A Neurofuzzy Network Knowledge Extraction and Extended Gram–Schmidt Algorithm for Model Subspace Decomposition

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

***Abstract*—This paper introduces 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 modeling of *a priori* unknown dynamical systems in the form of a set of fuzzy rules. The first contribution of the paper is the introduction of a one to one mapping between a fuzzy rule-base and a model matrix feature subspace using the T–S inference mechanism. This link enables the numerical properties associated with a rule-based matrix subspace, the relationships amongst these matrix subspaces, and the correlation between the output vector and a rule-base matrix subspace, to be investigated and extracted as rule-based knowledge to enhance model transparency. The matrix subspace spanned by a fuzzy rule is initially derived as the input regression matrix multiplied by a weighting matrix that consists of the corresponding fuzzy membership functions over the training data set. Model transparency is explored by the derivation of an equivalence between an A-optimality experimental design criterion of the weighting matrix and the average model output sensitivity to the fuzzy rule, so that rule-bases can be effectively measured by their identifiability via the A-optimality experimental design criterion. The A-optimality experimental design criterion of the weighting matrices of fuzzy rules is used to construct an initial model rule-base. An extended Gram–Schmidt algorithm is then developed to estimate the parameter vector for each rule. 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. This new approach is computationally simpler than the conventional Gram–Schmidt algorithm for resolving high dimensional regression problems, whereby it is computationally desirable to decompose complex models into a few submodels rather than a single model with large number of input variables and the associated curse of dimensionality problem. Numerical examples are included to demonstrate the effectiveness of the proposed new algorithm.**

***Index Terms*—Least squares, mixtures of experts, neurofuzzy networks, orthogonal decomposition, subspace.**

I. Introduction

ASSOCIATIVE memory networks [such as B-spline networks, radial basis functions (RBFs), support vector machines (SVM)] have been extensively developed [1]–[4]. While common with other neural networks, they can approximate with arbitrary accuracy complex nonlinear systems, they are additionally linear in their adjustable parameters or weights, offering the potential for online learning with provable convergence and stability properties [1], [2]. Most conventional neural networks lead only to "black box" model representation. For problems which require insight into the underlying phenomenology, model transparency is critical, i.e., internal system behavior interpretability and/or knowledge (rule) representation of the underlying process. A model with good transparency properties helps users to understand the system behaviors, oversee critical system operating regions, and/or extract physical laws or relationships that underpin the system. Exceptionally a neurofuzzy network system has the desirable properties of a compact support, a partition of unity, locality, logicality, and transparency via fuzzy rules. The transparency property of a neurofuzzy system is essential for the fuzzy rule extraction capabilities of the derived process model. The inherent transparency of a neurofuzzy network lies in the property of unity of support, i.e., the model output can be decomposed into a convex combination of the outputs of individual rules, so that the basis function can be interpreted as a fuzzy membership function of individual rules. Based on the fuzzy rules inference and model representation of Takagi and Sugeno (T–S) [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 function bases in a neurofuzzy system can be directly related to linguistic fuzzy logic rules under limited conditions, so that any model based on numerical information can be equivalently related to an associated set of fuzzy logic rules.

The problem of *the curse of dimensionality* [8] has been a main obstacle in nonlinear modeling 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), RBF, Karneva distributed memory maps, and all neurofuzzy networks (e.g., adaptive network based fuzzy inference system (ANFIS) [9], T–S 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

Manuscript received December 19, 2001; revised September 10, 2002 and October 9, 2002.

X. Hong is with the Department of Cybernetics, University of Reading, Reading RG6 6AY, U.K. (e-mail: x.hong@reading.ac.uk).

C. J. Harris is with the Image, Speech and Intelligent Systems Group, Department of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, U.K. (e-mail: cjh@ecs.soton.ac.uk).

Digital Object Identifier 10.1109/TFUZZ.2003.814842

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]. To enable the applicability of the OLS algorithm in neurofuzzy systems, a NeuDec algorithm has been developed to incorporate the OLS algorithm with an experimental design optimality criteria for the efficient model structure determination and estimation [11], [12].

In practice, data based neurofuzzy model construction algorithms have to utilize finite data sets to generate parsimonious models, such that the final model parameterization is adequately based on the amount of data, its distribution, and associated model identifiability. 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 multidimensional input variables), resulting in opaque models during the learning process. Since modeling is inevitably iterative it can be greatly enhanced if the modeler 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 modeling [13], various fuzzy similarity measures combined with genetic algorithms [14], [15].

In this paper, a new neurofuzzy model construction and parameter estimation algorithm in the form of fuzzy rules, is introduced based on the T–S inference mechanism and a new extended Gram–Schmidt orthogonal decomposition algorithm, for the modeling of *a priori* unknown dynamical systems based on finite data sets. A functional inference of a fuzzy rule as a matrix feature subspace is introduced based on an extension of the T–S inference mechanism to achieve a rule-based neurofuzzy system with exceptional rule extraction capabilities throughout the modeling process. Model transparency during learning is achieved because the proposed algorithm is a rule-based learning approach, based on the matrix feature subspace that is uniquely related to a fuzzy rule, enabling the numerical properties associated with a rule-based matrix subspace, the relationships between these matrix subspaces, and the associated correlation between the output vector and a rule-base matrix subspace, to be investigated and extracted as rule-based knowledge.

In optimum experimental design, the A-optimality experimental design criteria is usually a function of the eigenvalues of the model regression matrix, which in turn reflects the variance of parameter estimates [16] or the identifiability of an associated parameter vector. The subspace matrix of a fuzzy rule is derived as the input regression matrix weighted by a weighting matrix consisting of the corresponding fuzzy membership functions over the data set. Here, model transparency is explored by the derivation of an equivalence between an A-optimality experimental design criterion of this weighting matrix and the average model output sensitivity to the fuzzy rule, so that extracted rule-bases can be effectively measured by their identifiability via the A-optimality design criterion. The A-optimality experimental design criterion of the weighting matrices of fuzzy rules is used to construct an initial model base, i.e., any model base not satisfying an identifiability condition is excluded. This is advantageous in increasing model transparency during the initial model rule-base construction stage. An extended Gram–Schmidt algorithm is then developed and applied to estimate the parameter vector for each derived rule. The proposed new algorithm decomposes the model bases via an orthogonal subspace decomposition approach, with the advantage of relating rule-bases directly to the matrix feature subspaces so as to enhance model transparency with the capability of interpreting the rule-base in terms of its energy level. The computed output variance explained by the associated rules (the associated energy level) can also used as model final structure determination, as well as extracted as rule-based knowledge transparent to users.

This paper is organized as follows. Section II introduces a general class of neurofuzzy systems as a modeling approach. Section III introduces the proposed modeling approach, with theoretic analysis into the associated model transparency. Numerical examples are provided in Section IV to illustrate the effectiveness of the approach and Section V is devoted to conclusions.

II. Preliminaries

This section briefly presents a general class of neurofuzzy systems as a nonlinear data modeling 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

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

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.

Utilizing the principle of divide and conquer, model (1) can be simplified by decomposing it into a set of K local models $f_i(\mathbf{x}^{(i)}(t), \boldsymbol{\Theta}_i)$, $i = 1, \ldots, K$, where K is to be determined, each of which operates on a local region depending on the submeasurement vector $\mathbf{x}^{(i)} \in \Re^{n_i}$, a subset of the input vector $\mathbf{x}$, i.e., $\mathbf{x}^{(i)} \in \mathcal{X}_i \in \Re^{n_i}$, $(n_i < n)$, $\mathcal{X}_1 \cup \cdots \cup \mathcal{X}_K = \mathcal{X}$.

