to the fact that AIC or other information based criteria are usually simplified measures derived as an approximation formula that is particularly sensitive to model complexity.

Manuscript received ?, 2002; revised May 16, 2003. This work was supported by the EPSRC, UK.

Xia Hong is with Cybernetic Intelligence Research Group,Department of Cybernetics University of Reading, Reading, RG6 6AY, UK. Chris J. Harris and Sheng Chen are the Department of Electronics and Computer Science University of Southampton, Southampton SO17 1BJ, UK

In order to achieve a model structure with improved model generalization, it is natural that a model generalization capability cost function should be used in the overall model searching process, rather than only being applied as a measure of model complexity. Optimum experimental designs have been used [14] to construct smooth network response surfaces based on the setting of the experimental variables under well controlled experimental conditions. In optimum design, model adequacy is evaluated by design criteria that are statistical measures of goodness of experimental designs by virtue of design efficiency and experimental effort. Quantitatively, model adequacy is measured as function of the eigenvalues of the design matrix. In recent studies [15], [16], the authors have outlined efficient learning algorithms, in which composite cost functions were introduced to optimize the model approximation ability using the forward orthogonal least squares (OLS) algorithm [10], and simultaneously determined model adequacy using an A-optimality design criterion (i.e. minimizes the variance of the parameter estimates), or a D-optimality criterion (i.e. optimizes the parameter efficiency and model robustness via the maximization of the determinant of the design matrix). It was shown that the resultant models can be improved based on A- or D-optimality. These algorithms lead automatically to an unbiased model parameter estimate with an overall robust and parsimonious model structure. Combining a locally regularized orthogonal least squares (LROLS) model selection [17] with D-optimality experimental design further enhances model robustness [18].

Due to the inherent transparency properties of a neurofuzzy network, a parsimonious model construction approach should lead also to a logical rule extraction process that increases model transparency, as simpler models inherently involve fewer rules which are in turn easier to interpret. One drawback of most current neurofuzzy learning algorithms is that learning is based upon a set of one-dimensional regressors, or basis functions (such as B-splines, Gaussians, etc), but not upon a set of fuzzy rules (usually in the form of multi-dimensional input variables), resulting in opaque models during the learning process. Since modelling is inevitably iterative it can be greatly enhanced if the modeller can interpret or interrogate the derived rule base during learning itself, allowing him/her to terminate the process when his/her objectives are achieved. There are valuable recent developments on rule based learning and model construction, including a linear approximation approach combined with uncertainty modelling [19], various fuzzy similarity measures combined with genetic algorithms [20], [21]. Recently the authors have introduced a new neurofuzzy model construction and parameter estimation algorithm from observed finite data sets, based on a Takagi and Sugeno (T-S) inference mechanism and a new extended Gram-Schmidt orthogonal decomposition algorithm, for the modelling of *a priori* unknown dynamical systems in the form of a set of fuzzy rules [22], which, based on a Takagi and Sugeno (T-S) inference mechanism, establishes a one to one mapping between a fuzzy rule base and a model matrix feature subspace.

In this paper, a new neurofuzzy model construction and parameter estimation algorithm has been introduced. Based on a Takagi and Sugeno (T-S) inference mechanism a one to one mapping between a fuzzy rule base and a model matrix feature subspace is established [22]. This link enables rule based knowledge to be extracted from matrix subspace to enhance model transparency. In order to achieve maximized model robustness and sparsity, a new robust extended Gram-Schmidt algorithm has been introduced via two effective and complementary approaches of regularization and D-optimality experimental design. This new algorithm decomposes the model rule bases via an orthogonal subspace decomposition approach, so as to enhance model transparency with the capability of interpreting the derived rule base energy level. A locally regularized orthogonal least squares algorithm tailored for rule regularization has been combined with a D-optimality for subspace selection. By using a weighting for the D-optimality cost function, the entire model construction procedure becomes automatic. The proposed algorithm enhances the previous algorithm [22] via the combined LOLS and D-optimality for robust rule selection, and is based on the extension of the combined LOLS and D-optimality algorithm [18] from conventional regressor regression to orthogonal subspace regression.

This paper is organized as follows. Section 2 introduces a general class of neurofuzzy systems as a modelling approach. Section 3 introduces the proposed new algorithm, with analysis into the associated model transparency, robustness enhancement via D-optimality and rule based regularization. Numerical examples are provided in Section 4 to illustrate the effectiveness of the approach and Section 5 is devoted to conclusions.

II. A Neurofuzzy Modelling Approach

This section briefly presents a general class of neurofuzzy systems as a nonlinear data modelling approach within a coherent framework of both mathematical representation for learning and linguistic logic rule representation for model transparency. Given a finite data set $D_N = \{\mathbf{x}(t), y(t)\}_{t=1}^{N}$ of observed input/output data pairs, consider the identification of a general nonlinear system that generates this data

$$y(t) = f(\mathbf{x}(t), \boldsymbol{\Theta}) + e(t), \tag{1}$$

where

$$\mathbf{x}(t) = [x_1, x_2, \cdots, x_n]^T \in \mathcal{X} \in \Re^n \tag{2}$$

is an observed system input vector, $f(\bullet)$ is *a priori* unknown. The observation noise $e(t)$ is assumed uncorrelated with variance σ^2. $\boldsymbol{\Theta}$ is an unknown parameter vector associated with an appropriate but yet to be determined model structure.

Model (1) can be simplified by decomposing it into a set of K local models $f_i(\mathbf{x}^{(i)}(t), \boldsymbol{\Theta}_i)$, $i = 1, \cdots, K$, where K is to be determined, each of which operates on a local region depending on the sub-measurement vector $\mathbf{x}^{(i)} \in \Re^{n_i}$, a subset of the input vector $\mathbf{x}$, i.e. $\mathbf{x}^{(i)} \in \mathcal{X}_i \in \Re^{n_i}$, $(n_i < n)$, $\mathcal{X}_1 \cup \cdots \cup \mathcal{X}_K = \mathcal{X}$. Each of the local models $f_i(\mathbf{x}^{(i)}(t), \boldsymbol{\Theta}_i)$ can be represented by a set of linguistic rules

$$\begin{aligned} \text{IF} \quad & \mathbf{x}^{(i)} \text{ is } A^{(i)} \\ \text{THEN} \quad & y(t) = f_i(\mathbf{x}^{(i)}(t), \boldsymbol{\Theta}_i), \end{aligned} \tag{3}$$

where the fuzzy set $A^{(i)} = [A_1^{(i)}, \cdots, A_{n_i}^{(i)}]^T$ denotes a fuzzy set in the n_i-dimensional input space, $\Re^{n_i}$ and is given as an array of linguistic values, based on a predetermined input spaces partition into fuzzy sets via some prior system knowledge of the operating range of the data set. Usually if $\mathbf{x}^{(j)} = \mathbf{x}^{(k)}$, for $j \neq k$, then $A^{(j)} \cap A^{(k)} = \emptyset$, where $\emptyset$ denotes empty set. $\cup_{i=1}^{K} A^{(i)}$ defines a complete fuzzy partition of the input space $\mathcal{X}$. For an appropriate input space decomposition, the local models can have essentially local linear behavior. In this case, using the well known Takagi-Sugeno fuzzy inference mechanism [5], the output of system (1) can be represented by

$$f(\mathbf{x}(t), \mathbf{\Theta}) = \sum_{i=1}^{K} N_i(\mathbf{x}^{(i)}(t)) f_i(\mathbf{x}^{(i)}(t), \mathbf{\Theta}_i), \tag{4}$$

where $f_i(\mathbf{x}^{(i)}(t), \mathbf{\Theta}_i)$ is a linear function of $\mathbf{x}^{(i)}$, given by

$$f_i(\mathbf{x}^{(i)}(t), \mathbf{\Theta}_i) = \mathbf{x}^{(i)}(t)^T \mathbf{\Theta}_i \tag{5}$$

and $\mathbf{\Theta}_i \in \Re^{n_i}$ denotes parameter vector of the ith fuzzy rule or local model. $N_i(\mathbf{x}^{(i)})$ is a fuzzy membership function of the rule (3), subject to a unity of support condition: $0 \leq N_i(\mathbf{x}^{(i)}) \leq 1$, $\sum_{i=1}^{K} N_i(\mathbf{x}^{(i)}) = 1$. Each of the linguistic rules (3) can be evaluated via the known fuzzy membership function $N_i(\mathbf{x}^{(i)}(t))$.

