FITSK: Online Local Learning With Generic Fuzzy Input Takagi–Sugeno–Kang Fuzzy Framework for Nonlinear System Estimation

Kian Hong Quah and Chai Quek, *Member, IEEE*

***Abstract*—Existing Takagi–Sugeno–Kang (TSK) fuzzy models proposed in the literature attempt to optimize the global learning accuracy as well as to maintain the interpretability of the local models. Most of the proposed methods suffer from the use of offline learning algorithms to globally optimize this multi-criteria problem. Despite the ability to reach an optimal solution in terms of accuracy and interpretability, these offline methods are not suitably applicable to learning in adaptive or incremental systems. Furthermore, most of the learning methods in TSK-model are susceptible to the limitation of the curse-of-dimensionality. This paper attempts to study the criteria in the design of TSK-models. They are: 1) the interpretability of the local model; 2) the global accuracy; and 3) the system dimensionality issues. A generic framework is proposed to handle the different scenarios in this design problem. The framework is termed the generic fuzzy input Takagi–Sugeno–Kang fuzzy framework (FITSK). The FITSK framework is extensible to both the zero-order and the first-order FITSK models. A zero-order FITSK model is suitable for the learning of adaptive system, and the bias-variance of the system can be easily controlled through the degree of localization. On the other hand, a first-order FITSK model is able to achieve higher learning accuracy for nonlinear system estimation. A localized version of recursive least-squares algorithm is proposed for the parameter tuning of the first-order FITSK model. The local recursive least-squares is able to achieve a balance between interpretability and learning accuracy of a system, and possesses greater immunity to the curse-of-dimensionality. The learning algorithms for the FITSK models are online, and are readily applicable to adaptive system with fast convergence speed. Finally, a proposed guideline is discussed to handle the model selection of different FITSK models to tackle the multi-criteria design problem of applying the TSK-model. Extensive simulations were conducted using the proposed FITSK models and their learning algorithms; their performances are encouraging when benchmarked against other popular fuzzy systems.**

***Index Terms*—Degree of localization, localized learning, nonlinear system estimation, Takagi–Sugeno–Kang fuzzy models, zero and first-order TSK models.**

I. Introduction

THE Takagi–Sugeno–Kang (TSK) fuzzy model [1]–[3] is a class of fuzzy models that assumes local model representations with local function dynamics at the consequent or rule-layer of the models. The idea of such a fuzzy model is to consider the output by performing fuzzy interpolations of these simpler local functional models in the neighboring fuzzy partitions. The main advantage of the TSK-model over other classes of fuzzy models is its ability to model a system accurately; either globally or locally. The accurate global learning ability motivates the practical applications of TSK-model in nonlinear system estimation [4]. The local learning ability provides a course of interpretability of the local models in the localized subspaces [5]–[8].

Existing TSK-model learning algorithms proposed in the literature can be classified using two criteria. The first criterion is essentially the locality of learning. The second criterion is being an online or an offline learning algorithm. The locality of learning depends on the model's learning objective function, which is a minimization problem of the global or the local learning errors. The global parameter tuning considers the minimization of the global error of the model, such as the ANFIS's Kalman filter algorithm [9]. This benefits the global accuracy of the system, but suffers from degradation in local interpretability. This degradation exhibits an erratic local behavior, which causes the local models to have difficulty in interpretability [5]. In contrast, the local parameter tuning takes the form of the minimization problem of the local learning errors. This ensures that the local models possess accurate representations and improves on the local interpretability of the system. Examples of such local parameter tuning are the direct update of zero-order TSK model and the local learning algorithm in Yen's model [5]. However, minimization of local learning errors without consideration of global behavior results in degradation of global learning accuracy.

An offline learning algorithm assumes that the data is presented in a batch form and can be repeatedly accessed. This form of learning is easy to guarantee its success in reaching an optimal solution based on its learning objective function. Furthermore, it has the flexibility in recalling the stored training examples to improve the quality of learning. However, the offline learning is inapplicable to adaptive or incremental learning systems. Moreover, it is prone to the *curse-of-dimensionality* problem [10]. Examples of the offline learning methods are the singular value decomposition approach [5] and the constraint-optimization method [8]. Summarizing these, the existing TSK-models encountered one or more of the following major problems despite being good modeling tools for nonlinear system estimation. They are; namely: 1) offline learning; 2) low interpretability of local model; 3) degradation in global learning accuracy; and 4) inapplicable to higher dimensionality system.

Manuscript received November 18, 2004; revised May 23, 2005. This paper was recommended by Associate Editor Hideyuki Takagi.

The authors are with the Centre for Computational Intelligence, School of Computer Engineering, Nanyang Technological University, Singapore 639798 (e-mail: ashcquek@ntu.edu.sg).

Digital Object Identifier 10.1109/TSMCB.2005.856715

Hence, a novel TSK-model framework that is immune to the above deficiencies is proposed in this paper. The framework is termed *the generic fuzzy input Takagi-Sugeno-Kang fuzzy framework* (FITSK). The FITSK framework can be readily extended to different classes of TSK-models to alleviate the above deficiencies based on different design criteria. All the learning algorithms of FITSK assume online learning paradigms with fast convergence speed.

This paper is organized as follows. Section II describes the general structure of the FITSK framework, the zero-order and the first-order FITSK models, and their online learning algorithms. Section III presents the simulation results of the FITSK models with five different nonlinear system estimation tasks. Section IV concludes this paper.

II. FITSK

This section introduces the generic FITSK framework; the FITSK framework takes the form of a superset of all the TSK fuzzy models [1]–[3]. A special case of the FITSK framework is mapped to a zero-order FITSK model (denoted by FITSK^0) on the assumption that the local models' function dynamics are zero-order (constant term). Subsequently, a degree of localization of the zero-order FITSK model is introduced to balance the bias-variance of the learning of the FITSK model. A first-order FITSK model (denoted by FITSK^1) is another adaptation of the FITSK framework, where the local models are first-order linear functions of the input vector. There are two classes of parameter learning for the FITSK^1 model. They are the global recursive least-squares algorithm and the local recursive least-squares algorithm. The latter is computationally more effective than the former. A degree of localization is introduced in the local recursive least-squares algorithm to control the interpretability of the FITSK^1 model. A guideline is proposed to select the choice of different FITSK models with respect to the following three design criteria; namely 1) model interpretability; 2) learning accuracy; and 3) data dimensionality. The guideline allows the choice of the FITSK models with different settings to generate a balanced system on the basis of these criteria.

A. Generic Structure of FITSK

Fig. 1 depicts the generic structure of FITSK. The FITSK framework consists of six layers of nodes. The six layers are the input fuzzifier $(\mathbf{IF})$, the input linguistic $(\mathbf{IL})$, the conjunction $(\mathbf{C})$, the normalization $(\mathbf{N})$, the functional or rule $(\mathbf{F})$, and the summation $(\mathbf{S})$ layers. The input fuzzifier nodes in the first layer fuzzify the input vector into a corresponding vector of fuzzy membership functions. Subsequently, the input linguistic layer measures the similarity of each fuzzy membership functions with its corresponding linguistic nodes. The conjunction layer computes the conjunction of all incoming signals and generates the corresponding weights. The normalization layer normalizes these weights. Each node of the functional or rule layer implements a local model, which is a form of the TSK rule. The outputs of the functional layer are weighted with their incoming normalized weights. Finally, the summation layer summarizes all the incoming signals. Therefore, the outputs of the FITSK are computed by interpolating the local models with the normalized fuzzy measures of the input vector. This paragraph briefly summarizes the operations of the generic FITSK framework; subsequent segment of this section formularizes these mathematically.

Each input fuzzifier node $\mathbf{IF}_i$, $i \in \{1, 2, \ldots, I\}$, has a single input. Vector $X = [x_1, \ldots, x_i, \ldots, x_I]^T$ represents the inputs to the FITSK. Each output node $\mathbf{S}_n$, where $n \in \{1, 2, \ldots, N\}$ computes a single output denoted by y_n. Vector $Y = [y_1, \ldots, y_n, \ldots, y_N]^T$ denotes the outputs of the FITSK with respect to the input vector X. The variables i, j, k, l, m, and n are used to refer to arbitrary nodes in layers 1, 2, 3, 4, 5, and 6, respectively.

The total number of nodes for layers 1, 3, 4, and 6 are denoted with variables I, K, L, and N respectively. Each input fuzzifier node $\mathbf{IF}_i$ may have different number of input linguistic $(\mathbf{IL})$ nodes J_i. Hence, the total number of nodes for layer 2 is $\sum_{i=1}^{I} J_i$. Layer 3 consists of conjunction nodes $\mathbf{C}_k$, where $k \in \{1, 2, \ldots, K\}$. Each layer 3 node $\mathbf{C}_k$ is directly connected to a layer 4 normalization node $\mathbf{N}_l$. Hence, the total number of nodes in layers 3 and 4 are identical, $K = L$. Each layer 4 node $\mathbf{N}_l$ is associated with M number of layer 5 functional nodes, $\mathbf{F}_{l,m}$. M is identical to the number of layer 6 nodes, $M = N$. Hence, the total number of nodes in layer 5 is $L \times M$. Each output node at layer 6 is a summation node, $\mathbf{S}_n$. Each summation node is connected with L number of layer 5 functional nodes. The total number of nodes at layer 6, N, depends on the dimension of output vector Y. The FITSK model adopts the TSK's fuzzy model and the trainable parameters are found in layer 5 of the model.