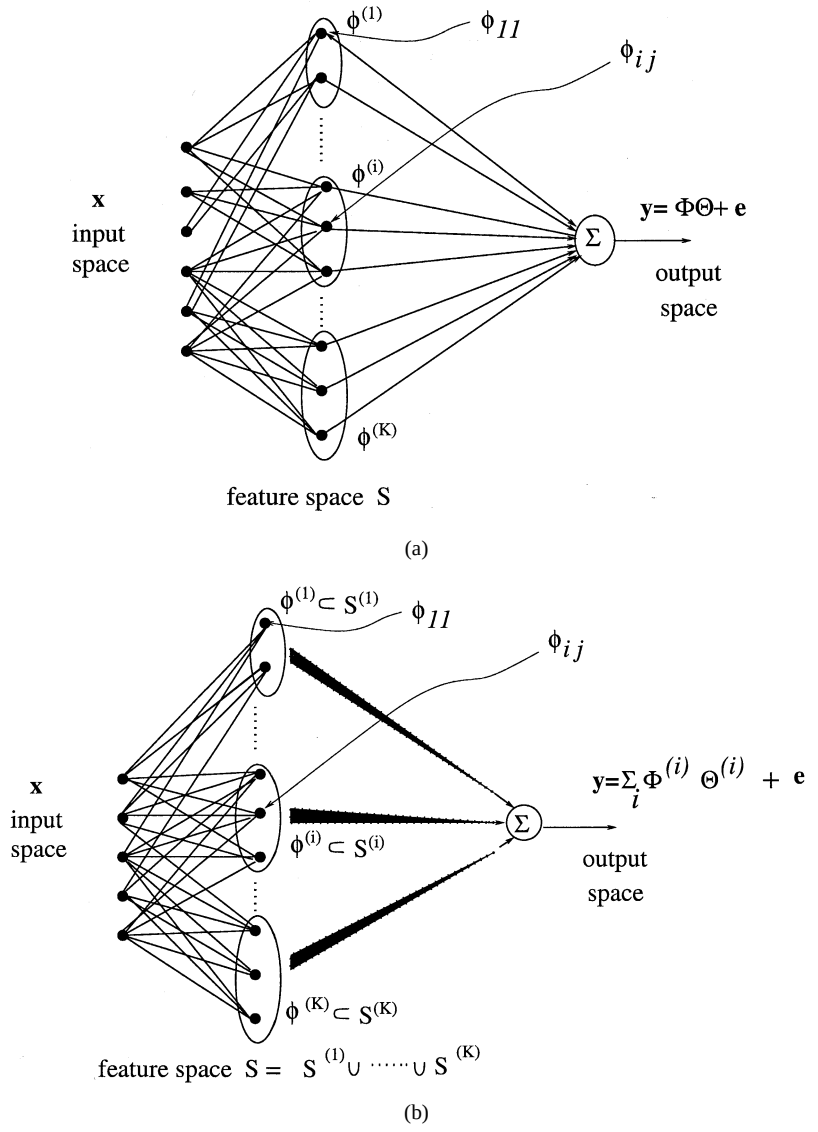

Fig. 1. Two views on a neurofuzzy network. (a) Based on p basis functions. (b) Based on K fuzzy rules consisting of n_i input variables.

Each of the local models $f_i(\mathbf{x}^{(i)}(t), \mathbf{\Theta}_i)$ can be represented by a set of linguistic rules

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

where the fuzzy set $A^{(i)} = [A_1^{(i)}, \ldots, A_{n_i}^{(i)}]^T$ 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 T–S fuzzy inference mechanism [5], the output of (1) can be represented by

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

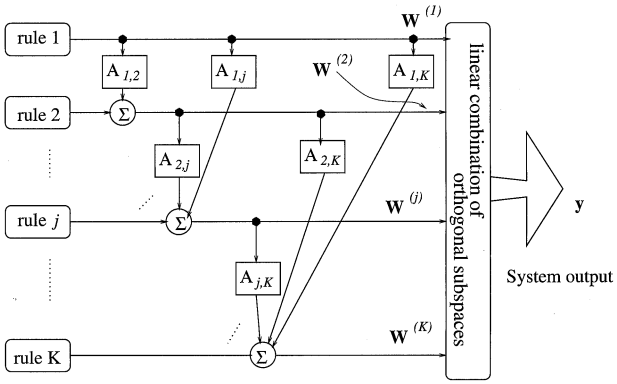

Fig. 2. Orthogonal subspaces based on fuzzy rule-bases.

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

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

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

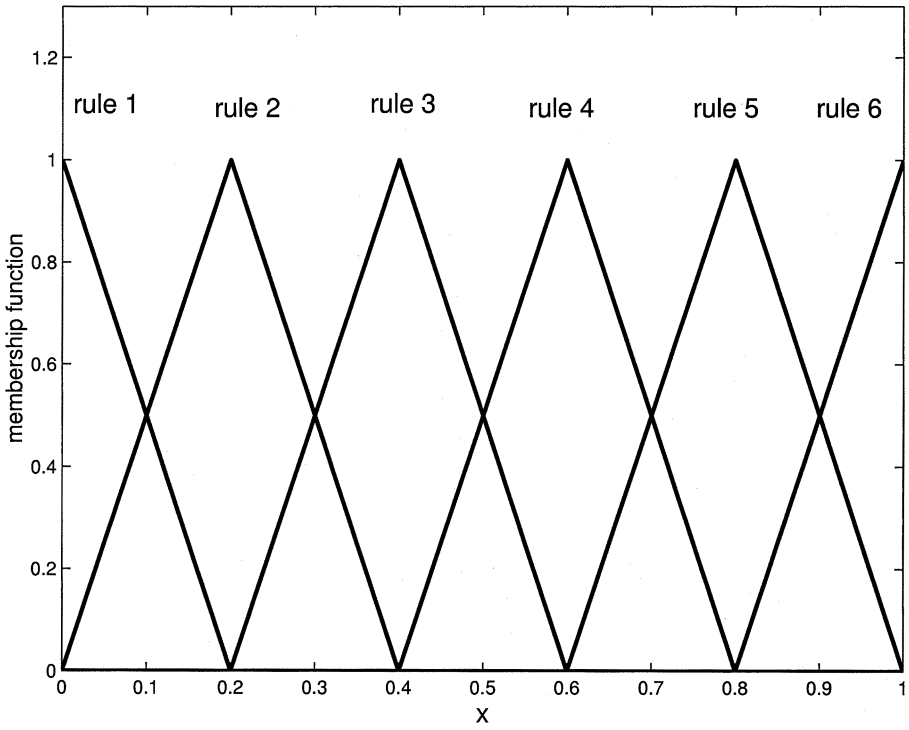

Fig. 3. Fuzzy membership functions for x in Example 1.

(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 [17] 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_js 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 $\boldsymbol{\tau} = \{\tau_1, \tau_2, \ldots, \tau_{L+m+1}\}$. The knot sequence forms a partitioning of the input domain into $(L+m)$ disjoint intervals. The basis functions set can be defined by the recursive equation [17]

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

with

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

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

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

for $i = 1, \ldots, M$, where $M = \prod_{i=1}^{n} L_i$, $\mathbf{x}^{(i)} = [x_1^{(i)}, x_2^{(i)}, \ldots, x_{n_i}^{(i)}]^T \in \Re^{n_i}$. $k_j = 1, 2, \ldots, L_j$, 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 analysis of variance (ANOVA) representation of multivariable functions uses lower dimensional tensor products of models inputs, such that the fuzzy membership functions (8) is replaced by

$$N_i\left(\mathbf{x}^{(i)}\right) = \prod_{j \in 1,\, n_i} B_{j,m}^{k_j}\left(x_j^{(i)}\right) \tag{9}$$

with the number of multiplication terms limited in practice to a low number (e.g., lower than three). For practical application, not only is the ANOVA approach effective in overcoming the curse of dimensionality, because the resultant rule-bases based on ANOVA is significant lower if the input dimension is high, it has additional advantage of model transparency because a lower input dimension than three can be visualized and interpreted [18].