Consider a neurofuzzy network using B-spline functions [23] as membership functions. A general one-dimensional B-spline model $f'(x)$ can be formed as a linear combination of L B-spline basis functions, $B_m^j(x)$, as

$$f'(x) = \sum_{j=1}^{L} \theta_j B_m^j(x) \tag{6}$$

The coefficients θ_j's represent the set of adjustable parameters associated with the set of basis functions. $B_m^j(x)$'s, which are polynomials of a given degree m and are uniquely defined by an ordered sequence of real values denoted as a knot vector $\tau = \{\tau_1, \tau_2, \cdots, \tau_{L+m+1}\}$. The knot sequence forms a partitioning of the input domain into $(L + m)$ disjoint intervals. The basis functions set can be defined by recursive equation [23]

$$\begin{aligned} B_m^j(x) &= \frac{x - \tau_j}{\tau_{j+m} - \tau_j} B_{m-1}^j(x) \\ &\quad + \frac{\tau_{j+m+1} - x}{\tau_{j+m+1} - \tau_{j+1}} B_{m-1}^j(x) \end{aligned} \tag{7}$$

with

$$B_0^j(x) = \begin{cases} 1 & \tau_j \leq x < \tau_{j+1} \\ 0 & otherwise \end{cases}$$

Multidimensional B-spline basis functions are formed by a direct multiplication of univariate basis functions via

$$N_i(\mathbf{x}^{(i)}) = \prod_{j=1}^{n_i} B_{j,m}^{k_j}(x_j^{(i)}) \tag{8}$$

for $i = 1, \cdots, M$, where $M = \prod_{i=1}^{n} L_i$, $\mathbf{x}^{(i)} = [x_1^{(i)}, x_2^{(i)}, \cdots, x_{n_i}^{(i)}]^T \in \Re^{n_i}$. $k_j = 1, 2, \cdots, L_j$, L_j is the number of B-spline basis functions defined in $x_j^{(i)}$, the jth component of $\mathbf{x}^{(i)}$.

Note that for a complete model base, the number of rules $M \gg K$ increases exponentially as the input dimension increases, (which is commonly known as the curse of dimensionality). To alleviate this disadvantage, input dimension or variable reduction can be used. Notably an ANOVA (Analysis of Variance) representation of multivariable functions uses lower dimensional tensor products of models inputs, e.g. in many practical applications, the number of multiplication terms maybe limited to as low as 3, yet maintaining sufficient modelling capability [1]. For practical applications, not only is the ANOVA approach effective in overcoming the curse of dimensionality, it has additional advantage of model transparency because a lower input dimension than 3 can be visualized and interpreted [24].

Substitute (5) & (4) into (1)

$$\begin{aligned} y(t) &= \sum_{i=1}^{K} \phi_i(\mathbf{x}^{(i)}(t))^T \mathbf{\Theta}_i + e(t) \\ &= \phi(\mathbf{x}(t))^T \mathbf{\Theta} + e(t), \end{aligned} \tag{9}$$

where $\phi_i(\mathbf{x}(t)) = [\phi_{i1}(t), \cdots, \phi_{in_i}(t)]^T = N_i(\mathbf{x}^{(i)}(t))\ \mathbf{x}^{(i)} \in \Re^{n_i}$. $\phi(\mathbf{x}(t)) = [\phi_1(\mathbf{x}^{n_1}(t))^T, \cdots, \phi_K(\mathbf{x}^{n_K}(t))^T]^T \in \Re^p$. $\mathbf{\Theta} = [\mathbf{\Theta}_1^T, \cdots, \mathbf{\Theta}_K^T]^T \in \Re^p$, where $p = \sum_{i=1}^{K} n_i$.

For the finite data set $D_N = \{\mathbf{x}(t), y(t)\}_{t=1}^{N}$, (9) can be written in a matrix form as

$$\begin{aligned} \mathbf{y} &= \sum_{i=1}^{K} \Phi^{(i)} \mathbf{\Theta}^{(i)} + \mathbf{e} \\ &= \mathbf{\Phi\Theta} + \mathbf{e} \end{aligned} \tag{10}$$

where $\mathbf{y} = [y(1), y(2), \cdots, y(N)]^T \in \Re^N$ is the output vector, $\Phi^{(i)} = [\phi_i(\mathbf{x}(1)), \cdots, \phi_i(\mathbf{x}(N))]^T \in \Re^{N \times n_i}$ is the regression matrix associated with the ith fuzzy rule, $\mathbf{e} = [e(1), \cdots, e(N)]^T \in \Re^N$ is the model residual vector. $\mathbf{\Phi} = [\Phi^{(1)}, \cdots, \Phi^{(K)}] \in \Re^{N \times p}$ is the full regression matrix.

An effective way of overcoming the curse of dimensionality is to start with a moderate sized rule base according to the actual data distribution. In this paper, the selection of K local models as an initial model base is based on model identifiability via an A-optimality design criterion [14] with the advantage of enhanced model transparency to quantify and interpret fuzzy rules and their identifiability.

III. Rule Based Model Construction and Learning Algorithms

A. Rule based learning and initial model base construction

Rule based knowledge, i.e. information associated with a fuzzy rule, is highly appropriate for users to understand a derived data based model. Most current learning algorithms in neurofuzzy model are based on an ordinary p-dimensional linear in the parameter model. Model transparency during learning cannot be automatically achieved unless these regressors have a clear physical interpretation, or are directly associated with physical variables. Alternatively, a neurofuzzy network is inherently transparent for rule based model construction. In (10), each of $\Phi^{(i)}$ is constructed based on a unique fuzzy membership function $N_i(\cdot)$, providing a link between a fuzzy rule base and a matrix feature subspace

spanned by $\Phi^{(i)}$. Rule based knowledge can be easily extracted by exploring this link.

Definition 1: Basis of a subspace: If n_i vectors $\phi_j^{(i)} \in \Re^N$, $j = 1, 2, \cdots, n_i$, satisfy the nonsingular condition that $\mathbf{\Phi}^{(i)} = [\phi_1^{(i)}, \cdots, \phi_{n_i}^{(i)}] \in \Re^{N \times n_i}$ has a full rank of n_i, they span a n_i-dimensional subspace $S^{(i)}$, then $\mathbf{\Phi}^{(i)}$ is the basis of the subspace $S^{(i)}$.

Definition 2: Fuzzy rule subspace: Suppose the $\Phi^{(i)}$ is nonsingular, clearly $\Phi^{(i)}$ is the basis of a n_i-dimensional subspace $S^{(i)}$, which is a functional representation of the fuzzy rule (3) by using Takagi-Sugeno fuzzy inference mechanism with a unique label $N_i(\cdot)$. $S^{(i)}$ is defined as a fuzzy rule subspace of the ith fuzzy rule.

$\Phi^{(i)}$, the sub-matrix associated with the ith rule, can be expanded as

$$\Phi^{(i)} = \mathbf{N}^{(i)} X^{(i)} \tag{11}$$

where $\mathbf{N}^{(i)} = diag\{N_i(1), \cdots, N_i(N)\} \in \Re^{N \times N}$, $X^{(i)} = [\mathbf{x}^{(i)}(1), \mathbf{x}^{(i)}(2), \cdots, \mathbf{x}^{(i)}(N)]^T \in \Re^{N \times n_i}$. (11) shows that each rule base is simply constructed by a weighting matrix multiplied to the regression matrix of original input variables. The weighting matrix $\mathbf{N}^{(i)}$ can be regarded as a data based spatial prefiltering over the input region. Without loss of generality, it is assumed that $X^{(i)}$ is nonsingular, and $N > n_i$, as $rank(X^{(i)}) = n_i$. As

$$\text{rank}(\Phi^{(i)}) = \min[\text{rank}(\mathbf{N}^{(i)}), \text{rank}(X^{(i)})] \tag{12}$$

For $\Phi^{(i)}$ to be nonsingular, then $\text{rank}(\mathbf{N}^{(i)}) > n_i$, this means that for the input region denoted by $N_i(\cdot)$, its basis function needs to be excited by at least n_i data points.

The A-optimality design criteria for the weighting matrix $\mathbf{N}^{(i)}$ which is given by [14], [22]

$$J_A(\mathbf{N}^{(i)}) = \frac{1}{N} \sum_{t=1}^{N} N_i(t), \tag{13}$$

provides an indication for each fuzzy rule on its identifiability and hence a metric for selecting appropriate model rules. The derived model rules can then be rearranged in descending order of identifiability, followed by utilizing only the first K experts with identifiability to construct a model rule base set.

B. Orthogonal subspace decomposition and regularization in orthogonal subspace

For ease of exposition, we initially introduce some notations and definitions that are used in the development of the new extended Gram-Schmidt orthogonal decomposition algorithm.

Definition 3: Orthogonal subspaces: For a p-dimensional matrix space $S \in \Re^{N \times p}$, two of its subspaces $\mathcal{W}^{(i)} \in \Re^{N \times n_i} \subset S$ and $\mathcal{W}^{(j)} \in \Re^{N \times n_j} \subset S$, ($n_i < p$, $n_j < p$) are orthogonal if and only if any two vectors $\mathbf{w}^{(i)}$ and $\mathbf{w}^{(j)}$ that are located in the two subspaces respectively, i.e. $\mathbf{w}^{(i)} \in \mathcal{W}^{(i)}$ and $\mathbf{w}^{(j)} \in \mathcal{W}^{(j)}$, are orthogonal, that is, $[\mathbf{w}^{(i)}]^T \mathbf{w}^{(j)} = 0$, for $i \neq j$.