The output of a node is denoted as Z with the subscripts specifying its origin, for example Z_{IF_i} is the output of node $\mathbf{IF}_i$. All the outputs of a layer are propagated to the inputs of the connecting nodes at the next layer with unity link-weight.

The generic operations of the FITSK can be defined as follows.

Layer 1 (Input fuzzifier layer)

$$Z_{\mathrm{IF}_i} = \mu_i(x_i) \tag{2.1}$$

where $\mu_i(.)$ is the fuzzy membership function of input fuzzifier node $\mathbf{IF}_i$

Layer 1 nodes fuzzify a singleton input by adopting the input x_i as the center of the fuzzy membership function, such that the membership function achieves maximum value at x_i and decreases while moving away from the center. However, when a nonsingleton (fuzzy) input is presented, the node simply directs the nonsingleton input as the output of the node.

Layer 2 (Input Linguistic layer)

$$Z_{\mathrm{IL}_{i,j}} = \mu_{i,j}\left(Z_{\mathrm{IF}_i}\right) = \mu_{i,j}\left(\mu_i(x_i)\right) \tag{2.2}$$

where $\mu_{i,j}(.)$ is a fuzzy linguistic label of node IF_i.

Layer 2 nodes take in the fuzzy membership function of layer 1 outputs as the inputs, and compute a *fuzzy subsethood measure* [4] for each corresponding fuzzy linguistic label. The fuzzy subsethood measure defines the degree that fuzzy set A is a subset of fuzzy set B as

$$S(A, B) = \frac{|A \cap B|}{|A|}. \tag{2.3}$$

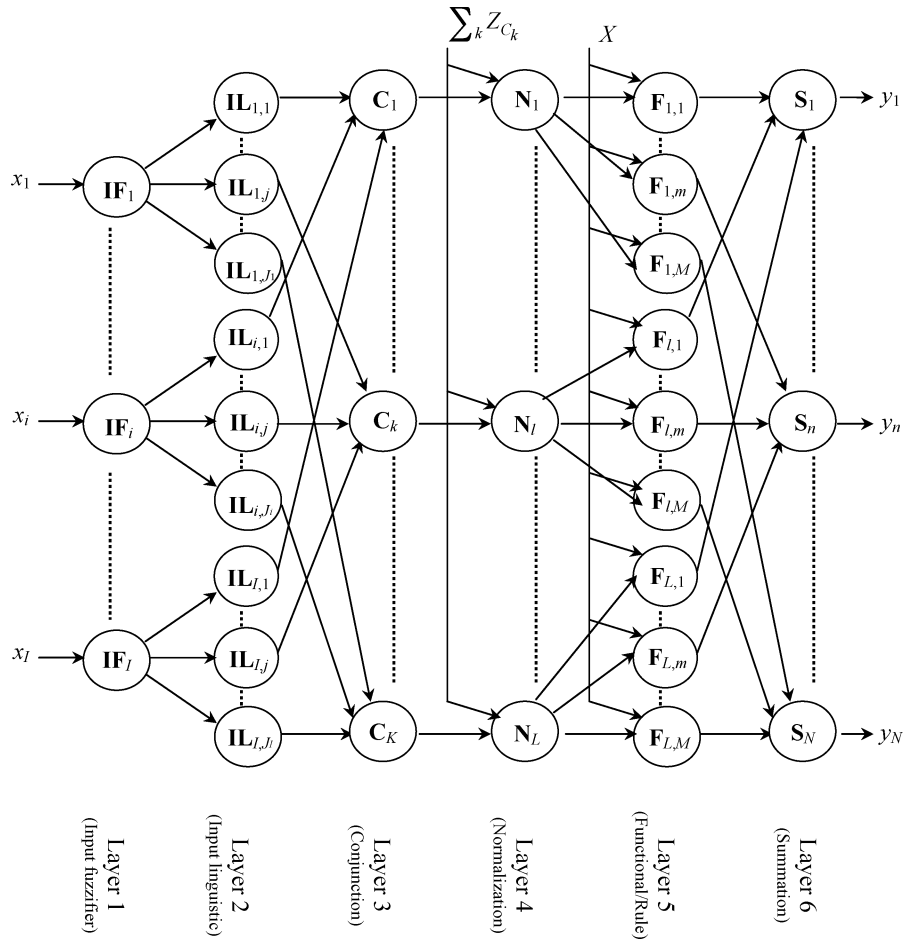

Fig. 1. Generic structure of FITSK.

If the min operator is used for intersection operation, then (2.3) is described in detail in (2.4)

$$S(A,B) = \frac{\int_{x\in U} \min\left(\mu_A(x), \mu_B(x)\right)}{\int_{x\in U} \mu_A(x)} \tag{2.4}$$

or can be approximated by (2.5)

$$S(A,B) = \frac{\max\limits_{x\in U}\left(\min\left(\mu_A(x), \mu_B(x)\right)\right)}{\max\limits_{x\in U}\left(\mu_A(x)\right)}. \tag{2.5}$$

Layer 3 (Conjunction layer):

$$Z_{\mathrm{C}_k} = \prod_{i=1}^{I} Z_{\mathrm{IL}_{i,j}}. \tag{2.6}$$

Layer 3 nodes compute the conjunction of all the incoming signals using a product or a minimum inference.

Layer 4 (Normalization layer)

$$Z_{\mathrm{N}_l} = \frac{Z_{\mathrm{C}_k}}{\sum_{k=1}^{K} Z_{\mathrm{C}_k}} = w_l \tag{2.7}$$

where w_l is the normalized weight.

Layer 5 (Functional/Rule layer):

$$Z_{\mathrm{F}_{l,m}} = Z_{\mathrm{N}_l} f_{\mathrm{F}_{l,m}}(X) = w_l f_{\mathrm{F}_{l,m}}(X) \tag{2.8}$$

where $f_{\mathrm{F}_{l,m}}(X)$ is a local model function of node $\mathbf{F}_{l,m}$.

Layer 6 (Summation layer):

$$y_n = Z_{\mathrm{S}_n} = \sum_{l=1}^{L} Z_{\mathrm{F}_{l,m}}. \tag{2.9}$$

Two motivations drive the development of the FITSK model. The first is to craft a systematic and generic framework for an arbitrary order TSK model [1]–[3] with arbitrary parameter learning capability. The second is to map the generic model with a nonsingleton (fuzzy) input fuzzifier such that the *degree of localization* of the local model can be controlled with ease. A special case of the generic FITSK is the ANFIS model [9] which considers a singleton input fuzzifier at layer 1, a product inference scheme at layer 3, a first-order polynomial local model function at layer 5, and a single output at layer 6.

B. A Zero-Order FITSK Model

The zero-order FITSK model is a special case of the generic FITSK framework with zero-order polynomial functions (constant terms) at layer 5, and product inference in layer 3. The specific formulation at layer 1 and 2 are described as follows.

Layer 1 (Input fuzzifier layer): The choice of fuzzy membership function is flexible. In this formulation, a Gaussian membership function is chosen

$$Z_{\mathrm{IF}_i} = \mu_i(x_i) = e^{\frac{-(x-x_i)^2}{2\sigma_i^2}} \tag{2.10}$$

where

x spreads over the universe of possible range of x_i;

σ_i standard deviation of the Gaussian function.

The standard deviation σ_i reflects the width of the membership function, and is dependant on the accuracy and generalization to be achieved by a specific system. This is a *bias-variance dilemma* problem. Smaller value of the standard deviation (bias) implies higher focusing region for the local model in the local subspace, and directly proportions to higher learning accuracy. Conversely, a larger value (variance) anticipates higher noise or generalization capability in the neighboring subspaces.

Layer 2 (Input linguistic layer): Each linguistic node is a singleton label to simplify the computation of fuzzy subsethood measure. Let $s_{i,j}$ denotes the centre of the singleton of the j^{th} linguistic label of the i^{th} input node, $\mathbf{IL}_{i,j}$, with unity height. The computations involving layer 1 and layer 2 nodes can be simplified as

$$Z_{\text{IL}_{i,j}} = e^{\frac{-(s_{i,j}-x_i)^2}{2\sigma_i^2}}. \tag{2.11}$$

The system designer would have to specify the number of labels in each dimension. The location of $s_{i,j}$ is evenly distributed over the value range of the input dimension i. The resultant resolution r_i is computed as

$$r_i = \frac{\text{range}_i}{\text{no. of labels}_i - 1} \tag{2.12}$$

where range_i is the value range of the input dimension i and no. of labels_i is the total number of labels in dimension i.

Layer 5 (Functional layer)

$$\begin{aligned} f_{\text{F}_{l,m}}(X) &= c_{\text{F}_{l,m}} \\ Z_{\text{F}_{l,m}} &= w_l c_{\text{F}_{l,m}} \end{aligned} \tag{2.13}$$

where $c_{\text{F}_{l,m}} \in \Re$ is the zero-order function (constant term) of the local model in the node $\mathbf{F}_{l,m}$.