Substitute (5) and (4) into (1)

$$\begin{aligned} y(t) &= \sum_{i=1}^{K} \phi_i\left(\mathbf{x}^{(i)}(t)\right)^T \boldsymbol{\Theta}_i + e(t) \\ &= \phi(\mathbf{x}(t))^T \boldsymbol{\Theta} + e(t) \end{aligned} \tag{10}$$

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

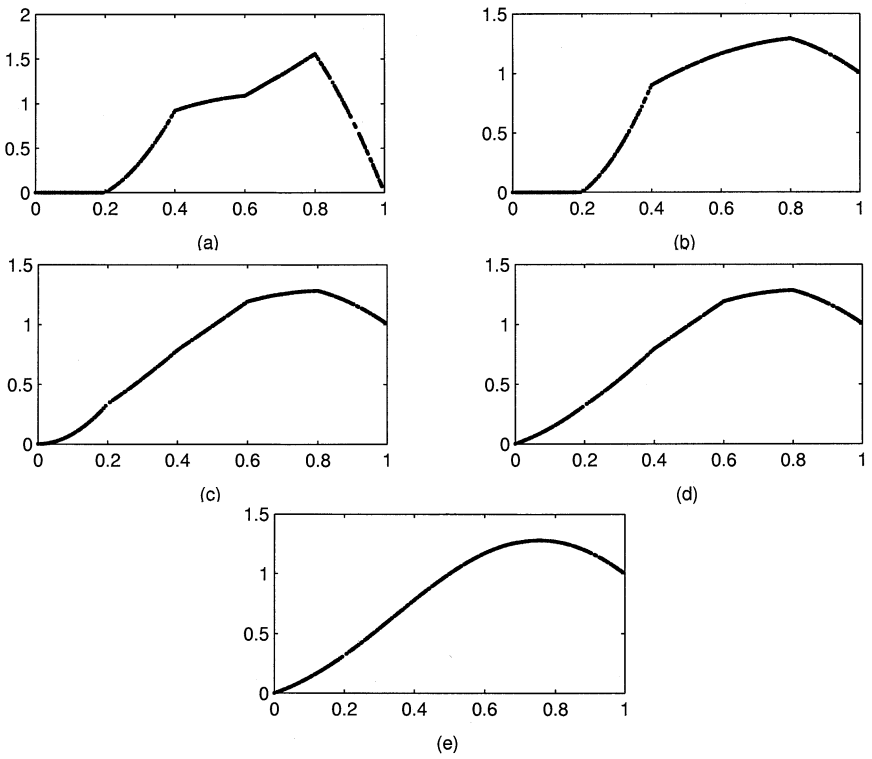

Fig. 4. Modeling processes as a forward selection of rule-bases in Example 1.

For the finite data set $D_N = \{\mathbf{x}(t), y(t)\}_{t=1}^{N}$, (10) can be written in a matrix form as

$$\mathbf{y} = \sum_{i=1}^{K} \Phi^{(i)}\boldsymbol{\Theta}^{(i)} + \mathbf{e} = \boldsymbol{\Phi}\boldsymbol{\Theta} + \mathbf{e} \tag{11}$$

where $\mathbf{y} = [y(1), y(2), \ldots, y(N)]^T \in \Re^N$ is the output vector, $\Phi^{(i)} = [\phi_i(\mathbf{x}(1)), \ldots, \phi_i(\mathbf{x}(N))]^T \in \Re^{N\times n_i}$ is the regression matrix associated with the ith fuzzy rule, $\mathbf{e} = [e(1), \ldots, e(N)]^T \in \Re^N$ is the model residual vector. $\boldsymbol{\Phi} = [\Phi^{(1)}, \ldots, \Phi^{(K)}] \in \Re^{N\times p}$ 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. This strategy is the normal basis of a wide range of algorithms such as clustering, kernel methods, and forward regression. In clustering and kernel methods, the data samples are themselves a potential model rule-base, and often used as the centers of a radial basis function. In this paper, the selection of K local models as an initial model base is based on model identifiability via the A-optimality design criterion with the advantage of enhanced model transparency to quantify and interpret fuzzy rules and their identifiability. Compared to conventional clustering, kernel methods, the proposed approach is based on a submatrix that is uniquely linked to fuzzy rule. This is advantageous in increasing model transparency because the numerical information associated with the submatrix is subsequently rule-based knowledge. In the following section, it will be shown that an A-optimality design criterion associated with a submatrix based on a fuzzy rule itself provides identifiability of the fuzzy rule. Then an initial model rule-base construction is introduced based on the A-optimality experimental design criterion that measures the identifiability of the system rule-base. This is based on the construction of an appropriate initial rule-base which is persistently excited by data. Because a persistent excitation by data is a prerequisite condition of system identification, rules lacking data excitation will be excluded from the initial model base.

TABLE I
FUZZY RULES IDENTIFIABILITIES IN EXAMPLE 1

Rule Index i	1	2	3	4	5	6
$\frac{1}{N}\sum_{t=1}^{N} N_i(t)$	0.0983	0.1839	0.1704	0.2150	0.2221	0.1104

TABLE II
SYSTEM ERROR REDUCTION RATIO BY THE SELECTED RULES IN EXAMPLE 1

Rule Index i	5	4	3	6	2	1
$[ERR]_i(t)$	0.5596	0.2622	0.0984	0.0645	0.0149	0.0003

III. RULE-BASED MODEL CONSTRUCTION AND LEARNING ALGORITHMS

While the object of a neurofuzzy network is to model and represent processes linguistically, the learning or training of the model is often carried out by conventional neural networks training algorithms, and in particular, linear learning algorithms such as least squares. Typically associative neural networks such as RBF, B-splines networks construction algorithm consists of two stages, an unsupervised stage (uses only system input but not output information) of an initial model base

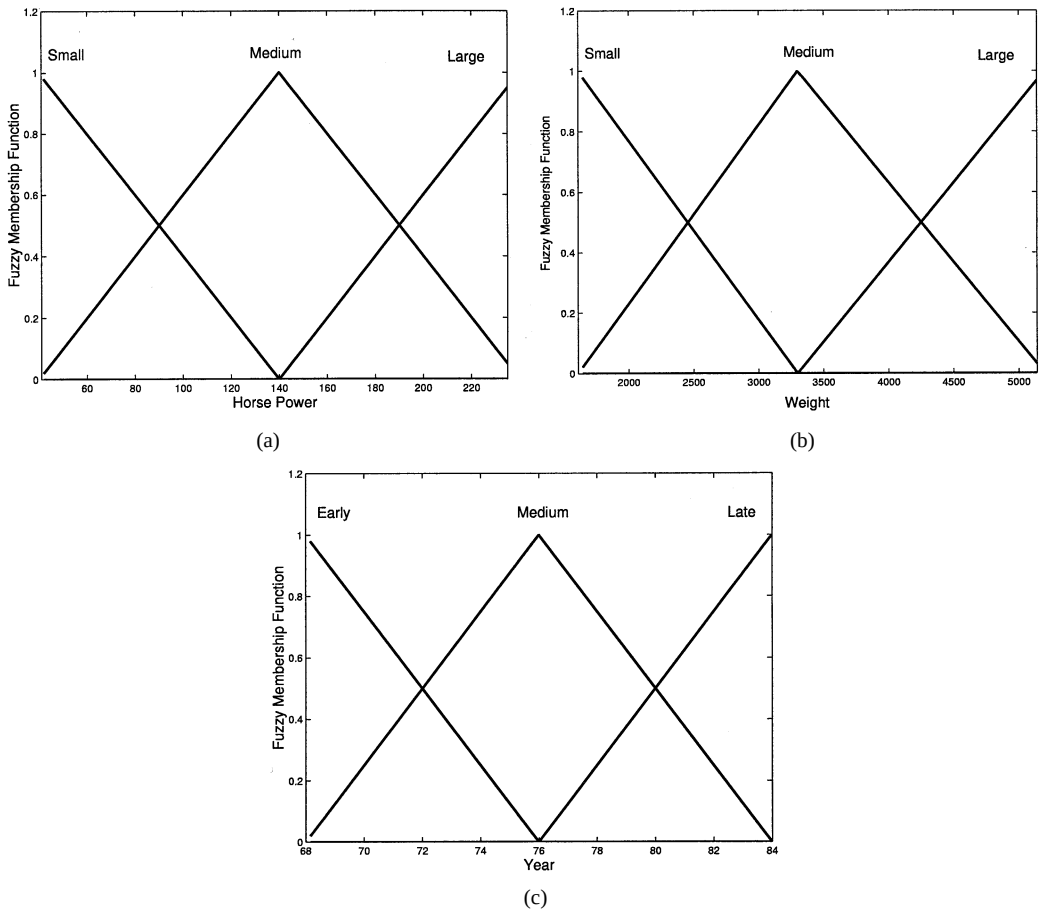

Fig. 5. Univariate fuzzy membership functions. (a) Horsepower. (b) Weight. (c) Year.

construction, followed by a supervised stage (use both system input and output information) for refined model structure detection, (with the approximation of system output as the main objective, together with a minimal model complexity). For general linear in the parameter systems, a forward orthogonal least squares (OLS) algorithm based on Gram–Schmidt orthogonal decomposition can be used to determine the significant terms and parameter estimates, and the model structure [10]. The mechanism underpinning the OLS method is to decompose the correlations amongst regressors in a linear regression equation in a forward manner, to achieve a minimal model structure whilst maximizing model approximation or generalization ability. The OLS algorithm potentially has an inherent model transparency in extracting energy levels in the selected models regressors, if only the regressors can be associated with meaningful system variables.