The p-dimensional space S, ($p = \sum_{i=1}^{K} n_i$), can be decomposed by K orthogonal subspaces $\mathcal{W}^{(i)}$, $i = 1, \cdots, K$, given by[25], [26]

$$\mathcal{W}^{(1)} \oplus \cdots \oplus \mathcal{W}^{(K)} = S \in \Re^{p \times N}, \tag{14}$$

where $\oplus$ denotes sum of orthogonal sets. From Definition 1, if there are any linear uncorrelated n_i vectors located in $\mathcal{W}^{(i)}$, denoted as $\mathbf{w}_i^{(i)} \subset \mathcal{W}^{(i)}$, $i = 1, \cdots, n_i$, then the matrix $\mathbf{W}^{(i)} = [\mathbf{w}_1^{(i)}, \cdots, \mathbf{w}_{n_i}^{(i)}]$, forms a basis of $\mathcal{W}^{(i)}$. Note that these n_i vectors need not to be mutually orthogonal, i.e. $[\mathbf{W}^{(i)}]^T \mathbf{W}^{(i)} = \mathbf{D}^{(i)} \in \Re^{n_i \times n_i}$, where $\mathbf{D}^{(i)}$ is not required to be diagonal.

Clearly if two matrix subspaces $\mathcal{W}^{(i)}$, $\mathcal{W}^{(j)}$ have the basis of full rank matrices $\mathbf{W}^{(i)} \in \Re^{N \times n_i}$, $\mathbf{W}^{(j)} \in \Re^{N \times n_j}$, then they are orthogonal if and only if

$$[\mathbf{W}^{(i)}]^T \mathbf{W}^{(j)} = \mathbf{0}_{n_i \times n_j} \tag{15}$$

where $\mathbf{0}_{n_i \times n_j} \in \Re^{n_i \times n_j}$ is a zero matrix.

Definition 4: Vector decomposition to subspace basis: If K orthogonal subspaces $\mathcal{W}^{(i)}$, $i = 1, \cdots, K$, are defined by a series of K matrices $\mathbf{W}^{(i)}$, $i = 1, \cdots, K$ as subspace basis based on Definition 3, then an arbitrary vector $\hat{\mathbf{y}} \in \Re^N \in S$ can be uniquely decomposed as

$$\begin{aligned} \hat{\mathbf{y}} &= \sum_{i=1}^{K} \sum_{j=1}^{n_i} c_{i,j} \mathbf{w}_j^{(i)} \\ &= \sum_{i=1}^{K} \mathbf{W}^{(i)} \mathbf{c}_i \end{aligned} \tag{16}$$

where $c_{i,j}$'s are combination coefficients. $\mathbf{c}_i = [c_{i,1}, \cdots, c_{i,n_i}]^T \in \Re^{n_i}$.

As the result of the orthogonality of $[\mathbf{w}^{(i)}]^T \mathbf{w}^{(j)} = 0$, (for $i \neq j$), from (16),

$$\hat{\mathbf{y}}^T \hat{\mathbf{y}} = \sum_{i=1}^{K} \mathbf{c}_i^T \mathbf{D}^{(i)} \mathbf{c}_i \tag{17}$$

Clearly the variance of the vector $\hat{\mathbf{y}}$ projected into each subspace can be computed as $\mathbf{c}_i^T \mathbf{D}^{(i)} \mathbf{c}_i$, for $i = 1, \cdots, K$.

Consider the nonlinear system (1) given as a vector form by (10). By introducing an orthogonal subspace decomposition $\mathbf{\Phi} = \mathbf{WA}$, (10) can be written as

$$\begin{aligned} \mathbf{y} &= \mathbf{W}\mathbf{c} + \mathbf{e} \\ &= \sum_{i=1}^{K} \mathbf{W}^{(i)} \mathbf{c}_i + \mathbf{e} \end{aligned} \tag{18}$$

where $\mathbf{W} = [\mathbf{W}^{(1)}, \cdots, \mathbf{W}^{(K)}]$ spans a p-dimensional space S with $\mathbf{W}^{(i)}$, $i = 1, \cdots, K$ spanning its subspaces $\mathcal{W}^{(i)}$, as defined via Definition 3. The auxiliary parameter vector $\mathbf{c} = \mathbf{A\Theta} = [\mathbf{c}_1^T, \cdots, \mathbf{c}_K^T]^T \in \Re^p$, where $\mathbf{A}$ is a block upper

triangular matrix

$$\mathbf{A} = \begin{bmatrix} A_{1,1} & A_{1,2} & \cdots & A_{1,K} \\ 0 & A_{2,2} & \cdots & A_{2,K} \\ & \cdots & \cdots & \\ 0 & \cdots & A_{i,j} & \cdots \\ & \cdots & \cdots & \\ 0 & \cdots & \cdots & A_{K,K} \end{bmatrix} \in \Re^{p \times p} \tag{19}$$

in which $A_{i,j} \in \Re^{n_i \times n_j}$. $A_{i,i} = \mathbf{I}_{n_i \times n_i}$, a unit matrix $\in \Re^{n_i \times n_i}$.

Definition 5: The extended Gram-Schmidt orthogonal decomposition algorithm [22]: An orthogonal subspace decomposition for model (18) can be realized based on an extended Gram-Schmidt orthogonal decomposition algorithm as follows, Set $\mathbf{W}^{(1)} = \Phi^{(1)}$, $A_{1,1} = \mathbf{I}_{n_1 \times n_1}$, and, for $j = 2, \cdots, K$, set $A_{j,j} = \mathbf{I}_{n_j \times n_j}$,

$$\mathbf{W}^{(j)} = \Phi^{(j)} - \sum_{i=1}^{j-1} \mathbf{W}^{(i)} * A_{i,j}, \tag{20}$$

where

$$A_{i,j} = \left[\mathbf{D}^{(i)}\right]^{-1} [\mathbf{W}^{(i)}]^T \Phi^{(j)} \in \Re^{n_i \times n_j} \tag{21}$$

for $i = 1, \cdots, j-1$.

Definition 6: Locally regularized least squares cost function in orthogonal subspaces: The orthogonal subspace based regularized least squares uses the following error criterion:

$$J_R(\mathbf{c}, \boldsymbol{\lambda}) = \mathbf{e}^T\mathbf{e} + \mathbf{c}^T\boldsymbol{\Lambda}\mathbf{c} \tag{22}$$

where $\boldsymbol{\lambda} = [\lambda_1, \cdots, \lambda_M]^T$, $\lambda_k > 0$, $k = 1, 2, \cdots, M$ are regularization parameters, and the diagonal matrix $\boldsymbol{\Lambda} = diag\{\lambda_1 \mathbf{I}_{n_1 \times n_1}, \lambda_2 \mathbf{I}_{n_2 \times n_2}, \cdots, \lambda_M \mathbf{I}_{n_M \times n_M}\}$, $\mathbf{I}$ is a unit matrix. The regularized least squares estimates of $\mathbf{c}$, is given by [27]

$$\mathbf{c} = (\mathbf{W}^T\mathbf{W} + \boldsymbol{\Lambda})^{-1}\mathbf{W}^T\mathbf{y} \tag{23}$$

An appropriate choice of $\boldsymbol{\lambda}$ can smooth parameter estimates (noise rejection), and $\boldsymbol{\lambda}$ can be optimized by using a separate procedure, such as Bayesian hyper-parameter optimization [18], or a genetic algorithm. In this paper, it is assumed that an appropriate $\boldsymbol{\lambda}$ is predetermined to simplify the procedure. The regularized least squares solution of (18) is given by

$$\mathbf{c}_i = [\mathbf{D}^{(i)} + \lambda_i \mathbf{I}]^{-1}[\mathbf{W}^{(i)}]^T\mathbf{y} \tag{24}$$

which follows from the fact that $\mathbf{W}^{(i)}$, $i = 1, \cdots, K$ are mutually orthogonal subspaces basis, and $\mathbf{D}^{(i)} = [\mathbf{W}^{(i)}]^T\mathbf{W}^{(i)}$. From (16), if the system output vector $\mathbf{y}$ is decomposed as a term $\hat{\mathbf{y}}$ by projecting onto orthogonal subspaces $\mathbf{W}^{(i)}$, $i = 1, \cdots, K$, and an uncorrelated term $\mathbf{e}(t)$ that is unexplained by the model, such that the projection onto each subspace basis (or a percentage energy contribution of these subspaces towards the construction of $\mathbf{y}$) can be readily calculated via

$$[err]_i = \frac{\mathbf{c}_i^T\mathbf{D}^{(i)}\mathbf{c}_i}{\mathbf{y}^T\mathbf{y}} \tag{25}$$

The output variance projected onto each subspace can be interpreted as the contribution of each fuzzy rule in the fuzzy system, subject to the existence of previous fuzzy rules. To include the most significant subspace basis with the largest $[err]_i$ as a forward regression procedure is a direct extension of conventional forward OLS algorithm [10]. The output variance projected into each subspace can be interpreted as the output energy contribution explained by a new rule demonstrating the significance of the new rule towards the model. At each regression step, a new orthogonal subspace basis is formed by using a new fuzzy rule and the existing fuzzy rules in the model, with the rule basis with the largest $[err]_i$ to be included in the final model until

$$1 - \sum_{i=1}^{n_f} [err]_i < \rho \tag{26}$$

satisfies for an error tolerance ρ to construct a model with $n_f < K$ rules. The parameter vectors $\boldsymbol{\Theta}_i$, $i = 1, \cdots, n_f$ can be computed by the following back substitution procedure: Set $\boldsymbol{\Theta}_{n_f} = \mathbf{c}_{n_f}$, and, for $i = n_f - 1, \cdots, 1$

$$\boldsymbol{\Theta}_i = \mathbf{c}_i - \sum_{j=i+1}^{n_f} A_{i,j} * \boldsymbol{\Theta}_j \tag{27}$$