Let the vector $D = [d_1, \ldots, d_n, \ldots, d_N]^T$ denote the actual outputs of the FITSK with respect to the input vector X. The training assumes a direct adaptive learning process, which has more plasticity rather than stability. The learning process adapts to new information rapidly by forgoing past information. Thus, the zero-order FITSK model is effective in modeling nonlinear adaptive system such as time-series data or temporally adaptive system. The learning process for an arbitrary zero-order local model is listed as

$$\begin{aligned} \Delta c_{\text{F}_{l,m}}^t &= \alpha w_l \left(d_n^t - y_n^t\right) \\ c_{\text{F}_{l,m}}^{t+1} &= c_{\text{F}_{l,m}}^t + \Delta c_{\text{F}_{l,m}}^t \end{aligned} \tag{2.14}$$

where t is the time index and $\alpha \in (0, 1]$ is the positive step size learning constant.

If the stability of the learning system is crucial, the learning constant can be modified to a monotonically decreasing function to tighten the plasticity of the system.

The design of input fuzzifier with singleton linguistic formulation is driven by two motivations. The first is to exclude the need of online clustering process for the linguistic labels. The second motivation is to ease the control of the *degree of localization* in the neighboring subspaces.

Clustering process for the Mamdani's fuzzy model [11] is an unsupervised learning process by partitioning the input and output spaces into reasonable clusters [12]–[14]. However, unsupervised clustering process for TSK fuzzy model identification is unjustifiable because the quality of the cluster is dependant on the input data as well as the corresponding output dynamics of the system. Several clustering techniques [3], [15] have been proposed for TSK model identification in a supervised learning manner, in which the quality of the output is quantified with respect to the input clusters formation. However, they suffer from adopting a heuristic search approach and are computationally expensive [3], [15]. This demonstrates that the supervised clustering problem for identification of TSK fuzzy model structure is a combinatorial complex problem. As a result, the FITSK model presupposes an even space partitioning of the singleton linguistic labels instead of performing an effective online clustering technique. Normal TSK model maintains the learning accuracy by fuzzy interpolation of neighboring local models. The FITSK is able to maintain a similar fuzzy interpolation by having fuzzy input with singleton linguistic labels, without the need of having a clustering process to structurally define the optimal linguistic labels. This is because the fuzzification is performed prior to the linguistic layer, and the membership functions are well-defined in the sense that the centers of the membership functions represent the exact location of the input vector.

The standard deviation in (2.11) is defined as a function of the input value range in (2.15)

$$\sigma_i = sr^{ld} \times \frac{\text{range}_i}{5} \tag{2.15}$$

where

$sr \in \Re$ shrinking rate and $sr \in (0, 1]$;

range_i value range of the input dimension i;

ld degree of localization.

The maximum of the standard deviation is simply estimated as $\text{range}_i/5$ such that the membership function is able to contain the entire range of the input space. The minimum of the standard deviation is chosen with respect to the concept of ε-completeness. The level ε at the crossover point of neighboring fuzzy membership functions implies a strong belief in the positive sense of the fuzzy control rules, which are associated with the controller. The degree of belief is chosen to be greater than 0.5 [16], [17] to achieve this. Thus, the weight or matching degree of at least one of the label of an input dimension is selected to be greater than 0.5.

Equations (2.16) and (2.17) summarize the derivation of a minimum of the standard deviation, which is derived from (2.11)

$$\begin{aligned} 0.5 &\le e^{\frac{-(s_{i,j}-x_i)^2}{2\sigma_i^2}} \\ \Rightarrow \sigma_i &\ge 0.8493 \times (s_{i,j} - x_i). \end{aligned} \tag{2.16}$$

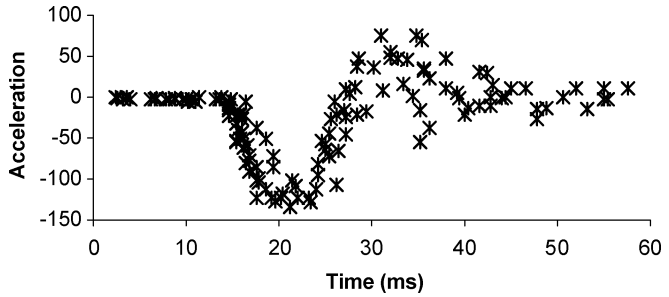

Fig. 2. Motorcycle crash dataset.

The boundary condition for $(s_{i,j} - x_i)$ appears when x_i is in the middle of two adjacent labels, which is half of the resolution r_i. This is shown in (2.17)

$$\begin{aligned}(s_{i,j} - x_i) &\geq \frac{r_i}{2} \\ \therefore\ \sigma_i &\geq 0.8493 \times \frac{r_i}{2} \\ \Rightarrow \sigma_i &\geq 0.4247 r_i. \end{aligned} \tag{2.17}$$

Therefore, the standard deviation is set to be at least 0.4247 times of the resolution in that dimension.

Equation (2.15) is able to control the bias-variance of the learning system by varying the degree of localization, ld. This is illustrated with a motorcycle crash dataset from [18]. The dataset consists of a one-dimensional time-series of 133 accelerometer readings in an experiment on the efficacy of crash helmets, as shown in Fig. 2.

Experiment was performed by randomly splitting the dataset into 67 training dataset and 66 testing dataset; setting the shrinking rate, $sr = 0.95$; and the no. of labels to 20. Each set of the results was acquired by setting a degree of localization, ld, and being trained for 200 iterations.

Fig. 3 illustrates that the bias-variance of the learning system by simply modifying the degree of localization, which directly controls the width of the membership function. The training accuracy of the system is proportional to the width of the membership function. However, the testing error increases when the degree of localization is more than 16, representing a scenario of overfitting or overtraining.

Thus, the bias-variance dilemma is a system dependant design issue. The FITSK model simplifies the design process by using the degree of localization to control the bias-variance. The degree of localization defines the locality or the vagueness of the input vector to neighboring subspaces. Higher degree of localization refers to higher accuracy (bias) and less generalization (variance), and vice-versa. The locality of local models can be controlled with ease by varying the degree of localization of the FITSK model without the need of repositioning or redefining the linguistic labels.

A possible usage of varying the degree of localization is to relate it to the progress of training in an adaptive system, such that the degree of localization is gradually increased as learning progresses. This ensures that the generalization capability of the learning system is guaranteed when learning commences, and gradually moves toward higher learning accuracy.

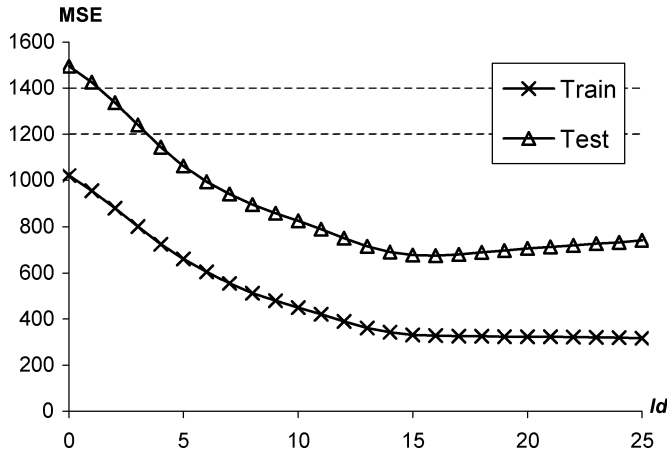

Fig. 3. Training and testing MSEs of FITSK.

C. A First-Order FITSK Model

The FITSK framework takes the form of a first-order FITSK model by adopting first-order local models at layer 5 (functional layer) of the framework. This is achieved by substituting a first-order linear function into (2.8) of the generic FITSK framework

$$\begin{aligned} Z_{\mathrm{F}_{l,m}} &= w_l f_{\mathrm{F}_{l,m}}(X) \\ f_{\mathrm{F}_{l,m}}(X) &= b_{0,(l,m)} + b_{1,(l,m)} x_1 + \ldots \\ &\quad + b_{i,(l,m)} x_i + \ldots + b_{I,(l,m)} x_I \end{aligned} \tag{2.18}$$

where $B_{(l,m)} = [b_{0,(l,m)}, b_{1,(l,m)}, \ldots, b_{i,(l,m)}, \ldots, b_{I,(l,m)}]^T$ refers to the parameter vector of an arbitrary node, $\mathbf{F}_{l,m}$, in layer 5 of the FITSK. The inference performed by the FITSK model is an interpolation of all the relevant local linear models based on the relevance of input data to the fuzzy subspaces associated with the linear model. The output nodes at layer 6 take the form of (2.19)

$$\begin{aligned} y_n &= \sum_{l=1}^{L} w_l f_{\mathrm{F}_{l,m}}(X) \\ &= \sum_{l=1}^{L} w_l \left(b_{0,(l,m)} + b_{1,(l,m)} x_1 + \ldots \right. \\ &\qquad \left. + b_{i,(l,m)} x_i + \ldots + b_{I,(l,m)} x_I \right). \end{aligned} \tag{2.19}$$

For simplicity of discussion, assume that the FITSK has two inputs x_1 and x_2, and a single output y, with only two rules, then

$$\begin{aligned} y &= \sum_{l=1}^{2} w_l f_{l,1}(x_1, x_2) \\ &= w_1 \left(b_{0,(1,1)} + b_{1,(1,1)} x_1 + b_{2,(1,1)} x_2 \right) \\ &\quad + w_2 \left(b_{0,(2,1)} + b_{1,(2,1)} x_1 + b_{2,(2,1)} x_2 \right). \end{aligned} \tag{2.20}$$

Consider that the learning system is presented with P number of training examples, $(X^{(1)}, d^{(1)}), (X^{(2)}, d^{(2)}), \ldots, (X^{(p)}, d^{(p)}), \ldots, (X^{(P)}, d^{(P)})$, where

p index;
$X^{(p)}$ p^{th} input vector, $p \in \{1, 2, \ldots, P\}$;
$d^{(p)}$ p^{th} desired output value.