However, conventional least squares including the forward OLS algorithm, if applied to a neurofuzzy model based on the T–S inference mechanism, loses its model transparency during learning. This is due to the fact that learning is based upon a set of one-dimensional regressors, or basis functions (such as B-splines, Gaussians, etc.), and not upon a set of fuzzy rules (usually in the form of multidimensional input variables). Since modeling is inevitably iterative it can be greatly enhanced if the modeler can interpret or interrogate the derived rule-base during learning itself, allowing the injecting of user knowledge as well as premature cessation when the model satisfies the users requirements. The model construction algorithm developed in the following is also composed of two stages, an unsupervised stage of an initial model base construction followed by a supervised stage of fine model structure detection based on an extension of forward OLS algorithm, with both stages using rule-based learning in order to maintain the model transparency during the learning phase. In Section III-B, it is shown that each fuzzy rule can be mapped into a submatrix within the full regression matrix. The identifiability of a fuzzy rule is discussed in associated with the nonsingularity condition of the associated submatrix, and then used in the initial model construction via the A-optimality design criteria. Compared to the conventional forward OLS algorithm based on Gram–Schmidt orthogonal decomposition, the extended Gram–Schmidt algorithm developed in Section III-B extends the orthogonalization of regressors to the orthogonalization of subspaces spanned by submatrices which in turn have a one-to-one mapping with fuzzy rules. The model transparency can be achieved by extracting energy level associated fuzzy rule-bases. Note that all the advantages of linear learning are maintained because the proposed method is still essentially a linear learning approach.

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

TABLE III
FUZZY RULES IDENTIFIABILITIES IN EXAMPLE 2. (a) RULES ABOUT HORSEPOWER. (b) RULES ABOUT WEIGHT. (c) RULES ABOUT YEAR. (d) RULES ABOUT HORSE POWER AND WEIGHT. (e) RULES ABOUT HORSE POWER AND YEAR. (f) RULES ABOUT WEIGHT AND YEAR. (THE BRACKET INDICATES RULES REMOVED FROM THE RULE BASE DUE TO LOW IDENTIFIABILITIES SHOWN BY THE A-OPTIMALITY CRITERIA; AND THE STAR "*" INDICATES RULES INCLUDED IN THE FINAL MODEL)

Rules (Horse Power)	Small	Medium	Large
$\frac{1}{N}\sum_{t=1}^{N} N_i(t)$	0.4119*	0.5315*	0.0566*

(a)

Rules (Weight)	Small	Medium	Large
$\frac{1}{N}\sum_{t=1}^{N} N_i(t)$	0.3303	0.5457	0.1241

(b)

Rules (Year)	Early	Medium	Late
$\frac{1}{N}\sum_{t=1}^{N} N_i(t)$	0.1987	0.6052	0.1961

(c)

$\frac{1}{N}\sum_{t=1}^{N} N_i(t)$		Rules (Horse Power)		
		Small	Medium	Large
Rules	Small	0.2066	0.1234	(0.0003)
	Medium	0.0983	0.1999	(0.0258)
(weight)	Large	(0.0054)	0.0882	0.0305

(d)

$\frac{1}{N}\sum_{t=1}^{N} N_i(t)$		Rules (Horse Power)		
		Small	Medium	Large
Rules	Early	0.0596	0.1113	0.0277
	Medium	0.2468	0.1839*	0.0280
(Year)	Late	0.1055*	0.0897	(0.0009)

(e)

$\frac{1}{N}\sum_{t=1}^{N} N_i(t)$		Rules (Weight)		
		Small	Medium	Large
Rules	Early	0.0537	0.1042	0.0407
	Medium	0.1911	0.3367	0.0774
(Year)	Late	0.0854*	0.1048	(0.0059)

(f)

in the parameter model, as shown in Fig. 1(a). Model transparency during learning cannot be automatically achieved unless these regressors have a clear physical interpretation, or are directly associated with physical variables. Under a neurofuzzy model based on the T–S mechanism, the regressors in an ordinary p-dimensional linear in the parameter model as shown in Fig. 1(a) are not based upon a set of fuzzy rules, therefore, this is unhelpful in extracting rule-based knowledge.

Alternatively, a neurofuzzy network is inherently transparent for rule-based model construction. In (11), 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. Numerical properties associated with a rule-based matrix subspace, the relationships among these matrix subspaces, and the correlation between the output vector and a rule-base matrix subspace, are easy to investigate and be extracted as rule-based knowledge. Fig. 1(b) provides a visual illustration of a rule-based system in which the system is a linear combination of fuzzy rules system, with each rule consisting of n_i regressors.

Definition 1: Basis of a Subspace: If n_i vectors $\boldsymbol{\phi}_j^{(i)} \in \Re^N$, $j = 1, 2, \ldots, n_i$, satisfy the nonsingular condition that $\boldsymbol{\Phi}^{(i)} = [\boldsymbol{\phi}_1^{(i)}, \ldots, \boldsymbol{\phi}_{n_i}^{(i)}] \in \Re^{N \times n_i}$ has a full rank of n_i, they span a n_i-dimensional subspace $S^{(i)}$, then $\boldsymbol{\Phi}^{(i)}$ 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 T–S fuzzy inference mechanism with a unique label $N_i(\cdot)$. $S^{(i)}$ is defined as a fuzzy rule subspace of the ith fuzzy rule.

Note that model transparency is inherent in the above neurofuzzy model representation due to a one-to-one link between fuzzy membership functions and fuzzy linguistic rules, with fuzzy membership providing an indication of the importance (confidence) of the derived linguistic rule, and $f_i(\cdot)$ is the output of subsystem (fuzzy rule) which is appropriate for transparent model construction. The IF part in the fuzzy rule (3) forms the fuzzy rule basis consisting of a n_i-dimensional input vector. For a data set $D_N = \{\mathbf{x}(t), y(t)\}_{t=1}^{N}$, the IF part of (3) can be expressed as a $N \times n_i$ matrix, whose components are in the form of a linguistic variable

$$\left\{N_i(t):\ \left(x_j^{(i)}(t) \text{ is } A_j^i\right)\right\}, \qquad j = 1, \ldots, n_i,\ t = 1, \ldots, N \tag{12}$$

where $N_i(t)$ is the confidence level of fuzzy rule (3). The T–S fuzzy inference mechanism simply numerically expresses the above linguistic variable matrix as the regression matrix of the ith local model $\Phi^{(i)}$, which is also one of the K submatrices of the regression matrix $\boldsymbol{\Phi}$. It is clear from Definitions 1 and 2 that $\Phi^{(i)}$ spans a n_i-dimensional feature subspace within in the p-dimensional feature space spanned by $\boldsymbol{\Phi}$. By considering that the matrix $\Phi^{(i)}$, representing an individual rule, spans a n_i-dimensional feature subspace within the p-dimensional feature space as spanned by $\boldsymbol{\Phi}$, representing the full rule-base consisting of K rules, conventional learning algorithm can be extended as rule-based learning algorithm in which model transparency can