The concept of orthogonal subspace decomposition based on fuzzy rule bases is illustrated in Figure 1. This figure illustrates (20) that forms the orthogonal bases. Because of the one to one mapping of a fuzzy rule to a matrix subspace, a series of orthogonal subspace basis are formed by using fuzzy rule subspace basis $\Phi^{(i)}$ in a forward regression manner, such that, $\{\mathcal{W}^{(1)} \oplus \mathcal{W}^{(2)} \oplus \cdots \mathcal{W}^{(i)}\} = \{S^{(1)} \cup S^{(2)} \cup \cdots S^{(i)}\}$, $\forall i$, whilst maximizing the output variance of the model at each regression step i. Note that the well known orthogonal schemes such as the classical Gram-Schmidt method construct orthogonal vectors as basis based on regression vectors (one dimensional), but the new algorithm extends the classical Gram-Schmidt orthogonal decomposition scheme to the orthogonalization of subspace bases (multidimensional). The extended Gram-Schmidt orthogonal decomposition algorithm is not only an extension from classical Gram-Schmidt orthogonal axis decomposition to orthogonal subspace decomposition, but also as an extension from basis function regression to matrix subspace regression, introducing a significant advantage of model transparency to interpret fuzzy rule energy level.

C. New extended Gram-Schmidt orthogonal decomposition algorithm with regularization and D-optimality in orthogonal subspaces

The above discussion has been largely introduced in [22], except that in [22], the $[err]_i$ was used for subset selection without parameter regularization ($\boldsymbol{\lambda} = \mathbf{0}$). Regularization can be used as an effective resort to overcome overfitting to noise. Note that the use of $[err]_i$ aims to optimize the model in terms of approximation capability, but not in terms of model robustness. In addition to parameter regularization, composite cost function such as least squares plus a penalty term based D-optimality experimental design criterion can be

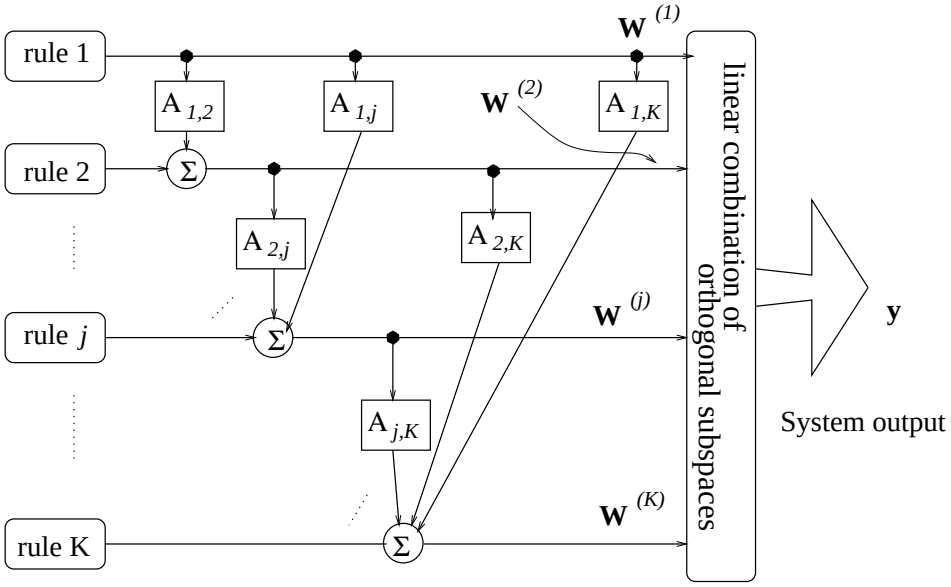

Fig. 1. Orthogonal subspace decomposition based on fuzzy rule bases.

used [16]. To enhance rule model robustness, the proposed algorithm combines the two separate previous works, the subspace based rule based model construction [22] and the combined LOLS and D-optimality algorithm [18] for robust rule based model construction. The combined LOLS and D-optimality algorithm [18] was not previously introduced as a rule based learning algorithm, hence some extensions to orthogonal subspace decomposition domain are necessary, as introduced in the following.

The concept of parameter regularization may be incorporated into a forward orthogonal least squares algorithm as a locally regularized orthogonal least square estimator for subspace selection by defining a regularized error reduction ratio due to the submatrix $\mathbf{W}_i$ as follows.

After some simplification, it can be shown that the criterion (22) can be expressed as

$$\mathbf{e}^T\mathbf{e} + \mathbf{c}^T\mathbf{\Lambda}\mathbf{c} = \mathbf{y}^T\mathbf{y} - \sum_{i=1}^{K} \mathbf{c}_i^T(\mathbf{D}^{(i)} + \lambda_i\mathbf{I})\mathbf{c}_i \tag{28}$$

where $\mathbf{D}^{(i)} = [\mathbf{W}^{(i)}]^T\mathbf{W}^{(i)}$. Normalizing (28) by $\mathbf{y}^T\mathbf{y}$ yields

$$\frac{\mathbf{e}^T\mathbf{e} + \mathbf{c}^T\mathbf{\Lambda}\mathbf{c}}{\mathbf{y}^T\mathbf{y}} = 1 - \sum_{i=1}^{K} \frac{\mathbf{c}_i^T(\mathbf{D}^{(i)} + \lambda_i\mathbf{I})\mathbf{c}_i}{\mathbf{y}^T\mathbf{y}}. \tag{29}$$

The regularized error reduction ratio $[err]_i$ due to the submatrix $\mathbf{W}^{(i)}$

$$[err]_i = \frac{\mathbf{c}_i^T(\mathbf{D}^{(i)} + \lambda_i\mathbf{I})\mathbf{c}_i}{\mathbf{y}^T\mathbf{y}}. \tag{30}$$

Definition 7: D-optimality experimental design cost function in orthogonal subspaces: In experimental design, the data covariance matrix $(\mathbf{\Phi}^T\mathbf{\Phi})$ is called the design matrix. The D-optimality design criterion maximizes the determinant of the design matrix for the constructed model. Consider a model with orthogonal subspaces with design matrix as $(\mathbf{W}^T\mathbf{W})$, and a subset of these subspaces are selected in order to construct a n_f-subspace $(n_f \ll K)$ model that maximizes the D-optimality $\det(\mathbf{W}_{n_f}^T\mathbf{W}_{n_f})$, where $\mathbf{W}_{n_f}$ is a column subset of $\mathbf{W}$ representing a constructed subset model with n_f sub-matrices selected from $\mathbf{W}$ (consisting of K sub-matrices). It is straightforward to verify that the maximization of $\det(\mathbf{W}_{n_f}^T\mathbf{W}_{n_f})$ is equivalent to the minimization of $J_D = -\log(\det(\mathbf{W}_{n_f}^T\mathbf{W}_{n_f}))$ [22]. Clearly

$$\begin{aligned} J_D &= -\log(\det(\mathbf{W}_{n_f}^T\mathbf{W}_{n_f})) \\ &= -\log\left[\prod_{i=1}^{n_f}\det(\mathbf{D}^{(i)})\right] \\ &= -\sum_{i=1}^{n_f}\log\left[\det(\mathbf{D}^{(i)})\right] \end{aligned} \tag{31}$$

It can be easily verify that the maximization of $\det(\mathbf{W}_{n_f}^T\mathbf{W}_{n_f})$ is identical to the maximization of $\det(\mathbf{\Phi}_{n_f}^T\mathbf{\Phi}_{n_f})$, where $\mathbf{\Phi}_{n_f}$ is a column subset of $(\mathbf{\Phi}$ representing a constructed subset model with n_f sub-matrices selected from $\mathbf{\Phi}$ (consisting of K sub-matrices) [22].

Definition 8: Combined Locally Regularized cost function and D-optimality in orthogonal subspaces: The combined LROLS and D-optimality algorithm based on orthogonal subspace decomposition is based on the combined criterion

$$J_c(\mathbf{c}, \boldsymbol{\lambda}, \beta) = J_R(\mathbf{c}, \boldsymbol{\lambda}) + \beta J_D \tag{32}$$

for model selection, where β is a fixed small positive weighting for the D-optimality cost. Equivalently a combined error reduction ratio defined as

$$[cerr]_i = \frac{\mathbf{c}_i^T(\mathbf{D}^{(i)} + \lambda_i\mathbf{I})\mathbf{c}_i + \beta\log\left[\det(\mathbf{D}^{(i)})\right]}{\mathbf{y}^T\mathbf{y}}. \tag{33}$$

is used for model selection, and the selection is terminated with a n_f-subspace model when

$$[cerr]_i \leq 0 \quad \text{for} \quad n_f + 1 \leq i \leq K \tag{34}$$

The introduction of D-optimality enhances model robustness and simplify the model selection procedure [18]. Given a proper $\boldsymbol{\lambda}$, the new extended Gram-Schmidt orthogonal subspace decomposition algorithm with regularization and D-optimality for rule based model construction is given in Appendix I.