Equation (2.20) can be easily arranged in a matrix form of

$$AB = C \tag{2.21}$$

or

$$\begin{bmatrix} w_1^{(1)} & w_1^{(1)}x_1^{(1)} & w_1^{(1)}x_2^{(1)} & w_2^{(1)} & w_2^{(1)}x_1^{(1)} & w_2^{(1)}x_2^{(1)} \\ w_1^{(2)} & w_1^{(2)}x_1^{(2)} & w_1^{(2)}x_2^{(2)} & w_2^{(2)} & w_2^{(2)}x_1^{(2)} & w_2^{(2)}x_2^{(2)} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ w_1^{(p)} & w_1^{(p)}x_1^{(p)} & w_1^{(p)}x_2^{(p)} & w_2^{(p)} & w_2^{(p)}x_1^{(p)} & w_2^{(p)}x_2^{(p)} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ w_1^{(P)} & w_1^{(P)}x_1^{(P)} & w_1^{(P)}x_2^{(P)} & w_2^{(P)} & w_2^{(P)}x_1^{(P)} & w_2^{(P)}x_2^{(P)} \end{bmatrix} \times \begin{bmatrix} b_{0,(1,1)} \\ b_{1,(1,1)} \\ b_{2,(1,1)} \\ b_{0,(2,1)} \\ b_{1,(2,1)} \\ b_{2,(2,1)} \end{bmatrix} = \begin{bmatrix} d^{(1)} \\ d^{(2)} \\ \vdots \\ d^{(p)} \\ \vdots \\ d^{(P)} \end{bmatrix} \tag{2.22}$$

with dimensions of matrices A, B, and C being $P \times L$, $L \times 1$, and $P \times 1$ respectively, where L is total the number of linear parameters.

Since the number of training examples is usually greater than the number of parameters to be identified, (2.22) is an over-determined problem and there is no exact solution. However, online estimation of the solution is possible with the formulation of recursive-least-squares algorithm (a special case of the Kalman filter algorithm) in order to perform iterative linear least-square estimation (LSE), as described in [9] and formally listed in [19].

Let a_p be the p^{h} row vector of matrix A, then B can be iteratively estimated using (2.23)

$$\begin{aligned} B_{p+1} &= B_p + S_{p+1}\mathrm{a}_{p+1}^T\left(d^{(p+1)} - \mathrm{a}_{p+1}B_p\right) \\ S_{p+1} &= S_p - \frac{S_p\mathrm{a}_{p+1}^T\mathrm{a}_{p+1}S_p}{1 + \mathrm{a}_{p+1}S_p\mathrm{a}_{p+1}^T} \\ p &= 0, 1, \ldots, P-1 \end{aligned} \tag{2.23}$$

with initial conditions of $B_0 = \mathbf{0}$ and $S_0 = \gamma\mathbf{I}$, where γ is a large positive number and $\mathbf{I}$ is the identity matrix of dimension $L \times L$.

The strengths of using the recursive least-squares algorithm for the first-order FITSK parameter identification are its online learning capability, fast convergence speed with minimal learning iterations, and optimal global learning performance [4], [7], [20]. However, it encounters two major problems. Firstly, the recursive-least-squares algorithm has high computational and space complexities. For example, a first-order TSK fuzzy model of I-input, L-rules will have an S matrix of dimension $L^2(I+1)^2$, and the computational cost is in order of $\mathrm{O}(L^2(I+1)^2)$. Secondly, the system identified using the algorithm has erratic local behavior and the local models have low local interpretability [4], [7]. This jeopardizes the initial motivation of using the TSK model to gain insights into the local models, such that interpretability of the TSK model is possible [4]. Therefore, the parameter learning algorithm of TSK model is a multi-objective identification problem; which would require a balance between good global learning accuracy and local model interpretability [7].

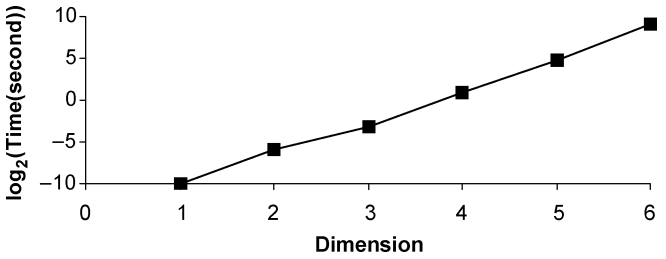

Fig. 4. Time-complexity of FITSK with recursive least-squares learning.

A simple experiment was performed to demonstrate the exponential growth in computational complexity with FITSK when adopting the recursive least-squares parameter learning (described in (2.23)). Consider a multidimensional function of the form described by (2.24)

$$y = \sum_{i=1}^{Dim} x_i^2 \tag{2.24}$$

where Dim is the dimension of the function.

Fig. 4 illustrates the time-complexity of first-order FITSK with recursive least-squares algorithm for parameter learning of function (2.24) with different number of input dimension. There are three labels for each dimension at layer 2 of FITSK in this experiment. The time-complexity of the learning process suffers from an exponential order as the dimension increases. This demonstrates the seriousness of the *curse-of-dimensionality* problem [10] in the FITSK model with recursive least-squares learning. Similar significance of the problem can be easily extensible to the general TSK model using a global learning algorithm, for example, the ANFIS model [9], [21].

With the hindsight from the above discussion, the modeling of the first-order FITSK should be positioned as a multi-objective problem; namely: 1) relaxing the computational complexity; 2) improving the local model interpretability; and 3) maintaining the global learning accuracy. These motivate the derivation of a localized version of recursive least-squares, which is design to achieve the three criteria, and simultaneously being an online algorithm with fast convergence to optimal performance.

Consider a local approximation of an arbitrary local model (denoted by (l, m)) in layer 5 of FITSK. Each training example is being influenced by the weight w_l. Thus, the local approximation can be represented as input-output relationships as shown in (2.25)

$$\begin{bmatrix} w_l^{(1)} & w_l^{(1)}x_1^{(1)} & w_l^{(1)}x_2^{(1)} \\ w_l^{(2)} & w_l^{(2)}x_1^{(2)} & w_l^{(2)}x_2^{(2)} \\ \vdots & \vdots & \vdots \\ w_l^{(p)} & w_l^{(p)}x_1^{(p)} & w_l^{(p)}x_2^{(p)} \\ \vdots & \vdots & \vdots \\ w_l^{(P)} & w_l^{(P)}x_1^{(P)} & w_l^{(P)}x_2^{(P)} \end{bmatrix} \begin{bmatrix} b_{0,(l,m)} \\ b_{1,(l,m)} \\ b_{2,(l,m)} \end{bmatrix} = \begin{bmatrix} w_l^{(1)}d^{(1)} \\ w_l^{(2)}d^{(2)} \\ \vdots \\ w_l^{(p)}d^{(p)} \\ \vdots \\ w_l^{(P)}d^{(P)} \end{bmatrix} \tag{2.25}$$

in a form of that can be identified by the least-squares algorithm, $A_{(l,m)}B_{(l,m)} = C_{(l,m)}$.

Letting $\mathrm{a}'_{p,(l,m)}$ be the p^{th} row vector of matrix $A_{(l,m)}$, and $\mathrm{a}_{p,(l,m)} = [1\, x_1^{(p)}\, x_2^{(p)}]$. To simplify the discussion, the subscript

(l,m) is omitted from (2.26) onwards since each local model is updated locally

$$\begin{aligned} \mathrm{a}'_p &= \left[w^{(p)} \;\; w^{(p)} x_1^{(p)} \;\; w^{(p)} x_2^{(p)} \right] \\ &= w^{(p)} \left[1 \;\; x_1^{(p)} \;\; x_2^{(p)} \right] = w^{(p)} \mathrm{a}_p \end{aligned} \tag{2.26}$$

then B can be iteratively estimated by substituting the solution from (2.23) with (2.26)

$$\begin{aligned} B_{p+1} &= B_p + \left(w^{(p)} \right)^2 S_{p+1} \mathrm{a}_{p+1}^T \left(d^{(p+1)} - \mathrm{a}_{p+1} B_p \right) \\ S_{p+1} &= S_p - \frac{\left(w^{(p)} \right)^2 S_p \mathrm{a}_{p+1}^T \mathrm{a}_{p+1} S_p}{1 + \left(w^{(p)} \right)^2 \mathrm{a}_{p+1} S_p \mathrm{a}_{p+1}^T} \\ & p = 0, 1, \ldots, P-1 \end{aligned} \tag{2.27}$$

and the local learning error is shown in (2.28)

$$\delta_{p+1} = \left(d^{(p+1)} - \mathrm{a}_{p+1} B_p \right) \tag{2.28}$$

which is an approximation of the global learning error as

$$\delta_{p+1} = \left(d^{(p+1)} - y^{(p+1)} \right) \tag{2.29}$$

Therefore, (2.27) can be simplified into

$$\begin{aligned} B_{p+1} &= B_p + \left(w^{(p)} \right)^2 S_{p+1} \mathrm{a}_{p+1}^T \delta_{p+1} \\ S_{p+1} &= S_p - \frac{S_p \mathrm{a}_{p+1}^T \mathrm{a}_{p+1} S_p}{\frac{1}{\left(w^{(p)} \right)^2} + \mathrm{a}_{p+1} S_p \mathrm{a}_{p+1}^T} \\ & p = 0, 1, \ldots, P-1 \end{aligned} \tag{2.30}$$

which is the optimal recursive solution to the least-squares problem $AB = C$. Therefore, the mean-squared-error (MSE) criterion is the optimal objective function for this least-squares-minimization problem. Subsequently, the optimal locally weighted MSE criterion of a local model is computed with (2.31)

$$Local\_MSE = \frac{1}{P} \sum_{p=1}^{P} \left(w^{(p)} \delta_p \right)^2. \tag{2.31}$$

The local_MSE criterion is the objective measure of the learning accuracy of a locally weighted local model. The total locally weighted MSE of all the local models is the sum of all local_MSEs.

$$Local\_MSE_{\text{total}} = \frac{1}{P} \sum_{l=1}^{L} \sum_{m=1}^{M} \sum_{p=1}^{P} \left(w_l^{(p)} \delta_{(l,m),p} \right)^2 \tag{2.32}$$

where

$L \times M$ total number of local models;
(l,m) indexing of the local model.