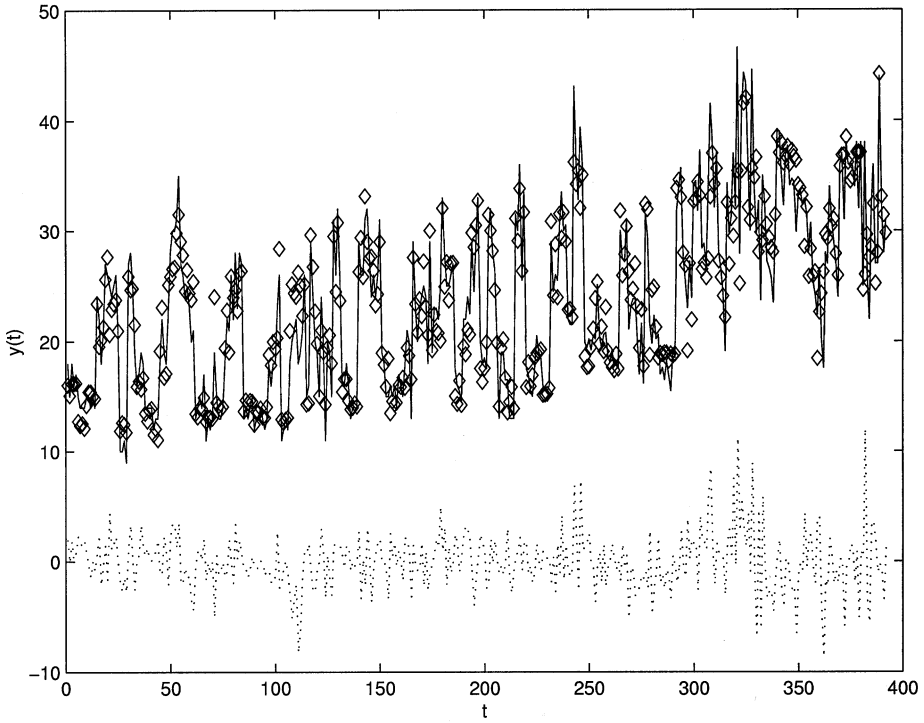

Fig. 6. Modeling of MPG data; solid line: actual MPG data; diamond: Model predictions; and dotted line: model residual.

be maintained during learning. This is achieved via extending linear in the parameter learning methods in a manner so as to extend scalar (an individual input variable) to multidimensional (a fuzzy rule consisting of n_i variables). This is also the main basis of Section III-B that extends variable regression to subspace regression, or fuzzy rule-base regressional construction based on functional subspace inference of the fuzzy rules.

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

$$\Phi^{(i)} = \mathbf{N}^{(i)} X^{(i)} \tag{13}$$

where $\mathbf{N}^{(i)} = \text{diag}\{N_i(1), \ldots, N_i(N)\} \in \Re^{N \times N}$, $X^{(i)} = [\mathbf{x}^{(i)}(1), \mathbf{x}^{(i)}(2), \ldots, \mathbf{x}^{(i)}(N)]^T \in \Re^{N \times n_i}$. Equation (13) 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 $\text{rank}(X^{(i)}) = n_i$. As

$$\text{rank}\left(\Phi^{(i)}\right) = \min\left[\text{rank}\left(\mathbf{N}^{(i)}\right), \text{rank}\left(X^{(i)}\right)\right]. \tag{14}$$

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.

As the numerical properties of $\mathbf{N}^{(i)}$ reflects the identifiability of the relevant fuzzy rule. By taking account the identifiability of a fuzzy rule into an initial model base construction is an effective weapon in overcoming the curse of dimensionality, as the model size can be automatically reduced by the number of data points, but not exponentially increasing with input dimension, if the rule-base with low identifiability (due to lack of data excitation) are excluded from the complete rule-base.

TABLE IV
SYSTEM ERROR REDUCTION RATIO BY THE SELECTED RULES IN EXAMPLE 2

Selected Rules	$[ERR]_i(t)$
Horse Power is Small	0.8918
Horse Power is Medium	0.0900
Horse Power is Large	0.0045
(Horse Power is Small) And (Year is Late)	0.0018
(Weight is Small) And (Year is Late)	0.0011
(Horse power is Medium) And (Year is Medium)	0.0005

The A-optimality design criteria for the weighting matrix $\mathbf{N}^{(i)}$ which is given by [16]

$$J_A\left(\mathbf{N}^{(i)}\right) = \frac{1}{N}\sum_{t=1}^{N} N_i(t) \tag{15}$$

provides an indication for each fuzzy rule on its identifiability.

Alternatively, consider the neurofuzzy system given by (4), and assumes that each submodel (rule) f_i is independent, the

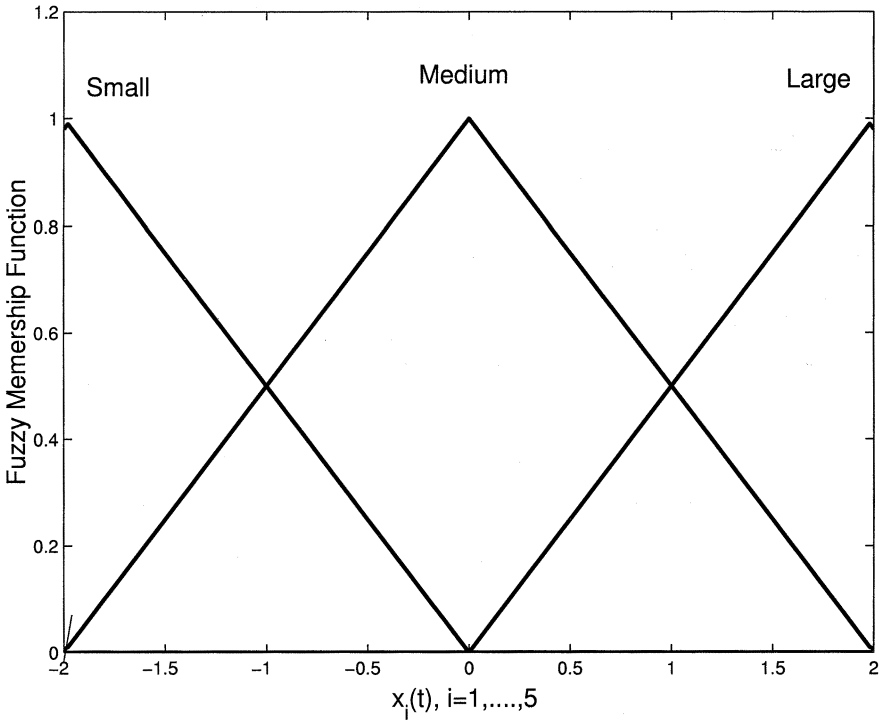

Fig. 7. Univariate fuzzy membership functions for all inputs in Example 3.

system output sensitivity of the composite model output $\hat{f}(\mathbf{x})$ with respect to associated submodel (rule) $\hat{f}_i(\mathbf{x})$ is given by

$$\frac{\delta \hat{f}(\mathbf{x})}{\delta \hat{f}_i(\mathbf{x})} = N_i(t). \tag{16}$$

For a given data sample $\{\mathbf{x}(t), y(t)\}$, $t = 1, 2, \ldots, N$, the average model output sensitivities to submodels (representing fuzzy rules), given by

$$\overline{\frac{\delta \hat{y}(\mathbf{x})}{\delta \hat{y}_j(\mathbf{x})}} = \frac{1}{N}\sum_{t=1}^{N} N_i(t) \tag{17}$$

which by (15) provides a metric for selecting appropriate model rules. The derived model rules can then be rearranged in descending order of average output sensitivity, followed by utilizing only the first K experts with greatest average sensitivity to construct a model rule-base set.

Notes: i) Comparing (15) and (17), it shows that the A-optimality design criteria of the weighting matrix $\mathbf{N}^{(i)}$ measures the average model output sensitivity over the input data set. Because parameters for a rule with near zero value of (17) (due to lack of data excitation) cannot be reliably estimated, the A-optimality design criteria for the weighting matrix $\mathbf{N}^{(i)}$ can be interpreted as the fuzzy rule identifiability.

ii) Note that for this rule, its input vector $\mathbf{x}^{(i)}$ is simply a subvector of the input vector $\mathbf{x}(t)$, that is, its input vector is a n_i-dimensional subset within the n-dimensional input space, called fuzzy partitioned input space [as spanned $\mathbf{x}^{(i)}(t), t = 1, \ldots, N$ within the n-dimensional input space]. Usually each subsystem (fuzzy rule) has a defined specific operating region depending on a subset of input variables, partitioning a specific operating region within the whole input space. Consequently, each subsystem usually is customized as a smaller sized model using a subset of input variables of those contained in $\mathbf{x}$, subsequently a much smaller network can be constructed to overcome the curse of dimensionality as well as provide model transparency.