IV. Numerical examples

Example 1: We start with a simple illustrative mapping example. Consider a nonlinear functional approximation of:

$$\begin{cases} z(x) = x^2 \exp(-3x), \\ y(x) = z(x) + N(0, 0.01^2). \end{cases}$$

500 data pairs $\{x, y\}$ are generated where the system input x is generated as a uniformly distributed random number ranged in [0,1]. Define a knot vector $[-0.2, 0, 0.2, 0.4, 0.6, 0.8, 1, 1.2]$, and use a piecewise linear B-spline fuzzy membership function to build a one-dimensional model, resulting $M = 6$ basis functions. These basis functions, as shown in Figure 2, corresponding to 6 fuzzy rules: (1) If (x = 0) (very small); (2) IF (x = 0.2) (small); (3) IF (x = 0.4) (medium-small); (4) IF (x = 0.6) (medium-large); (5) IF (x = 0.8) (large), and (6) If (x = 1) (very large).

By using the fuzzy model (4) for the approximation of $z(x)$, the neurofuzzy model is simply given as

$$\hat{z}(t) = \sum_{j=1}^{6} N_j(x(t))x(t)\theta_j, \tag{35}$$

where t denotes the data label, with each of the fuzzy rule $\Phi^{(j)} = N_j(x(t))x(t)$ spanning a one dimensional space, i.e. $n_j = 1$, $\forall j$. The identifiability of these fuzzy rules are computed based on (13) and are listed in Table I. Because this example only involves a scalar input variable, the extended Gram-Schmidt orthogonal decomposition algorithm reduces to the conventional OLS algorithm, with each rule subspace being spanned by a one-dimensional rule basis. The proposed algorithm produces rule based information of percentage energy increment (or the model error reduction ratio) by the selected rule to the model, as shown in Table II (in the order of selected rules), shown in two cases of with or without parameter regularization. Each rule contribution in reducing model error (or increasing the model energy level) provides model transparency for the fuzzy rules interpretability. To verify the model's approximation and robustness, Table III lists the mean squares error (MSE) of model in target to the noisy observations ($y(x)$) and the true function ($z(x)$), respectively. For this example, the modelling results are insensitive to a wide range of the parameter β associated with D-optimality ($\beta = 0 \sim 1 \times 10^{-6}$). However for ($\beta = 10^{-6}$), the model selection process automatically terminates at a 5 rule model (rule 1 is excluded). This insensitivity means that varying β within a certain range will all terminates the modelling within a suitable structural range. The modelling results of a model using 5 rules with $\lambda = 0.2$ is plotted in Figure 3. This example demonstrates that the proposed method has good approximation and some robustness improvement. Clearly the proposed modelling approach is additionally advantageous via its significant model transparency during the modelling process.

Example 2: Nonlinear 2-D surface modelling. The Matlab logo was generated by the first eigenfunction of the L-shaped membrane. A 51×51 meshed data set is generated by using Matlab commands

$$\begin{aligned} x &= linspace(0, 1, 51); \\ y &= linspace(0, 1, 51); \\ [X, Y] &= meshgrid(x, y); \\ Z &= membrane(1, 25); \end{aligned} \tag{36}$$

such that output Z is defined over an unit square input region $[0, 1]^2$. The data set $z(x, y)$, shown in Figure 5.(a), is used to model the target function (the first eigenfunction of the L-shaped membrane function).

For both x, y, define a knot vector $[-0.4, -0.2, 0, 0.25, 0.5, 0.75, 1, 1.2, 1.4]$, and use a piecewise quadratic B-spline fuzzy membership function to build a one-dimensional model, resulting $M = 6$ basis functions. These basis functions, as shown in Figure 4, correspond to 6 fuzzy rules: (1) If (*x or y*) is (very small) (VS); (2) IF (*x or y*) is (small)(S); (3) IF (*x or y*) is (medium-small)(MS); (4) IF (*x or y*) is (medium-large)(ML); (5) IF (*x or y*) is (large)(L), and (6) If (*x or y*) is (very large)(VL).

The univariate and bivariate membership functions (interaction between univariate membership function x and y via tensor product) are used as model set and shown in Table IV, in which, the identifiability of fuzzy rules are listed based on (13). From Table IV, it is seen that all the rules have been uniformly excited. There are 48 rules.

By using the fuzzy model (4) for the modelling of $Z(x, y)$, the neurofuzzy model is simply given as

$$\hat{Z}(t) = \sum_{j=1}^{48} N_j(\mathbf{x}(t))\mathbf{x}^T\boldsymbol{\Theta}_j, \tag{37}$$

where t denotes the data label, and $\mathbf{x}(t) = [x, y]^T$ is given by the meshed values of $[x, y]$ in the input region $[0, 1]^2$. Hence each of the fuzzy rule $\Phi^{(j)} = N_j(\mathbf{x}(t))\mathbf{x}(t)$ spans a 2 dimensional space, i.e. $n_j = 2$, $\forall j$. The proposed algorithm based on the extended Gram-Schmidt orthogonal decomposition has been applied, in which each rule subspace being spanned by a 2-dimensional rule basis is mapped into orthogonal matrix subspaces. The modelling results contain rule based information of percentage energy increment (or the model error reduction ratio) by the selected rule to the model as shown in Table. V for $\lambda_j = 10^{-4}$, $\beta = 0.01$. The MSE of the resultant 20-rule model is 3.4527×10^{-4}. In Table V, the selected rules are ordered in the sequence of being selected, and the model selection automatically terminates at a 20-rule model ($[cerr]_{21} < 0$). The model prediction of the 20-rule model is shown in Figure 5.(b). For this example, the modelling results are insensitive to value of λ_j. It has shown

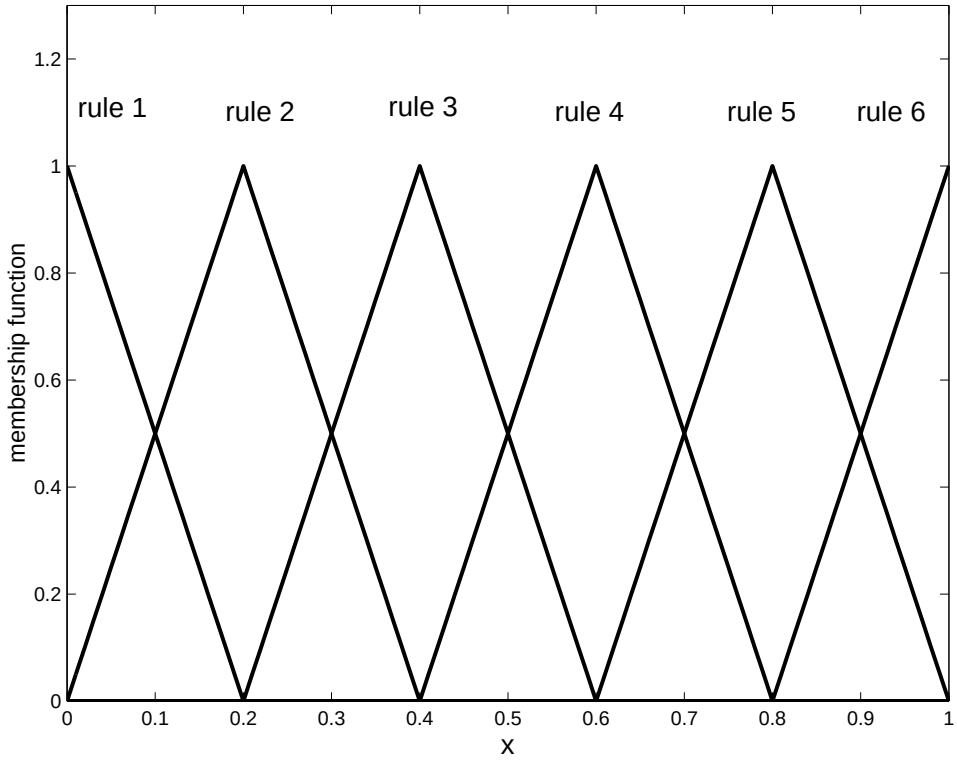

Fig. 2. The fuzzy membership functions for x in Example 1.

TABLE I

FUZZY RULES IDENTIFIABILITY IN EXAMPLE 1

Rule Index j	1	2	3	4	5	6
$\frac{1}{N}\sum_{t=1}^{N} N_j(t)$	0.0892	0.2127	0.2158	0.1816	0.2011	0.0996

TABLE II

SYSTEM ERROR REDUCTION RATIO BY THE SELECTED RULES IN EXAMPLE 1.

Rule Index j	5	3	4	6	2	1
$[err]_j(t),\ \lambda_j = 0$	0.4290	0.2959	0.1356	0.0565	0.0386	0.0000
$[err]_j(t),\ \lambda_j = 0.2$	0.4251	0.2860	0.1329	0.0556	0.0338	0.0000

TABLE III

MODEL MEAN SQUARES ERRORS (MSE) FOR NOISY OBSERVATIONS AND UNDERLYING FUNCTION.