On the other hand, (2.30) can be generalized with a degree of localization, ld

$$\begin{aligned} B_{p+1} &= B_p + \left(w^{(p)} \right)^{ld} S_{p+1} \mathrm{a}_{p+1}^T \delta_{p+1}, \\ S_{p+1} &= S_p - \frac{S_p \mathrm{a}_{p+1}^T \mathrm{a}_{p+1} S_p}{\frac{1}{\left(w^{(p)} \right)^{ld}} + \mathrm{a}_{p+1} S_p \mathrm{a}_{p+1}^T} \\ & p = 0, 1, \ldots, P-1 \end{aligned} \tag{2.33}$$

with initial conditions of $B_0 = \mathbf{0}$ and $S_0 = \gamma \mathbf{I}$, and a degree of localization of value 2 will always result in the optimal error as shown in (2.31).

Equation (2.33) is a proposed localized version of recursive least-squares algorithm. Each local model has its local copy of S matrix and B vector. The parameter tuning is computed locally instead of globally with respect to all the local models. The influence of an individual local model for a training example is measured with its weight. The degree of localization indicates the locality of the learning to the local model, and is indirectly proportional to the influence of training examples that are further away in the input spaces to a particular local model. A higher degree of localization will often result in better interpretability of local models.

Consider that K number of data points nearest to the centre of a local model in the local subspaces is responsible for influencing the interpretability of the local model. Then, the interpretability of the local models is measured with a K-central-nearest MSE (KCN_MSE) criterion, which is computed by (2.34)

$$KCN\_MSE = \frac{1}{N \times K} \sum_{n=1}^{N} \sum_{k=1}^{K} \delta_{n,k}^2 \tag{2.34}$$

where

$\delta_{n,k}$ error of the n^{th} local model w.r.t. the k^{th} data point;
K total number of data points under consideration;
k index.

The level of interpretability of local models is related to the degree of localization, and will be demonstrated with a simulation in Section III-C using the criterion described in (2.34).

The first-order FITSK fuzzy model with local recursive least-squares of I-input, L-rules will have L number of S matrices, each with dimension of $(I+1)^2$. The local recursive least-squares is repeated for each local model during the training process, the computational cost is in order of $\mathrm{O}(L(I+1)^2)$. Furthermore, the weight of most local models is zero or near-to-zero for high-dimensional FITSK model during a particular training instance. The computational cost can be further reduced by bypassing the training of these near-to-zero weight local models. This significantly reduces the computational and space complexities, especially for higher-dimension FITSK model.

Fig. 5 illustrates the time-complexity of a first-order FITSK with the local recursive least-squares algorithm for the parameter learning of the function described by (2.24) with different number of input dimension. There are 3 labels for each dimension at layer 2 of FITSK in this experiment. The time-complexity of the learning process is still of exponential order. However, the computational cost has been significantly reduced by comparing with the time-complexity of FITSK with global recursive least-squares as shown in Fig. 4. By comparing Fig. 5 and Fig. 4, the computational cost for a 10-dimensional data with local recursive least-squares algorithm is lower than the cost of the global recursive least-squares algorithm with 5-dimensional data. This demonstrates that the FITSK model

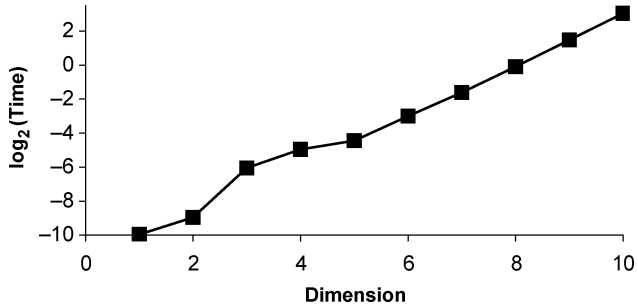

Fig. 5. Time-complexity of FITSK with local recursive least-squares learning.

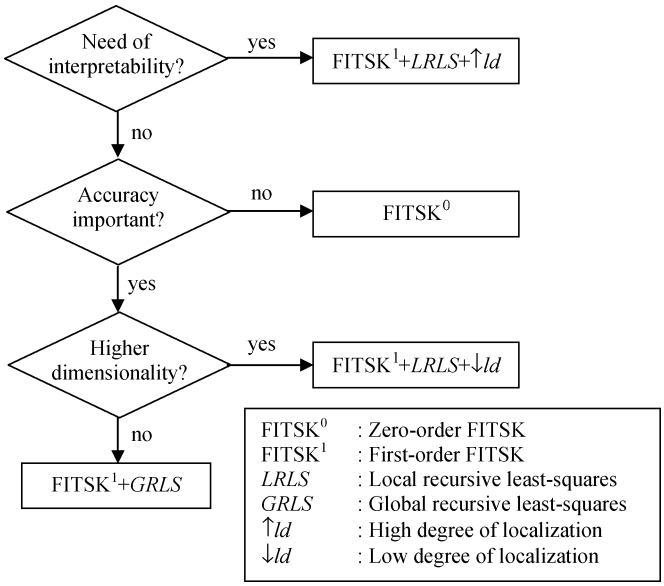

Fig. 6. Flowchart on 3-criteria model selection of FITSK.

adopting the local recursive least-squares algorithm has better immunity to the curse-of-dimensionality.

This research focuses on providing a general FITSK framework such that a specific model can be formulated with ease depending on different design criteria and considerations. The criteria for considerations are the interpretability, accuracy and dimensionality of the system; all these criteria are design or problem-dependant. Formulating this design problem as a multi-objective optimization problem is not justifiable practically since these objectives are system-dependant. Therefore, a guideline is provided to cater for the selection of FITSK model based on these three criteria.

Fig. 6 illustrates the design guideline on the model selection of FITSK based on three criteria. If the interpretability of a learning system is of the highest priority, a first-order FITSK model with localized version of recursive least-squares is adopted with a high degree of localization. This is to ensure that the learning is performed locally for each regional subspace. However, a zero-order FITSK model is chosen when both interpretability and accuracy is not important, and a straightforward and fast adaptive system is required. Finally, a first-order FITSK model is necessary on the condition that accuracy remains an important criterion. The global recursive least-squares learning algorithm is adopted when the system has low dimensionality since this approach is computational costly. On the other hand, for higher dimensional system, the local recursive least-squares is chosen, and since interpretability is not a main consideration, a low degree of localization will adequately accommodate such need.

TABLE I
SPECIFICATIONS OF TASKS

Task	No. of dimensions	Selected dimensions	Train datasets	Test datasets
(a)	4	2	25	25
(b)	5	2	35	35
(c)	10	3	50	50
(d)	4	4	500	500
(e)	1	1	133	---

III. Simulation Results and Analysis

Five nonlinear estimation tasks are studied in this section; the first three benchmarking tasks are taken from [22], they are; namely: 1) a nonlinear system; 2) a human-operated chemical plant; and 3) a daily price of a stock. These are benchmarked against five models [22] in a thesis [23] and the results are summarized in Section III-A. The fourth task is a Mackey–Glass time series prediction taken from [9], and is compared against the ANFIS model and TSK^0-FCMAC model as proposed by [24]. The last task is a motorcycle crashing simulation taken from [18], to comprehend the interpretability issue [5] of the first-order FITSK model. The detail descriptions of each task are to be referred to each respective research [9], [18], [22].

The specifications of these tasks are shown in Table I.

The third column of Table I shows the selected dimensions of the original problem via structure identification [22]. The MSE and Pearson correlation coefficient are used to benchmark the results, The MSE measures the performance accuracy in terms of training or testing result. The Pearson product moment correlation coefficient is a dimensionless index that ranges from -1.0 to 1.0 and reflects the extent of a linear relationship between two sequences of data.

Defining MSE and root-MSE between two data vectors as (3.1)

$$MSE(\vec{x},\vec{y}) = \frac{1}{n}(\vec{x}-\vec{y})^{\mathrm{T}}(\vec{x}-\vec{y})$$
$$RMSE(\vec{x},\vec{y}) = \sqrt{MSE(\vec{x},\vec{y})} \tag{3.1}$$

where

MSE MSE function;
$RMSE$ root-MSE function;
$\vec{x}, \vec{y}$ two data vectors;
n number of elements in the vector.