TABLE V
ILLUSTRATION OF SOME FUZZY RULES IDENTIFIABLITIES IN EXAMPLE 3. (a) RULES ABOUT $y(t-1)$. (b) RULES ABOUT $y(t-1)$ AND $y(t-2)$. (THE BRACKET INDICATES RULES REMOVED FROM THE RULE BASE DUE TO LOW IDENTIFIABLITIES SHOWN BY THE A-OPTIMALITY CRITERIA)

Rules $(y(t-1))$	Small	Medium	Large
$\frac{1}{N}\sum_{t=1}^{N} N_i(t)$	0.1589	0.6937	0.1414

(a)

$\frac{1}{N}\sum_{t=1}^{N} N_i(t)$		Rules $(y(t-1))$		
		Small	Medium	Large
Rules	Small	0.0234	0.1187	0.0164
	Medium	0.1221	0.4695	0.1025
$(y(t-2))$	Large	(0.0135)	0.1054	0.0225

(b)

B. New Extended Gram–Schmidt Orthogonal Decomposition Algorithm

Analogous to conventional two-stage learning procedures for associative neural networks such as RBF model construction [10], the proposed rule-based model construction approach also consists of two stage learning, but is rule-based. This section introduces the second stage of fine model structure detection that has a model transparency property during learning. The construction of the initial rule-base introduced in previous sec-

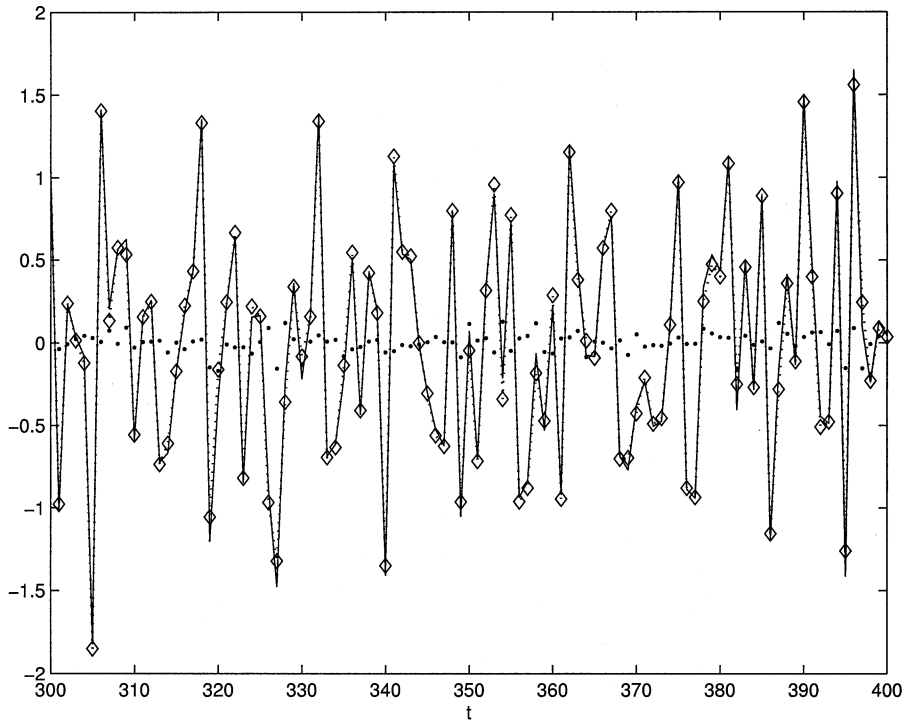

Fig. 8. Results of example 3; solid line: actual system output; diamond: model predictions; and dotted line: model residual.

tion is an unsupervised procedure without utilizing information from the system output, nor correlations between these derived rules. From a model with a structure of the initial rule-base, the system output information can then be utilized as model identification including parameter estimation that optimizes the model in its capacity to capture the system dynamics. Usually, if these fuzzy rules coexist in the model they are competitive such that some rules can become insignificant if some other rules are already in the model. Therefore in the second stage the model structure should be minimized but still maintains its modeling capability. Parsimonious model construction process is also a natural logical rule extraction process that increase model transparency simultaneously, because simpler models involve less rules and are easier to interpret. For general linear in the parameter systems, an OLS algorithm based on Gram–Schmidt orthogonal decomposition can be used to determine the model structure, its significant terms and associated parameter estimates [10]. Note that the forward OLS is inherently transparent in retrieving the energy levels associated with the selected regressors. One drawback of most current learning algorithms including the direct application of OLS to neurofuzzy systems is that the learning is based upon a set of basis functions and not upon a set of fuzzy rules, resulting in obscuring model transparency during the learning stage. Model transparency during the modeling process can be enhanced if the learning is rule-based. In the following a new extended Gram–Schmidt orthogonal decomposition algorithm is introduced that extends variable regression to subspace regression, which corresponds to fuzzy rule-base regressional construction due to the fuzzy rules functional subspace inference relationship. By exploring the one to one mapping between a fuzzy rule-base and a matrix feature subspace, the projection of the output vector onto a rule-base matrix subspace can be computed so as to enhance model transparency with the capability of interpreting the rule-base energy level. The significance of a new rule to an existing model can be effectively measured and extracted as rule-based knowledge; this is a direct extension of forward OLS algorithm that projects the output vector onto a regressor (basis function) so that any new regressor (basis function) significance can be readily evaluated relative the existing model basis, by modifying the conventional one dimensional regressor (basis

TABLE VI
SYSTEM ERROR REDUCTION RATIO BY THE SELECTED RULES IN EXAMPLE 3

Selected Rules	$[ERR]_i(t)$
($y(t-2)$ is Medium) And ($y(t-3)$ is Medium)	0.9275
($y(t-1)$ is Small) And ($y(t-3)$ is Small)	0.0352
($y(t-1)$ is Large) And ($y(t-3)$ is Small)	0.0202
$u(t-2)$ is Medium	0.0023
($y(t-3)$ is Medium) And ($u(t-2)$ is Medium)	0.0015
($y(t-2)$ is Small) and ($u(t-2)$ is Medium)	0.0016
$y(t-1)$ is Medium	0.0010

function) into submatrix spanned by linguistic fuzzy rule [$\Phi^{(i)}$ in (11)].

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 $\mathcal{W}^{(i)}$, $i = 1, \ldots, K$, given as [19], [20]

$$\mathcal{W}^{(1)} \oplus \cdots \oplus \mathcal{W}^{(K)} = S \in \Re^{p\times N} \tag{18}$$

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 $\mathbf{w}_i^{(i)} \subset \mathcal{W}^{(i)}$, $i = 1, \ldots, n_i$, then the matrix $\mathbf{W}^{(i)} = [\mathbf{w}_1^{(i)}, \ldots, \mathbf{w}_{n_i}^{(i)}]$, forms a basis of $\mathcal{W}^{(i)}$. Note that these n_i vectors need not to be mutually orthogonal, i.e., $[\mathbf{W}^{(i)}]^T\mathbf{W}^{(i)} = D^{(i)} \in \Re^{n_i\times n_i}$, where $D^{(i)}$ 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

$$\left[\mathbf{W}^{(i)}\right]^T \mathbf{W}^{(j)} = \mathbf{0}_{n_i\times n_j} \tag{19}$$

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 $\mathcal{W}^{(i)}$, $i = 1, \ldots, K$, are defined by a series of K matrices $\mathbf{W}^{(i)}$, $i = 1, \ldots, K$ as subspace basis based on Definition 1, then an arbitrary vector $\hat{\mathbf{y}} \in \Re^N \in S$ can be uniquely decomposed as

$$\hat{\mathbf{y}} = \sum_{i=1}^{K}\sum_{j=1}^{n_i} c_{i,j}\mathbf{w}_j^{(i)} \tag{20}$$

where $c_{i,j}$s are combination coefficients.