Modelling results	6 rule model	5 rule model
$E[y(x) - \hat{z}(x)]^2,\ \lambda_j = 0$	9.8765×10^{-5}	9.8676×10^{-5}
$E[y(x) - \hat{z}(x)]^2,\ \lambda_j = 0.2$	9.8866×10^{-5}	9.8853×10^{-5}
$E[z(x) - \hat{z}(x)]^2,\ \lambda_j = 0$	7.8336×10^{-7}	7.1251×10^{-7}
$E[z(x) - \hat{z}(x)]^2,\ \lambda_j = 0.2$	5.6524×10^{-7}	5.5833×10^{-7}

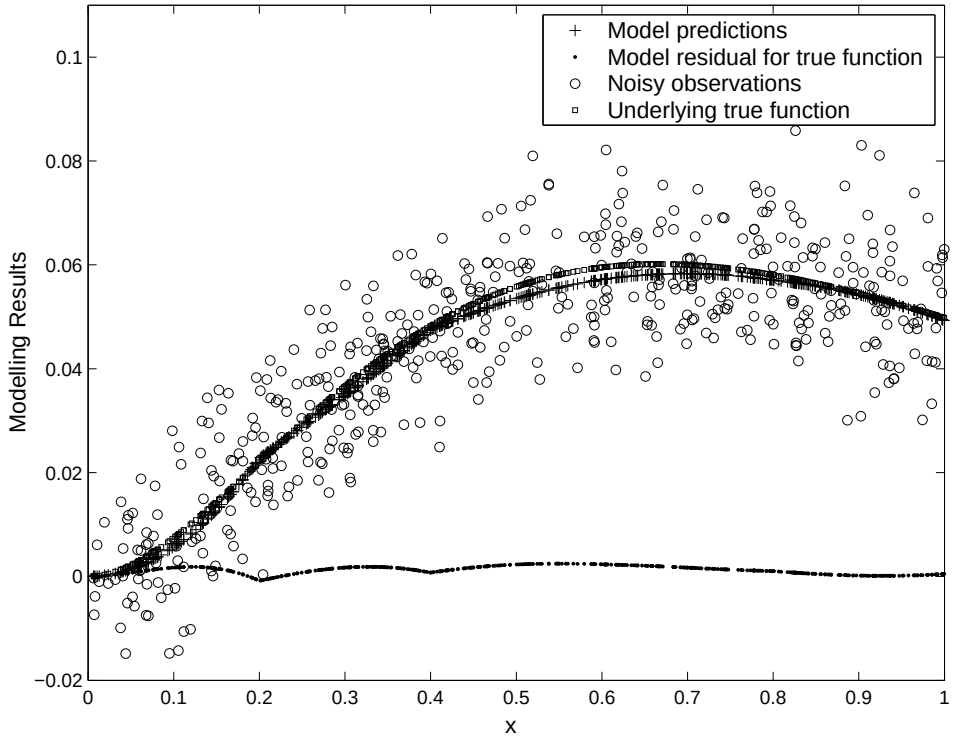

Fig. 3. The modelling results of a 5-rule model with $\lambda_j = 0.2$ for Example 1.

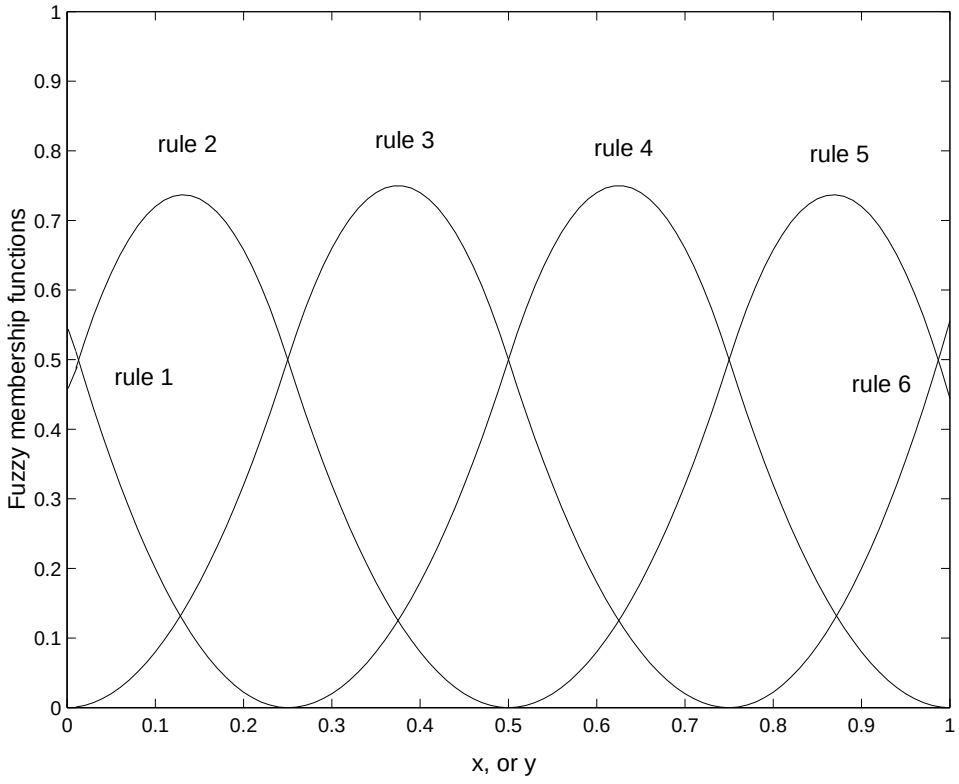

Fig. 4. The fuzzy membership functions for x, or y in Example 2.

TABLE IV
FUZZY RULES IDENTIFIABILITY IN EXAMPLE 2; (A)RULES ABOUT x; (B)RULES ABOUT y; (C)RULES ABOUT X AND Y (THE STAR '* ' INDICATES RULES INCLUDED IN THE FINAL MODEL)

(a)

Rules (x)	VS	S	MS	ML	L	VL
$\frac{1}{N}\sum_{t=1}^{N} N_j(t)$	0.0510*	0.2039*	0.2451*	0.2451	0.2039*	0.0510

(b)

Rules (y)	VS	S	MS	ML	L	VL
$\frac{1}{N}\sum_{t=1}^{N} N_j(t)$	0.0510*	0.2039	0.2451	0.2451*	0.2039*	0.0510*

(c)

$\frac{1}{N}\sum_{t=1}^{N} N_j(t)$		Rules (x)					
		VS	S	MS	ML	L	VL
Rules (y)	VS	0.0026*	0.0104	0.0125	0.0125	0.0104	0.0026*
	S	0.0104	0.0416	0.0500*	0.0500	0.0416	0.0104
	MS	0.0125	0.0500*	0.0601*	0.0601*	0.0500	0.0125*
	ML	0.0125	0.0500*	0.0601*	0.0601*	0.0500*	0.0125
	L	0.0104	0.0416	0.0500	0.0500	0.0416	0.0104
	VL	0.0026	0.0104*	0.0125	0.0125	0.0104	0.0026

that by using a weighting for the D-optimality cost function, the entire model construction procedure becomes automatic. It can be seen that the model has some limitations over the modelling of corner and edge of the surface due to the data being only piecewise smooth and piecewise nonlinear. This factor may contribute to the fact that regularization may not help in reducing misfit in some strong nonlinear behavior region. Global nonlinear modelling using B-spline for strong nonlinear behavior such as piecewise smooth and piecewise nonlinear data is under investigation.

Example 3: Consider the benchmark *Henon* time series given by

$$y(t) = 1.4 - y^2(t-1) + 0.3y(t-2) \tag{38}$$

500 data points are generated with an initial condition $y(0) = 0, y(1) = 0$. All the data points were used in the modelling by using the proposed approach. The modelling process is briefly described here. The input vector $\mathbf{x}(t) = [y(t-1), y(t-2)]^T$. For each input, define a knot vector $[-2.0, -1.9, -1.8, 0, 1.8, 1.9, 2.0]$, and use a piecewise quadratic B-spline fuzzy membership function to build a one-dimensional model, resulting $M = 4$ basis functions, corresponding to 6 fuzzy rules. That is, for $i = 1, 2$, (1) If (*y(t-i)*) is (small) (S); (1) If (*y(t-i)*) is (Medium Small) (MS);(1) If (*y(t-i)*) is (Medium Large) (ML);(1) If (*y(t-i)*) is (Large) (L); Then bivariate membership functions are formed by using tensor product.

The modelling results derived by the subspace forward regression process, with $\lambda = 0.001$, $\beta = 1 \times 10^{-4}$, is given in Table VI, with the final model consisting of 13 fuzzy rules. This table shows the energy level per rule extracted for this chaotic time series. Figure6 demonstrates the excellent approximation of the derived model. The final model MSE is 0.0041. This is very small compared to signal variance of 1.01.