The Pearson correlation coefficient is defined in (3.2)

$$R(\vec{x},\vec{y}) = \frac{C(\vec{x},\vec{y})}{\sqrt{C(\vec{x},\vec{x})C(\vec{y},\vec{y})}} \tag{3.2}$$

where

R Pearson correlation coefficient function;
$\vec{x}, \vec{y}$ two sequence of data vectors;
$C(.)$ covariance between two data vectors.

The two data vectors for performance evaluations are the desired and the actual outputs from the model respectively. The FITSK^0 model refers to the zero-order FITSK model with a

TABLE II
RELATIVE MSE FOR THREE TASKS

MSE	Sugeno's P&P-G	Sugeno's P	Sugeno's P-G	Mamdani's	Turksen's IVCRI	FITSK^0
Task 1	0.40	0.90	0.54	1.00	0.82	0.39
Task 2	0.15	0.33	1.00	0.34	0.13	0.02
Task 3	0.56	0.21	1.00	0.24	0.55	0.20

TABLE III
RELATIVE PEARSON CORRELATION COEFFICIENT FOR THREE TASKS

R	Sugeno's P&P-G	Sugeno's P	Sugeno's P-G	Mamdani's	Turksen's IVCRI	FITSK^0
Task 1	0.98	0.66	1.00	0.58	0.72	0.98
Task 2	0.98	0.94	0.99	0.94	1.00	1.00
Task 3	0.80	1.00	0.79	0.98	0.75	1.00

learning rate of 0.2, and standard deviation of 0.4247 times the resolution. The FITSK^1-GRLS denotes the first-order FITSK model with global recursive least-squares learning (based on (2.23)), and FITSK^1-LRLS represents the first-order FITSK model with local recursive least-squares learning (based on (2.33)).

A. Nakanishi's Nonlinear Estimation Tasks

The three nonlinear systems were described in [22]. They are; namely: 1) a nonlinear system; 2) a human operated chemical plant; and 3) a daily price of a stock. The result of FITSK^0 model is compared against five other models, the Sugeno's P&P-G, P, P-G, Mamdani's model, and Turksen's IVCRI model. These models differ in terms of modeling and theoretical basis, the details can be found in [22]. Turksen's PVAAR and IVAAR models are not included in the evaluation due to the large number of missing values in their results. This is because the performance results of these two models will significantly degrade by considering such missing values to be out-of-range, and are substituted with default value of zeros.

Detailed result discussions of these three tasks are presented in [23]. Table II summarizes the performances of models in the three tasks with MSE. The table normalizes each MSE with the best result within the row. Summarizing all the three tasks as discussed above, FITSK^0 is able to achieve the lowest MSE for all the tasks with significant accuracy improvements over all the other models. Table III summarizes the performances of models in the three tasks with Pearson correlation coefficient. The table normalizes each coefficient with the best result within the row. As shown in the table, the FITSK^0 model is the only model that is capable of achieving the optimal performances for all the three tasks. Models such as Sugeno's (P), (P-G) or Turksen's (IVCRI) performed well for a particular task and but not the others. This demonstrates that the FITSK^0 model is outperforming other popular models despite being a simple zero-order local model TSK network for nonlinear system estimation.

B. Mackey–Glass Time Series Prediction

The dynamics of Mackey–Glass differential delay equation is defined in (3.3).

$$\dot{x}(t) = \frac{0.2x(t-\tau)}{1+x^{10}(t-\tau)} - 0.1x(t) \tag{3.3}$$

This time series is a popular benchmark problem considered by several researchers [9], [24].

The time series is obtained similarly as in Jang's thesis [9]. The fourth-order Runge-Kutta method was applied to compute the numerical approximation with time step of 0.1, initial condition $x(0) = 1.2$, $\tau = 17$, and $x(t) = 0$ for $t < 0$, for a time period of $0 \leq t \leq 2000$. From the computed series, 1000 input-output data pairs from $118 \leq t \leq 1117$ were extracted with the following format: $[x(t-18), x(t-12), x(t-6), x(t); x(t+6)]$. The goal of the task is to use known values of the time series of past 18, 12, 6, and current time to predict the 6th instance ahead in time. The first 500 data pairs are used as training datasets while the remaining 500 data pairs are used as testing datasets.

The models to be studied are TSK^0-FCMAC [24], ANFIS [9], FITSK^0, FITSK^1-GRLS and FITSK^1-LRLS. FITSK^1-LRLS implements the first-order FITSK model with the proposed local recursive least-squares approach with a localization factor of 9. The models to be compared consist of TSK-type fuzzy models of zero-order (TSK^0-FCMAC, FITSK^0), and first-order (ANFIS, FITSK^1-GRLS, FITSK^1-LRLS).

The training datasets were randomly permutated into 100 different training datasets in TSK^0-FCMAC simulation, while the other models had three different permutations of training datasets. All the models are running on the same computer platform, with debug version of executables being generated from C++ source codes (C for ANFIS).

Table IV lists the testing results of Mackey–Glass series for the various models. The training process lasted for 20 epochs for the zero-order models. The second column of the table lists the number of labels of the respective models for each input dimension at the layer 2 of the models; and the number of rules of the models are computed as $(\mathrm{no.\ of\ fuzzy\ labels})^4$ since there are four input dimensions. TSK^0-FCMAC had 218 rules as those zero-weight rules were not included in the computation of the total number of rules.

Comparing the zero-order models, the FITSK^0 is able to achieve a lower testing error despite using lesser number of rules than the TSK^0-FCMAC, which is due to the Gaussian membership function adopted in layer 2 of the model. The average training time of FITSK^0 is acceptably low despite the usage of

TABLE IV
TESTING RESULTS OF MACKEY-GLASS SERIES

Model	No. of labels (layer 2)	Average RMSE	Std. dev. (RMSE)	Training epochs	Avg training time (s)	No. of rules
TSK^0-FCMAC	4	0.02229	0.00468	20	1.92	218
$FITSK^0$	3	0.01842	0.00004	20	0.99	81
ANFIS	2	0.00295	0.00000	1	0.27	16
$FITSK^1$-GRLS	2	0.00321	0.00000	1	0.33	16
$FITSK^1$-GRLS	3	0.00139	0.00000	1	9.78	81
$FITSK^1$-LRLS	2	0.03070	0.00000	1	0.09	16
$FITSK^1$-LRLS	3	0.01926	0.00000	1	0.13	81

Gaussian membership function, which is computational costly comparing with the trapezoidal or triangular fuzzy functions in TSK^0-FCMAC.

Next, it is appropriate to compare the two first-order models with two linguistic labels, the ANFIS model at the fourth row and FITSK^1-GRLS model at the fifth row. It is shown that both models have similar accuracy and training time since they both adopt the global recursive least-squares learning algorithm. The minor differences may be due to programming differences. The 6th row demonstrates the FITSK^1-GRLS model with three linguistic labels at layer 2, and thus larger number of rules (local models). As expected, the accuracy improves significantly over the previous 2 models as the number of local models increases. However, the drawback of this is the significant growth in computational cost, which increases from 0.33 s to 9.78 s.

The next two models are the FITSK^1-LRLS with two and three linguistic labels in each dimension; both adopting local recursive least-squares parameter learning with degree of localization being set to numeric value 2. The computational time of these two models demonstrate a significant reduction comparing with models with global recursive least-squares approach. On the other hand, the growth in computational cost due to the increased number of local models is insignificant (from 0.09 s to 0.13 s). This clarifies that the parameter learning algorithm with local recursive least-squares approach is insignificantly affected by the curse-of-dimensionality having in the global recursive least-squares approach. Moreover, by using the recursive least-squares parameter learning approaches, all the first-order models are able to achieve fast learning convergence within a single training epoch. This inherits the online learning capability from the recursive least-squares parameter identification technique. The online learning capability is made possible by discarding each training instances after it is processed without repeatedly reassessing it, and the weights are subsequently updated based on the current training instances [25].

C. Motorcycle Crashing Simulation

The final task is the motorcycle crashing simulation data taken from [18], which has been briefly introduced in Section II-B. The data is revisited here to illustrate the interpretability issue in the first-order TSK model. The same dataset has been used in [5] to study the interpretability issue of TSK fuzzy model.

Firstly, eight linguistic labels in layer 2 of the FITSK model are equally position across the space of range [2.4 57.6]. The standard deviation for the Gaussian membership function is chosen to be the minimum, which is 0.4247 of the resolution, as shown in (2.16). The global MSE of the system identified is 459.126 by using the global recursive least-squares parameter identification [(2.23)]. The global learning error is slightly better than the one identified by Yen's model [5], which is 460.62. Yen adopted a singular value decomposition method to identify redundant fuzzy partitions using a firing strength matrix. However, the method is a form of unsupervised clustering; no output value or output performance indication has been used to influence the result of clustering. Thus, the method could only identify the redundant fuzzy partitions among all the fuzzy partitions, but could not recognize the relevancy of the fuzzy partitions with respect to the output of the system. It can be examined from the result in [5] that, the fuzzy partitions identified by Yen's model is not optimal; i.e., there should be a fuzzy partition at a time near-to 26 ms since the output around this region changes drastically. However, the method is unable to monitor such drastic transition at the output space. Thus, Yen's fuzzy partition method performed no better than an equally spaced partition with a fuzzy input formulation as in the FITSK model. This is supported by the higher learning accuracy of the FITSK model than the one by Yen's model.