As the result of the orthogonality of $[\mathbf{w}^{(i)}]^T\mathbf{w}^{(j)} = 0$, (for $i \neq j$), from (20)

$$\|\hat{\mathbf{y}}\|^2 = \sum_{i=1}^{K}\left\|\sum_{j=1}^{n_i} c_{i,j}\mathbf{w}_j^{(i)}\right\|^2. \tag{21}$$

Clearly, the variance of the vector $\hat{\mathbf{y}}$ projected into each subspace can be computed as $\|\sum_{j=1}^{n_i} c_{i,j}\mathbf{w}_j^{(i)}\|^2$, for $i = 1, \ldots, K$.

Consider the nonlinear system (1) given as a vector form by (11). By introducing an orthogonal subspace decomposition $\mathbf{\Phi} = \mathbf{WA}$, (11) can be written as

$$\begin{aligned}\mathbf{y} &= \mathbf{Wc} + \mathbf{e}\\ &= \sum_{i=1}^{K}\mathbf{W}^{(i)}\mathbf{c}_i + \mathbf{e}\end{aligned} \tag{22}$$

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

$$\mathbf{A} = \begin{bmatrix} A_{1,1} & A_{1,2} & \cdots & A_{1,K} \\ 0 & A_{2,2} & \cdots & A_{2,K} \\ & \cdots & \cdots & \\ 0 & \cdots & A_{i,j} & \cdots \\ & \cdots & \cdots & \\ 0 & \cdots & \cdots & A_{K,K} \end{bmatrix} \in \Re^{p\times p} \tag{23}$$

in which $A_{i,j} \in \Re^{n_i\times n_j}$.

The extended Gram–Schmidt orthogonal decomposition algorithm is as follows.

Set $\mathbf{W}^{(1)} = \Phi^{(1)}$, $A_{1,1} = \mathbf{I}_{n_1\times n_1}$, and, for $j = 2, \ldots, K$, set $A_{j,j} = \mathbf{I}_{n_j\times n_j}$

$$\mathbf{W}^{(j)} = \Phi^{(j)} - \sum_{i=1}^{j-1}\mathbf{W}^{(i)} * A_{i,j} \tag{24}$$

where

$$A_{i,j} = \left[\left[\mathbf{W}^{(i)}\right]^T\mathbf{W}^{(i)}\right]^{-1}\left[\mathbf{W}^{(i)}\right]^T\Phi^{(j)} \in \Re^{n_i\times n_j} \tag{25}$$

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

Note $D^{(i)} = [\mathbf{W}^{(i)}]^T\mathbf{W}^{(i)}$, the least squares solution of (22) is given by

$$\mathbf{c}_i = \left[D^{(i)}\right]^{-1}\left[\mathbf{W}^{(i)}\right]^T\mathbf{y} \tag{26}$$

which follows from the fact that $\mathbf{W}^{(i)}$, $i = 1, \ldots, K$ are mutually orthogonal subspaces basis.

From (20), if the system output vector $\mathbf{y}$ is decomposed as a term $\hat{\mathbf{y}}$ by projecting onto orthogonal subspaces $\mathbf{W}^{(i)}$, $i = 1, \ldots, K$, 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 toward the construction of $\mathbf{y}$) can be readily calculated via

$$[\mathrm{ERR}]_i = \frac{\left\|\sum_{j=1}^{n_i} c_{i,j}\mathbf{w}_j^{(i)}\right\|^2}{\|\mathbf{y}\|^2}. \tag{27}$$

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 $[\mathrm{ERR}]_i$ 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 toward 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, as shown in Fig. 2, with the rule basis with the largest $[ERR]_i$ to be included in the final model until

$$1 - \sum_{i=1}^{n_f}[\mathrm{ERR}]_i < \rho \tag{28}$$

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

Set $\mathbf{\Theta}_{n_f} = \mathbf{c}_{n_f}$, and, for $i = n_f - 1, \ldots, 1$

$$\mathbf{\Theta}_i = \mathbf{c}_i - \sum_{j=i+1}^{n_f} A_{i,j} * \mathbf{\Theta}_j. \tag{29}$$

Notes: iii) 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. 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.

iv) For a high-dimension p, the proposed algorithm decomposes the system into a few (n_f) submodels rather than a single model with a significant large number of $\sum_{i=1}^{n_f} n_i$ orthogonal basis if conventional Gram–Schmidt method were applied. This is computationally simpler because each orthogonal subspace basis $\mathbf{W}^{(i)}$ is not required to be a diagonal matrix. The algorithm is also useful in many signal processing applications whereby it is often more desirable to decompose single model into a few submodels.

IV. Numerical Examples

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

$$y(x) = (\sin(\pi x) + 1)x. \tag{30}$$

500 data pairs $\{x, y\}$ are generated via (30), 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 Fig. 3, correspond 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 (30), the neurofuzzy model is simply given as

$$\hat{y}(t) = \sum_{i=1}^{6} N_i(x(t))x(t)\theta_i \tag{31}$$

where t denotes the data label, with each of the fuzzy rule $\Phi^{(i)} = N_i(x(t))x(t)$ spanning a one dimensional space, i.e., $n_i = 1$, $\forall i$. The identifiability of these fuzzy rules are computed based on (15) [or, equivalently, (17)] 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 forward selection procedure 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). Each rule contribution in reducing model error (or increasing the model energy level) provides model transparency for the fuzzy rules interpretability. The modeling results are given in Fig. 4, in which, the mean squares error (MSE) by using the model [Fig. 4(c)] with all six basis as 3.24×10^{-5}. If rule 1 is excluded, the model predicted output is shown in Fig. 4(c), with an MSE is 2.64×10^{-4}, demonstrating excellent approximation. Clearly the proposed modeling approach is additionally advantageous via its significant model transparency during the modeling process.

Example 2: Automobile Miles Per Gallon (MPG) Data: This data concerns city cycle fuel consumption in MPG[1] and its potential causal relation to various observed inputs. The original data set of 398 data points contains 392 complete data points. There are six inputs of various manufacturers cars; the number of cylinders, displacement, horsepower, weight, acceleration, and model year. In a previous study by ISIS [18], it has been shown that three inputs (horse power, weight and model year) are significant in modeling MPG. These three inputs are used in the initial rule-base construction for this study. By predetermining knot vectors for each of these variables over their data range, and using a piecewise linear B-spline fuzzy membership function, three univariate fuzzy membership functions are generated, as shown in Fig. 5. Then, the ANOVA approach is used to construct fuzzy rule-bases based on the interaction of these membership functions. The univariate and bivariate membership functions (interaction between any two univariate membership function via tensor product) used, are shown in Table III, in which, the identifiability of fuzzy rules are listed based on (15) [or, equivalently, (17)]. From Table III, it is seen that some of the rules have poor identifiability due to lack of data excitation. After removing these redundant rules from the rule-base, there are 31 initial rules.

By using the fuzzy model (4) for the modeling of MPG data, the neurofuzzy model is simply given as

$$\hat{y}(t) = \sum_{i=1}^{31} N_i(\mathbf{x}(t))\mathbf{x}(t)\mathbf{\Theta}_i \tag{32}$$

where t denotes the data label, and $\mathbf{x}(t)$ is given by the data values of [horsepower, weight, year]. Hence, each of the fuzzy rule $\Phi^{(i)} = N_i(\mathbf{x}(t))\mathbf{x}(t)$ spans a three-dimensional space, i.e., $n_i = 3, \forall i$. The extended Gram–Schmidt orthogonal decomposition algorithm decomposes each rule subspace being spanned by a three-dimensional rule basis into orthogonal matrix subspaces. The forward selection procedure 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 IV. In Table IV, the selected rules are ordered in the sequence of being selected, together with each rules contribution in reducing model error (or increasing the model energy level),

[1]Available online at ftp.ics.uci.edu/pub/machine-learning-databases

providing natural model transparency via the fuzzy rules. The final model has six fuzzy rules produces a MSE of 6.31; its model prediction result is shown in Fig. 6. It is clear that the proposed modeling approach can reveal significant model transparency during the modeling process, also it can construct a parsimonious set of rules with excellent model transparency and excellent approximation capability.