V. CONCLUSIONS

This paper has introduced a new robust neurofuzzy model construction algorithm for the modelling of *a priori* unknown dynamical systems in the form of a set of fuzzy rules. A one to one mapping between a fuzzy rule base and a model matrix feature subspace has been established by extending a Takagi and Sugeno (T-S) inference mechanism. Rule based knowledge are extracted from matrix subspace to enhance model transparency due to this mapping link. In order to achieve maximized model robustness and sparsity, a new robust extended Gram-Schmidt method has been introduced via two effective and complementary approaches of regularization and D-optimality experimental design. By combining a subspace approach and the concept of robust model construction, a locally regularized orthogonal least squares algorithm is extended for fuzzy rule regularization and subspace based information extraction, and by combined with a D-optimality for subspace based rule selection. Model rule bases are decomposed into orthogonal subspaces, so as to enhance model transparency with the capability of interpreting the derived rule base energy level, and are automatically selected for a model with robustness.

APPENDIX I
THE ALGORITHM

An extended classical Gram-Schmidt scheme combined with parameter regularization and D-optimality selective criterion in orthogonal subspaces can be summarized as the following procedure:

TABLE V

SYSTEM ERROR REDUCTION RATIO BY THE SELECTED RULES IN EXAMPLE 2.

Selected Rules	$(x\ is\ ML)$ and $(y\ is\ MS)$	$(x\ is\ MS)$ and $(y\ is\ S)$	$(x\ is\ ML)$ and $(y\ is\ ML)$	$(x\ is\ MS)$	$(y\ is\ ML)$
$[err]_j(t)$	0.7044	0.1181	0.0954	0.0292	0.0193
Selected Rules	$(x\ is\ VS)$	$(x\ is\ VL)$ and $(y\ is\ MS)$	$(x\ is\ MS)$ and $(y\ is\ MS)$	$(x\ is\ S)$ and $(y\ is\ MS)$	$(y\ is\ L)$
$[err]_j(t)$	0.0086	0.0056	0.0040	0.0028	0.0021
Selected Rules	$(y\ is\ VL)$	$(x\ is\ S)$	$(x\ is\ VS)$ and $(y\ is\ VS)$	$(x\ is\ MS)$ and $(y\ is\ ML)$	$(x\ is\ S)$ and $(y\ is\ ML)$
$[err]_j(t)$	0.0037	0.0016	0.0011	0.0004	0.0003
Selected Rules	$(x\ is\ L)$ and $(y\ is\ ML)$	$(x\ is\ VL)$ and $(y\ is\ VS)$	$(x\ is\ L)$	$(y\ is\ VS)$	$(x\ is\ S)$ and $(y\ is\ VL)$
$[err]_j(t)$	0.0002	0.0005	0.0000	0.0002	0.0001

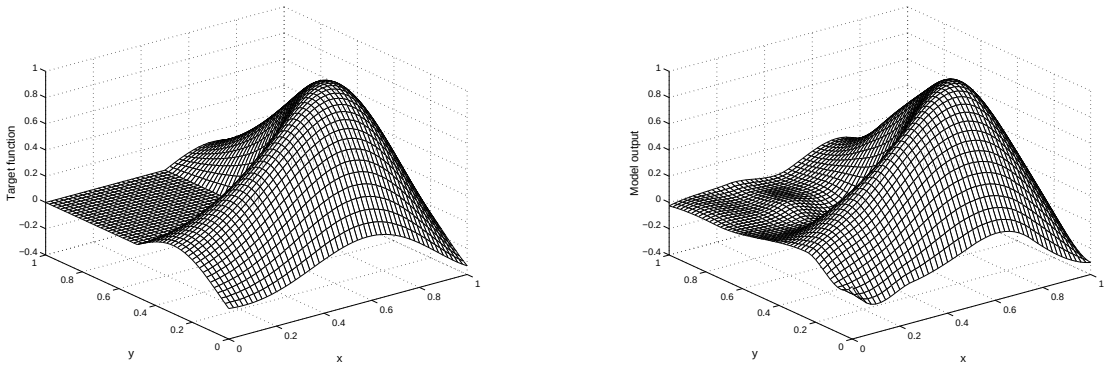

Fig. 5. Modelling results for Example 2.

TABLE VI

SYSTEM ERROR REDUCTION RATIO BY THE SELECTED RULES IN EXAMPLE 3.

Selected Rules	$y(t-2)\ (ML)$	$y(t-1)\ (L)$	$y(t-1)\ (ML)$	$y(t-2)\ (MS)$	$y(t-1)\ (S)$
$[err]_j(t)$	0.5509	0.2281	0.1024	0.0724	0.0307
Selected Rules	$y(t-2)\ (L)$	$y(t-1)\ (MS)$ and $y(t-2)\ (L)$	$y(t-1)\ (ML)$ and $y(t-2)\ ML)$	$y(t-1)\ (MS)$ and $y(t-2)\ (MS)$	$y(t-1)\ (ML)$ and $y(t-2)\ (L)$
$[err]_j(t)$	0.0102	0.0007	0.0005	0.0001	2×10^{-5}
Selected Rules	$y(t-1)\ (ML)$ and $y(t-2)\ (S)$	$y(t-1)\ (S)$ and $y(t-2)\ (S)$	$y(t-1)\ (S)$ and $y(t-2)\ (MS)$		
$[err]_j(t)$	3×10^{-5}	1×10^{-7}	2×10^{-7}		

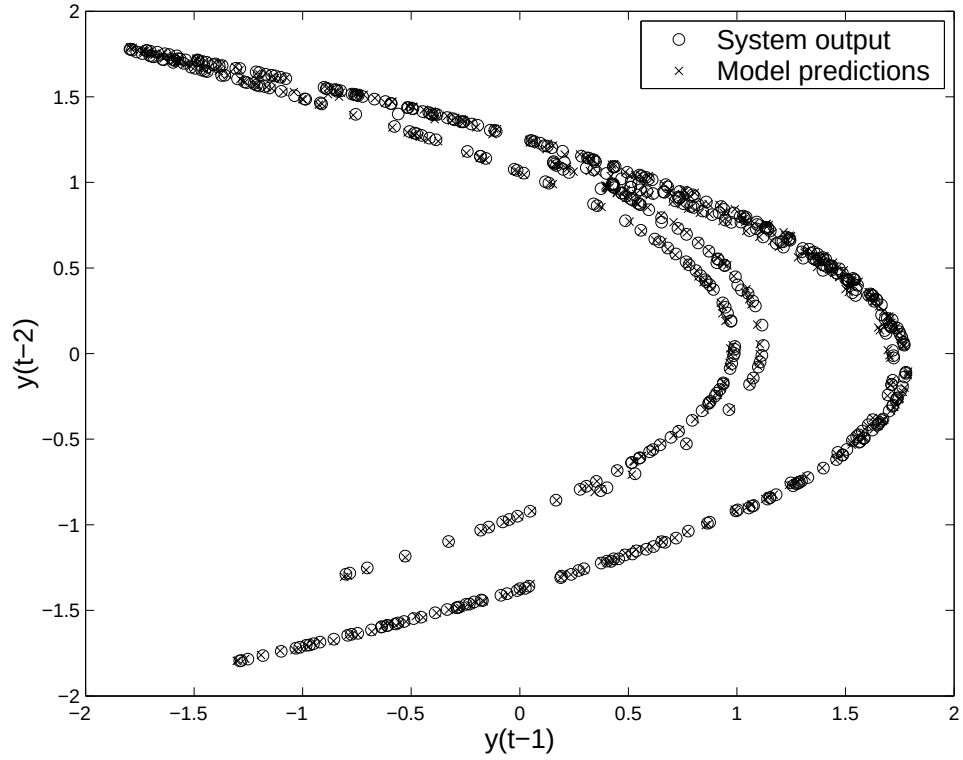

Fig. 6. Modelling results for Example 3.

1) At the jth forward regression step, where $j \geq 1$, for $j \leq l \leq K$, compute

$$A_{i,j}^{(l)} = \begin{cases} \left[\mathbf{D}^{(i)}\right]^{-1} (\mathbf{W}^{(i)})^T \mathbf{\Phi}^{(l)} \\ \quad \text{for } i = 1, \cdots, j-1 \quad \text{and if } j \neq 1 \\ \mathbf{I}_{n_l \times n_l} \qquad\qquad \text{if } i = j \end{cases}$$

$$\mathbf{W}_{(l)}^{(j)} = \begin{cases} \mathbf{\Phi}^{(l)} & \text{if } j = 1 \\ \mathbf{\Phi}^{(l)} - \sum_{i=1}^{j-1} \mathbf{W}^{(i)} A_{i,j}^{(l)} & \text{if } j > 1 \end{cases}$$

$$\mathbf{D}_{(l)}^{(j)} = [\mathbf{W}_{(l)}^{(j)}]^T \mathbf{W}_{(l)}^{(j)} \tag{39}$$

$$\mathbf{c}_j^{(l)} = [\mathbf{D}_{(l)}^{(j)} + \lambda_j \mathbf{I}]^{-1} [\mathbf{W}_{(l)}^{(j)}]^T \mathbf{y} \tag{40}$$

$$[cerr]_j^{(l)} = \frac{[\mathbf{c}_j^{(l)}]^T (\mathbf{D}_{(l)}^{(j)} + \lambda_j \mathbf{I}) \mathbf{c}_j^{(l)} + \beta \log\left[\det(\mathbf{D}_{(l)}^{(j)})\right]}{\mathbf{y}^T \mathbf{y}} \tag{41}$$