The interpretability of the first-order FITSK local models is investigated in Fig. 7. Fig. 7 illustrates the learning result of FITSK^1 with different settings. The original dataset is represented using black dots in the figure; the learnt-models are plotted with solid curves; and the corresponding eight local models are indicated by straight bold lines in the figure and denoted numerically from (1) to (8). Fig. 7(a) represents the training result of FITSK^1-GRLS, which is the first-order FITSK model with global recursive least-squares as the parameter learning algorithm. Fig. 7(b), (c), and (d) are first-order FITSK models with local recursive least-squares (FITSK^1-LRLS), the degree of localization of these models are 2, 6, and 10, respectively. By observation, it is easy to infer that the learnt systems have significant differences in the first, second, and fourth local models given different system settings. The global learning in Fig. 7(a) has good global learning error, but the interpretation of local models (1) and (2) are inappropriate. On the other extreme, the local models (1) and (2) in Fig. 7(d) have best interpretability given a high degree of localization ($ld = 10$). Local model (4) in Fig. 7(d) is misinterpreted to the neighboring region; however, if the observation is focused on a small regional subspace, it can be argued that the output data values are almost constant within this region. This coincides with the learning result of local model (4) in

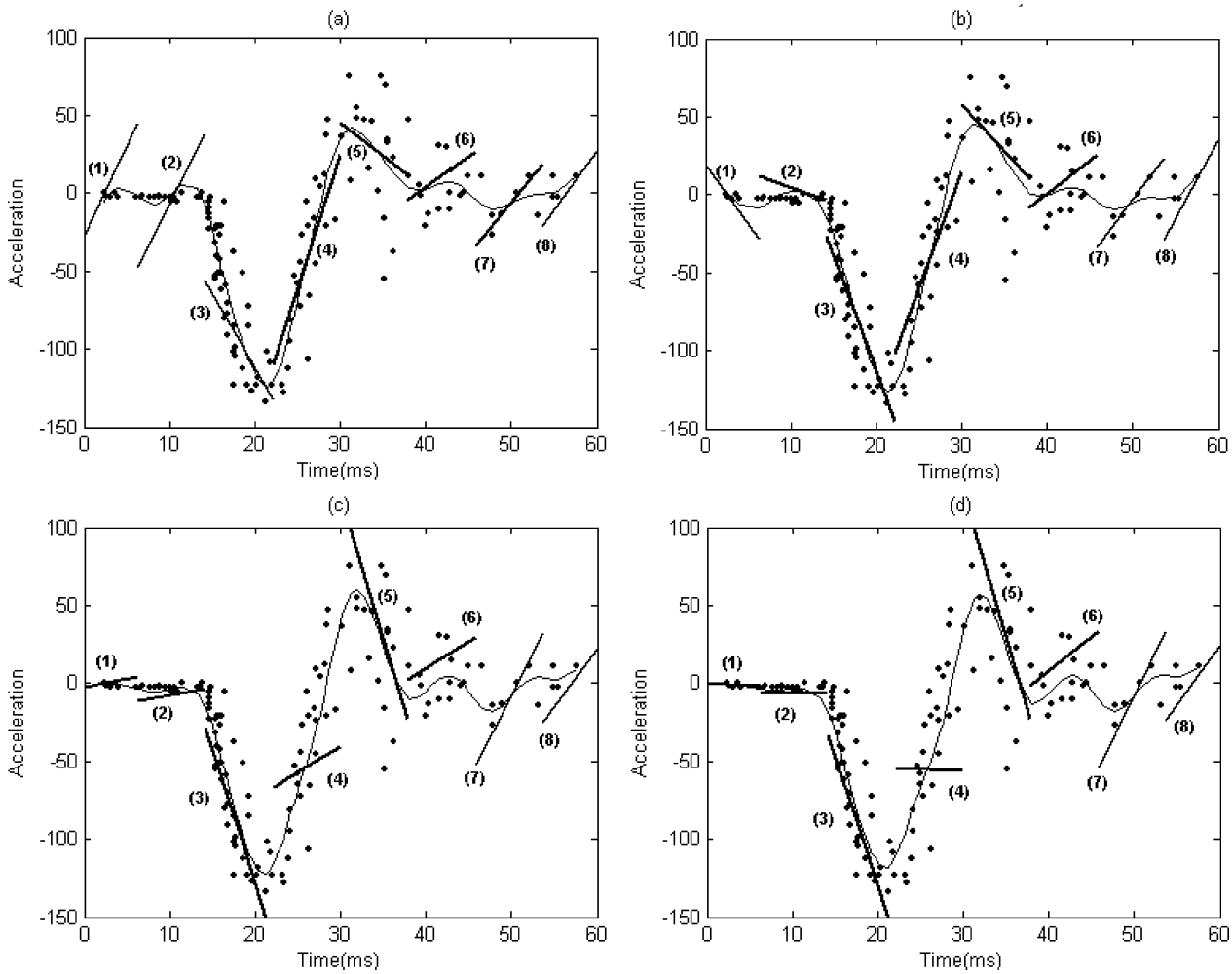

Fig. 7. Global and local learning results of first-order FITSK models. (a) Global learning (FITSK^1-GRLS); (b) local learning (FITSK^1-LRLS, $\mathrm{Id} = 2$); (c) local learning (FITSK^1-LRLS, $\mathrm{Id} = 6$); (d) local learning (FITSK^1-LRLS, $\mathrm{Id} = 10$).

Fig. 7(d), which is almost a constant term model. The results of Fig. 7(b) and (c) are in between the two extreme cases in Fig. 7(a) and (d), this is made possible with smaller settings of degree of localization for these models. By visual inspection, if the influence is confined to a regional subspaces proportional to the weight (matching degree) of the local models as shown in Fig. 7(b), then the FITSK^1-LRLS model with degree of localization 2 has the best local learning accuracy. The local models for this model have a good balanced between learning accuracy and model interpretability. A FITSK^1-LRLS with degree of localization equals 2 has optimal locally weighted MSE as formulated in (2.32).

The learning results of different FITSK^1 settings are illustrated with four different error criteria in Fig. 8. The criteria are the $Global\_MSE$ [(3.1)], $Local\_MSE2$ [(2.32)], $2CN_MSE$ [(2.34) with $K = 2$], and $Local\_MSE1$ being defined by [8], as shown in (3.4). The $Local\_MSE1$ is similar to $Local\_MSE2$ except that the weight of $Local\_MSE2$ is order of 2

$$Local\_MSE1 = \frac{1}{P}\sum_{n=1}^{N}\sum_{p=1}^{P} w_n^{(p)}\delta_{n,p}^2. \tag{3.4}$$

The learning errors of FITSK^1-GRLS are plotted in $ld = 0$, and $ld = [1, 10]$ represents the FITSK^1-LRLS models with

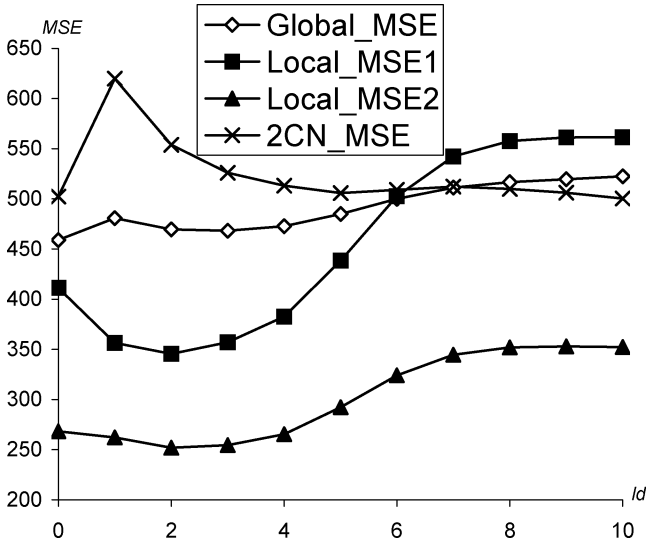

Fig. 8. Learning errors of FITSK^1 with different degree of localization.

degree of localization ranging from 1 to 10 accordingly. As shown with the $Global\_MSE$ criterion, the FITSK^1 with global recursive least-squares learning has the best global learning accuracy. However, the accuracy gradually reduces

as the degree of localization increases. On the other hand, the $Local\_MSE1$ and $Local\_MSE2$ criteria demonstrate that the local recursive least-squares learning with degree of localization 2 has the optimal locally weighted LSE. This is supported by the observation from Fig. 7(b), which demonstrates that a degree of localization of 2 will result in FITSK^1 model with a balanced learning accuracy and local model interpretability. The $2CN\_MSE$ criterion is the strictest formulation of the general KCN_MSE criterion. It uses the two data points nearest to the centre of a local model to measure the interpretability of a local model. The $2CN\_MSE$ criterion illustrates that, the interpretability of the local models increases as proportional to the degree of localization. This coincides with Fig. 7(d), which shows that the local models have best interpretability to the smallest confined region at the centre of the local models.

These conclude that (2.34) is a good indicator of interpretability, and (2.32) is a good criterion to measure the locally weighted least-squares accuracy of the local models. The simulation demonstrates that the FITSK^1 model with local recursive least-squares can be easily reformulated to have different degree of interpretability and local model accuracy by varying the degree of localization; and is flexible to the system designer based on different design considerations.