Example 3: A nonlinear dynamic system. Consider the following benchmark dynamic system given by [21], [22]:

$$y(t) = \frac{y(t-1)y(t-2)y(t-3)u(t-2)[y(t-3)-1]+u(t-1)}{1+y^2(t-2)+y^2(t-3)} \tag{33}$$

where the system input $u(t)$ is given as a uniformly distributed random signal in the range $[-2, 2]$. 500 data points were generated. The input vector is predetermined as a five-input vector as $\mathbf{x}(t) = [y(t-1), y(t-2), y(t-3), u(t-1), u(t-2)]^T$. According to the data distribution and using apiecewise linear B-spline fuzzy membership function, three univariate fuzzy membership functions are generated based on a knot vector $[-3, -2, 0, 2, 3]$ for all of the five inputs, as shown in Fig. 7. The ANOVA approach is used to construct fuzzy rule-bases based on the interaction of these membership functions. The univariate and bivariate membership functions (interaction between any two different univariate membership function via tensor product) are used. The identifiability of fuzzy rules are computed based on (15) [or equivalently (17)]. Although 5 univariate input variables produce 5×3 univariate rule-bases and 10×9 bivariate bases, for simplicity of presentation, an example of univariate rule-bases (based on an input with three rules) and that of a bivariate rule-bases (based on two inputs with nine rules) are shown in Table V. From Table V, some of the rules are of poor identifiability due to lack of data excitation and are removed from the rule-base, to give an initial rule-base of 102 rules.

By using the fuzzy model (4) for the modeling this nonlinear dynamical system, the neurofuzzy model is simply given as

$$\hat{y}(t) = \sum_{i=1}^{102} N_i(\mathbf{x}(t))\mathbf{x}(t)\mathbf{\Theta}_i \tag{34}$$

where t denotes the time index. Hence, each of the fuzzy rule $\Phi^{(i)} = N_i(\mathbf{x}(t))\mathbf{x}(t)$ spans a five-dimensional space, i.e., $n_i = 5, \forall i$. The extended Gram–Schmidt orthogonal decomposition algorithm decomposes each rule subspace being spanned by a five-dimensional rule basis into orthogonal matrix subspaces. The forward selection procedure 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 VI. In Table VI, the selected rules which are ordered in the selection sequence, and their contribution in reducing model error (or increasing the model energy level), and provide appropriate model transparency for the derived fuzzy rules set. The model with six fuzzy rules explains 99% of system output variance and produces a MSE of 0.006. The model prediction results are shown in Fig. 8 (for visual clarity only a section of data typical of the remaining data is shown). Again the results based on this nonlinear system demonstrate that the proposed approach has significant model transparency during the modeling process, as well as the capability to construct a parsimonious rule-based system with excellent approximation capability.

V. Conclusion

This paper has introduced a new neurofuzzy construction and parameter estimation algorithm for the modeling of *a priori* unknown dynamical system from data. Based on a T–S inference mechanism, the first contribution of this paper is the introduction of a one to one mapping between a fuzzy rule-base and a matrix feature subspace. The second contribution of the paper is the introduction of an extended Gram–Schmidt algorithm that decomposes the model rule-bases via an orthogonal subspace decomposition approach. In consequence, the proposed approach greatly has enhanced the model transparency during the modeling process as well as in the resultant model. Numerical examples have demonstrated the effectiveness of the new algorithm. The proposed algorithm is applicable to a wide range of signal processing problems for both dynamic and nondynamic nonlinear processes where there is a requirement for the evolution of an explanatory fuzzy rule-base and a final parsimonious rule-base that offers high approximation and good generalization capability.

References

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

[2] M. Brown and C. J. Harris, *Neurofuzzy Adaptive Modeling and Control*. Upper Saddle River, NJ: Prentice-Hall, 1994.

[3] K. M. Bossley, "Neurofuzzy modeling approaches in system identification," Ph.D. dissertation, Dept. ECS, Univ. Southampton, Southampton, U.K., 1997.

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

[5] T. Takagi and M. Sugeno, "Fuzzy identification of systems and its applications to modeling and control," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-15, pp. 116–132, Jan. 1985.

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

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

[8] R. Bellman, *Adaptive Control Processes*. Princeton, NJ: Princeton Univ. 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 nonlinear system identification," *Int. J. Control*, vol. 50, pp. 1873–1896, 1989.

[11] X. Hong and C. J. Harris, "Nonlinear model structure detection using optimum experimental design and orthogonal least squares," *IEEE Trans. Neural Networks*, vol. 12, pp. 435–439, Apr. 2001.

[12] X. Hong and C. J Harris, "Neurofuzzy design and model construction of nonlinear dynamical processes from data," *Proc. Inst. Electr. Eng.—Control Theory and Applications*, vol. 148, no. 6, pp. 530–538, 2001.

[13] T. Taniguchi, K. Tanaka, H. Ohtake, and H. O. Wang, "Model construction, rule reduction and rubust compensation for generalized form of Takagi–Sugeno fuzzy systems," *IEEE Trans. Fuzzy Syst.*, vol. 9, pp. 525–538, Aug. 2001.

[14] Y. Jin, "Fuzzy modeling of high dimensional systems: Complexity reduction and interpretability improvement," *IEEE Trans. Fuzzy Syst.*, vol. 8, pp. 212–221, Feb. 2000.

[15] H. Roubos and M. Setnes, "Compact and transparent fuzzy models and classifiers through iterative complexity reduction," *IEEE Trans. Fuzzy Syst.*, vol. 9, pp. 516–524, Aug. 2001.
[16] A. C. Atkinson and A. N. Donev, *Optimum Experimental Designs*. Oxford, U.K.: Clarendon, 1992.
[17] P. Dierckx, *Curve and Surface Fitting with Splines*. Oxford, U.K.: Clarendon, 1995.
[18] S. R. Gunn, M. Brown, and K. Bossley, "Network performance assessment for neurofuzzy data modeling," *Intell. Data Anal.*, pp. 313–323, 1997.
[19] K. W. Gruenberg and A. J. Weir, *Linear Geometry*. New York: Van Nostrand, 1967.
[20] T. Soderström and P. Stoica, *System Identification*. Upper Saddle River, NJ: Prentice-Hall, 1989.
[21] K. S. Narendra and K. Parthasarathy, "Identification and control of dynamic systems using neural networks," *IEEE Trans. Neural Networks*, vol. 1, pp. 4–27, Feb. 1990.
[22] J. H. Nie and T. H. Lee, "Rule-based modeling: Fast construction and optimal manipulation," *IEEE Trans. Syst., Man, Cybern.*, vol. 26, pp. 728–738, Nov. 1996.

Xia Hong (SM'02) received the B.Sc. and M.Sc. degrees from the National University of Defense Technology, P.R. China, and the Ph.D. degree from the University of Sheffield, Sheffield, U.K., in 1984, 1987, and 1998, respectively, all in automatic control.

She worked as a Research Assistant with the Beijing Institute of Systems Engineering, Beijing, China from 1987 to 1993. She worked as a Research Fellow in the Department of Electronics and Computer Science, the University of Southampton from 1997 to 2001. She is currently a lecturer with the Department of Cybernetics, University of Reading, Reading, U.K. She is actively engaged in research into neurofuzzy systems, data modeling, and learning theory and their applications. Her research interests include system identification, estimation, neural networks, intelligent data modeling, and control. She has published over 30 research papers and coauthored a research book.

Dr. Hong was awarded a Donald Julius Groen Prize (Best Paper in Control) by IMechE in 1999.

Chris J. Harris received the B.Sc. degree from the University of Leicester, Leicester, U.K., in 1967, the M.A. degree from Oxford University, Oxford, U.K., in 1972, and the Ph.D. and D.Sc. degrees from the University of Southampton, Southampton, U.K., in 1976 and 2002, respectively.

He previously held appointments at the Universities of Hull, UMIST, Oxford, and Cranfield, as well as being employed by the U.K. 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, multisensor data fusion, and systems integration. He has authored or coauthored 12 books and over 300 research papers.

Dr. Harris is the Associate Editor of numerous international journals. 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.