Find

$$[cerr]_j^{(l_j)} = \max\{[cerr]_j^{(l)}, \quad j \leq l \leq K\}. \tag{42}$$
$$(NB: \text{ for rule base selection})$$

and select

$$\begin{aligned} A_{i,j} &= A_{i,j}^{(l_j)}, \quad \text{for } i = 1, \cdots, j \\ \mathbf{W}^{(j)} &= \mathbf{W}_{(l_j)}^{(j)} = \mathbf{\Phi}^{(l_j)} - \sum_{i=1}^{j-1} \mathbf{W}^{(i)} A_{i,j} \\ [cerr]_j &= [cerr]_j^{(l_j)} \\ \mathbf{c}_j &= \mathbf{c}_j^{(l_j)} \\ \mathbf{D}^{(j)} &= [\mathbf{W}^{(j)}]^T \mathbf{W}^{(j)} \\ [err]_j &= \frac{[\mathbf{c}_j]^T \mathbf{D}^{(j)} \mathbf{c}_j}{\mathbf{y}^T \mathbf{y}} \end{aligned} \tag{43}$$
$$(NB: \text{ for selected rule base energy level information extraction})$$

The selected sub-matrix $\mathbf{\Phi}^{(l_j)}$ exchanges columns with sub-matrix $\mathbf{\Phi}^{(j)}$. For notational convenience, all the sub-matrices will still be referred as $\mathbf{\Phi}^{(j)}, j = 1, \cdots, K$, according to the new column sub-matrix order j in $\mathbf{\Phi}$, even if some of the column sub-matrices have been interchanged.

2) The procedure is monitored and terminated at the derived $j = n_f$ step, when $[cerr]_{n_f} \leq 0$, for a predetermined $\beta > 0$. Otherwise, set $j = j + 1$, goto step 1.
3) Calculate the original parameters according to (27).

ACKNOWLEDGMENT

XH gratefully acknowledges that part of this work was supported by EPSRC in the UK. The authors would like to thank the referees for the constructive comments.

REFERENCES

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

[2] M. Brown and C. J. Harris, *Neurofuzzy Adaptive Modelling and Control*, Prentice Hall, Hemel Hempstead, 1994.

[3] K. M. Bossley, *Neurofuzzy Modelling Approaches in System Identification*, Ph.D. thesis, Dept of ECS, University of Southampton, 1997.

[4] R. Murray-Smith and T. A. Johansen, *Multiple Model Approaches to Modelling and Control*, Taylor and Francis, 1997.

[5] T. Takagi and M. Sugeno, "Fuzzy identification of systems and its applications to modelling and control," *IEEE Trans. on Systems, Man, and Cybernetics*, vol. 15, pp. 116–132, 1985.

[6] M. Feng and C. J. Harris, "Adaptive neurofuzzy control for a class of state-dependent nonlinear processes," *International Journal of Systems Science*, vol. 29, no. 7, pp. 759–771, 1998.

[7] H. Wang, M. Brown, and C. J. Harris, "Modelling and control of nonlinear, operating point dependent systems via associative memory networks," *J. Dynamics and Control*, vol. 6, pp. 199–218, 1996.

[8] R. Bellman, *Adaptive Control Processes*, Princeton University Press, 1966.

[9] J. S. R. Jang, C.T. Sun, and E. Mizutani, *Neuro-Fuzzy and Soft Computing: A Computational Approach to Learning and Machine Intelligence*, Upper Saddle River, NJ : Prentice Hall, 1997.

[10] S. Chen, S. A. Billings, and W. Luo, "Orthogonal least squares methods and their applications to non-linear system identification," *International Journal of Control*, vol. 50, pp. 1873–1896, 1989.

[11] S. Chen, Y. Wu, and B. L. Luk, "Combined genetic algorithm optimization and regularized orthogonal least squares learning for radial basis function networks," *IEEE Trans. on Neural Networks*, vol. 10, pp. 1239–1243, 1999.

[12] M. J. L. Orr, "Regularisation in the selection of radial basis function centers," *Neural Computation*, vol. 7, no. 3, pp. 954–975, 1995.

[13] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. on Automatic Control*, vol. AC-19, pp. 716–723, 1974.

[14] A. C. Atkinson and A. N. Donev, *Optimum Experimental Designs*, Clarendon Press, Oxford, 1992.

[15] X. Hong and C. J Harris, "Nonlinear model structure detection using optimum experimental design and orthogonal least squares," *IEEE Transactions on Neural Networks*, vol. 12, no. 2, pp. 435–439, 2001.

[16] X. Hong and C. J Harris, "Nonlinear model structure design and construction using orthogonal least squares and d-optimality design," *IEEE Transactions on Neural Networks*, vol. 13, no. 5, pp. 1245–1250, 2001.

[17] S. Chen, "Local regularization assisted orthogonal least squares regression," *Submitted to International Journal of Control*, 2003.

[18] S. Chen, X. Hong, and C. J. Harris, "Sparse kernel regression modelling using combined locally regularised orthogonal least squares and d-optimality experimental design," *IEEE Trans. on Automatic Control*, p. Accepted., 2003.

[19] T. Taniguchi, K. Tanaka, H. Ohtake, and H. O. Wang, "Model construction, rule reduction and rubust compensation for generalised form of takagi-sugeno fuzzy systems," *IEEE Trans. on Fuzzy Systems*, vol. 9, no. 4, pp. 525–538, 2001.

[20] Y. Jin, "Fuzzy modelling of high dimensional systems: complexity reduction and interpretability improvement," *IEEE Trans. on Fuzzy Systems*, vol. 8, no. 2, pp. 212–221, 2000.

[21] H. Roubos and M. Setnes, "Compact and transparent fuzzy models and classifiers through iterative complexity reduction," *IEEE Trans. on Fuzzy Systems*, vol. 9, no. 4, pp. 516–524, 2001.

[22] X. Hong and C. J Harris, "A neurofuzzy network knowledge extraction and extended gram-schmidt algorithm for model subspace decomposition," *IEEE Transactions on Fuzzy Systems*, vol. Accepted, 2002.

[23] P. Dierckx, *Curve and Surface Fitting with Splines*, Clarendon Press, Oxford, 1995.

[24] S. R. Gunn, M. Brown, and K. Bossley, "Network performance assessment for neurofuzzy data modelling," in *Intelligent Data Analysis*, pp. 313–323. 1997.

[25] K. W. Gruenberg and A. J. Weir, *Linear Geometry*, D. Van Nostrand Company, Inc., 1967.

[26] T. Soderström and P. Stoica, *System Identification*, Prentice Hall, 1989.

[27] D. W. Marquardt, "Generalised inverse, ridge regression, biased linear estimation and nonlinear estimation," *Technometrics*, vol. 12, no. 3, pp. 591–612, 1970.

PLACE PHOTO HERE

Xia Hong received her university education at National University of Defense Technology, P.R.China (BSc, 1984, MSc, 1987), and University of Sheffield, UK (PhD,1998), all in automatic control. She worked as a research assistant in Beijing Institute of Systems Engineering, Beijing, China from 1987-1993. She worked as a research fellow in the Department of Electronics and Computer Science at University of Southampton from 1997-2001. She is currently a lecturer at Dept of Cybernetics, University of Reading. She is actively engaged in research into neurofuzzy systems, data modelling and learning theory and their applications. Her research interests include system identification, estimation, neural networks, intelligent data modelling and control. She has published over 30 research papers , an coauthored a research book. She was awarded a Donald Julius Groen Prize by IMechE in 1999.

PLACE PHOTO HERE

Chris Harris received university education at Leicester (BSc), Oxford (MA) and Southampton (PhD). He previously held appointments at the Universities of Hull, UMIST, Oxford and Cranfield, as well as being employed by the UK Ministry of Defence. His research interests are in the area of intelligent and adaptive systems theory and its application to intelligent autonomous systems, management infrastructures, intelligent control and estimation of dynamic processes, multi-sensor data fusion and systems integration. He has authored or co-authored 12 books and over 300 research papers, and he is the associate editor of numerous international journals including Automatica, Engineering Applications of AI, Int. J. General Systems Engineering, International J. of System Science and the Int. J. on Mathematical Control and Information Theory. He was elected to the Royal Academy of Engineering in 1996, was awarded the IEE Senior Achievement medal in 1998 for his work on autonomous systems, and the highest international award in IEE, the IEE Faraday medal in 2001 for his work in Intelligent Control and Neurofuzzy System.

PLACE PHOTO HERE

Sheng Chen obtained a BEng degree in control engineering from the East China Petroleum Institute in 1982, and a PhD degree in control engineering from the City University at London in 1986. He joined the University of Southampton in September 1999. He previously held research and academic appointments at the Universities of Sheffield, Edingburgh and Portsmouth. Dr Chen is a Senior Member of IEEE in the USA. His recent research works include adaptive nonlinear signal processing, modelling and identification of nonlinear systems, neural network research, finite-precision digital controller design, evolutionary computation methods and optimization. He has published over 200 research papers.