IV. Conclusions

This paper proposes a generic TSK framework, the FITSK framework. The framework is subsequently mapped to a zero-order FITSK model and a first-order FITSK model. The zero-order and first-order FITSK models assume online parameter learning formulae. In this way, the FITSK models are suitable for adaptive system learning and nonlinear system estimation.

The key strength of the FITSK framework is the flexibility of adopting different formulations and settings to achieve different design criteria. The three design criteria are 1) interpretability; 2) global accuracy; and 3) immunity to curse-of-dimensionality. The zero-order FITSK model is suitable for fast adaptive system without a major consideration of interpretability. The first-order TSK model with global recursive least-squares is appropriate for a low-dimensionality, low-interpretability system with optimal global learning accuracy. A proposed localized version of the recursive least-squares algorithm has been shown to achieve optimal accuracy for locally weighted LSE criterion. Furthermore, the local recursive least-squares has been demonstrated to have better immunity than the global recursive least-squares to the curse-of-dimensionality issue. The first-order FITSK model adopting this local recursive least-squares can be readily applied to higher-dimensionality system with lower computational resources comparing with the one using global recursive least-squares learning. The degree of localization in the local recursive least-squares provides a flexibility to balance the interpretability and system accuracy in the local models. A guideline is discussed based on these three design criteria for model selection of these FITSK models.

The performance of the FITSK models are evaluated using five nonlinear system simulations: 1) a nonlinear system; 2) human operation of a chemical plant; 3) daily price of a stock; 4) Mackey–Glass time series prediction; and 5) motorcycle crashing simulation. The first three simulations demonstrate that the zero-order FITSK model has encouraging performances in terms of learning error and correlation measure. The fourth simulation highlights the computational advantages of the local recursive least-squares in the first-order FITSK model. The final simulation illustrates the modeling flexibility of the first-order FITSK model to tackle the balancing between local model interpretability and learning accuracy.

A possible future work is to automate the model selection of FITSK, which is described in Fig. 6. This is especially important in automatically determining whether the system is likely to suffer from the curse-of-dimensionality issue. Currently, this process is performed manually based on the subjective perception of the designer. A formal approach should be adopted with supporting studies in the future.

The online leaning capability of the FITSK is currently demonstrated with a benchmarking Mackey–Glass time series prediction problem. Further research is desired by applying the FITSK to several process control applications. These include the fluidised bed combustor and the three tanks flow system; such that the FITSK is able to be contrasted against other popular models such as the distributed logic processors [26]–[28] and the blackboard-based BBIPS model [29].

References

[1] M. Sugeno and K. T. Kang, "Structure identification of fuzzy model," *Fuzzy Sets Syst.*, p. 28, 1988.

[2] T. Takagi and M. Sugeno, "Derivation of fuzzy control rules from human operator's control actions," in *Proc. IFAC Symp. Fuzzy Information, Knowledge Representation and Decision Analysis*, 1983, pp. 55–60.

[3] ——, "Fuzzy identification of systems and its application to modeling and control," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-15, no. 1, pp. 116–132, Feb. 1985.

[4] J. Yen and R. Langari, *Fuzzy Logic: Intelligence, Control, and Information*. Englewood Cliffs, NJ: Prentice-Hall, 1999.

[5] J. Yen, L. Wang, and C. W. Gillespie, "Improving the interpretability of TSK fuzzy models by combining global learning and local learning," *IEEE Trans. Fuzzy Syst.*, vol. 6, no. 4, pp. 530–537, Nov. 1998.

[6] J. Abonyi and R. Babuska, "Local and global identification and interpretation of parameters in Takagi–Sugeno fuzzy models," in *Proc. FUZZ-IEEE 2000*, vol. 2, 2000, pp. 835–840.

[7] T. A. Johansen, R. Shorten, and R. Murray-Smith, "On the interpretation and identification of dynamic Takagi–Sugeno fuzzy models," *IEEE Trans. Fuzzy Syst.*, vol. 8, no. 3, pp. 297–313, Jun. 2000.

[8] T. A. Johansen and R. Babuska, "Multiobjective identification of Takagi-Sugeno fuzzy models," *IEEE Trans. Fuzzy Syst.*, vol. 11, no. 6, pp. 847–860, Dec. 2003.

[9] J. S. Jang, "Neuro-Fuzzy Modeling: Architecture, Analyzes and Applications," Ph.D. dissertation, Dept. Elect. Eng. Comput. Sci., Univ. California, Berkeley, 1992.

[10] R. E. Bellman, *Dynamic Programming*. Princeton, NJ: Princeton Univ. Press, 1957.

[11] E. H. Mamdani and S. Assilian, "An experiment in linguistic synthesis with a fuzzy logic controller," *Int. J. Mach. Studies*, vol. 7, no. 1, 1975.

[12] J. C. Bezdek, *Pattern Recognition With Fuzzy Objective Function Algorithms*. New York: Plenum, 1981.

[13] T. K. Kohonen, "Self-organized formation of topologically correct feature maps," *Biol. Cybern.*, vol. 43, pp. 59–69, 1982.

[14] W. L. Tung and C. Quek, "GenSoFNN: a generic self-organizing fuzzy neural network," *IEEE Trans. Neural Netw.*, vol. 13, no. 5, pp. 1075–1086, Sep. 2002.

[15] M. Männle, "Identifying rule-based TSK fuzzy models," in *Proc. EUFIT*, Aachen, Germany, Sep. 1999.

[16] C. C. Lee, "Fuzzy Logic in control systems: fuzzy logic controller: part I," *IEEE Trans. Syst., Man, Cybern.*, vol. 20, no. 2, pp. 404–418, Mar./Apr. 1990.

[17] ——, "Fuzzy logic in control systems: fuzzy logic controller: part II," *IEEE Trans. Syst., Man, Cybern.*, vol. 20, no. 2, pp. 419–435, Mar./Apr. 1990.
[18] W. Hardle, *Applied Nonparametric Regression*. Cambridge, MA: Cambridge Univ. Press, 1990.
[19] C. T. Lin and C. S. Lee, *Neural Fuzzy Systems: A Neuro-Fuzzy Synergism on Intelligent System*. Englewood Cliffs, NJ: Prentice-Hall, 1996.
[20] X. Xu, H. He, and D. Hu, "Efficient reinforcement learning using recursive least-squares methods," *J. Artif. Intell. Res.*, vol. 16, pp. 259–292, 2002.
[21] J. S. Jang, "ANFIS: adaptive-network-based fuzzy inference system," *IEEE Trans. Syst., Man, Cybern.*, vol. 23, no. 3, pp. 665–685, May./Jun. 1993.
[22] H. Nakanishi, I. B. Turksen, and M. Sugeno, "A review and comparison of six reasoning methods," *Fuzzy Sets Syst.*, vol. 57, pp. 257–294, 1993.
[23] K. H. Quah, "Fuzzy Modeling in Reinforcement Learning," Ph.D. dissertation, Sch. Comput. Eng., Nanyang Technological Univ., Singapore, 2005.
[24] C. W. Ting, "Fuzzy Associative Memory Architecture," Ph.D. dissertation, Sch. Comput. Eng., Nanyang Tech. Univ., Singapore, 2002.
[25] W. S. Sarle. (1997) Neural Network FAQ, Part 2 of 7: Learning. [Online]Available: ftp://ftp.sas.com/pub/neural/FAQ2.html
[26] E. Ikonen and K. Najim, "Use of learning automata in distributed fuzzy logic processor training," *Proc. Inst. Elect. Eng., Control Theory Appl.*, vol. 144, no. 3, pp. 255–262, 1997.
[27] K. Najim and E. Ikonen, "Distributed logic processors trained under constraints using stochastic approximation techniques," *IEEE Trans. Syst., Man, Cybern. A*, vol. 29, no. 4, pp. 421–426, Jul. 1999.
[28] E. Ikonen, K. Najim, and U. Kortela, "Neuro-fuzzy modeling of power plant flue-gas emissions," *Eng. Applicat. Artif. Intell.*, vol. 13, pp. 705–717, 2000.
[29] K. W. Oh and C. Quek, "BBIPS: a blackboard-based integrated process supervision," *Eng. Applicat. Artif. Intell.*, vol. 14, pp. 703–714, 2001.

Kian Hong Quah received the B.A.Sc degree in computer engineering in 2000 from Nanyang Technological University, Singapore, where he is currently pursuing the Ph.D. degree in the Centre for Computational Intelligence, School of Computer Engineering.

He was a Research Engineer with Hewlett-Packard (Singapore) Pte. Ltd. from 2000 to 2001. His current research interests are in the areas of neural-fuzzy systems, fuzzy inference system, reinforcement learning, feature subset selection, and pattern recognition.

Chai Quek (M'97) received the B.Sc. degree in electrical and electronics engineering in 1986 and the Ph.D. degree in intelligent control in 1990 from Heriot Watt University, Edinburgh, U.K.

He is an Associate Professor and a member of the Centre for Computational Intelligence (formerly the Intelligent Systems Laboratory), School of Computer Engineering, Nanyang Technological University, Singapore. His research interests include intelligent control, intelligent architectures, AI in education, neural networks, fuzzy neural systems, neurocognitive informatics, and genetic algorithms.

He is a member of the IEEE Technical Committee on Computational Finance.