M-Estimator and D-Optimality Model Construction Using Orthogonal Forward Regression

Xia Hong, *Senior Member, IEEE* and
Sheng Chen, *Senior Member, IEEE*

***Abstract*—This correspondence introduces a new orthogonal forward regression (OFR) model identification algorithm using D-optimality for model structure selection and is based on an M-estimators of parameter estimates. M-estimator is a classical robust parameter estimation technique to tackle bad data conditions such as outliers. Computationally, The M-estimator can be derived using an iterative reweighted least squares (IRLS) algorithm. D-optimality is a model structure robustness criterion in experimental design to tackle ill-conditioning in model structure. The orthogonal forward regression (OFR), often based on the modified Gram–Schmidt procedure, is an efficient method incorporating structure selection and parameter estimation simultaneously. The basic idea of the proposed approach is to incorporate an IRLS inner loop into the modified Gram–Schmidt procedure. In this manner, the OFR algorithm for parsimonious model structure determination is extended to bad data conditions with improved performance via the derivation of parameter M-estimators with inherent robustness to outliers. Numerical examples are included to demonstrate the effectiveness of the proposed algorithm.**

***Index Terms*—Forward regression, Gram–Schmidt, identification, M-estimator, model structure selection.**

I. Introduction

Various neural networks such as radial basis function (RBF) networks can be expressed as a linear-in-the-parameters model structure where the system output is a linear combination of nonlinear basis functions. Provided that there is a separate mechanism for determining centers/widths of these basis functions, the basis weights/parameters can be trained using linear optimization techniques. The architecture or topology of the class of linear-in-the-parameters modeling networks enables them to be readily assessed in terms of their modeling capability, structure, learning, construction, and numeric stability, since the results of quadratic optimization and linear algebra are directly applicable. Moreover, by applying linear regression statistical techniques to the identification of this type of neural networks, it is possible to model the observational data in a statically optimal sense to achieve improved performance for a wide range of applications/tasks in the field of signal processing, dynamical system modeling, and control.

The general method of M–estimation [1] is well established in order to tackle outliers in observational data. As a generalization of maximum-likelihood estimation method for data with outliers, the M-estimator uses some cost functions which increase less rapidly than that of least squares estimators as the residual departs from zero, so the parameters estimator is more robust to outliers. Computationally, M-estimator can be derived using an iterative reweighted least squares (IRLS) algorithm. M-estimation has been applied successfully to time series prediction, image processing and pattern recognition [2]–[4]. Two major aspects of system identification are model structure determination and parameter estimation. While M-estimator is concerned with parameter robustness, conventional optimum experimental designs are concerned with model structure robustness [5]. In optimum experimental design, model adequacy is evaluated by statistical measures of goodness via experimental design criteria, e.g. A- and D-optimality. By quantitatively measuring the model adequacy as function of the eigenvalues of the design matrix, design efficiency and experimental effort of designs can be optimized.

The orthogonal forward regression (OFR) is an efficient algorithm to determine a parsimonious model structure [6]. Driven by requirements for improved model generalization, a few variants of OFR have been introduced in order to tackle ill-conditioning problem that may be associated with least squares parameter estimates [7]–[11]. Recently, variants of the forward OFR algorithms have been introduced by modifying the selective criteria to include A- and D-optimality in forward regression [12], [13] to form hybrid approaches applicable to neural networks modeling. Although these methods do not generally need the assumption of a normal error distribution, the parameter estimator may not be statistically optimal if the data exhibit bad conditions such as outliers.

Alternatively there exists a vast amount of work on sparse modeling including the well-known support vector machine (SVM) [14], which is often used in classification tasks [15] and can also be used in sparse regression modeling [16]. SVM is regarded as a robust modeling approach and based on a structural risk minimization (SRM) principle, that is to minimize an estimate of the upper bound of model generalization. The model sparsity and robustness of SVM can be achieved by incorporating an ε-insensitive function in the loss function, as proposed by Vapnik [14]. The ε-insensitive function and Huber loss function used in M-estimator shares a similarity of using l_1 norm (see Section II-A). Because the implementation of SVM sparse modeling with the ε-insensitive function is solved via constrained quadratic programming (QP), it is computationally expensive. It has been shown that OFR algorithm can be combined with SVM to improve model sparseness [16].

This paper presents a new model identification algorithm that combines the M-estimator with forward regression. Based on the modified Gram–Schmidt procedure for OFR, the proposed algorithm incorporates an IRLS inner loop into the modified Gram–Schmidt procedure to derive the M-estimator of model parameters. In combination with D-optimality for model structure selection, the proposed algorithm simultaneously derive robust model structure and parameter estimates for bad data conditions.

The paper is organized as follows. Section II initially introduces methodologies relevant to the proposed algorithm, including general nonlinear regression modeling based on OFR algorithm with D-optimality and the concept of the M-estimator. Section III introduces the model identification algorithm using forward regression with M-estimation. Numerical examples are used to demonstrate the efficacy of the algorithm in Section IV and conclusions are given in Section V.

II. Preliminaries

A linear regression model (RBF neural network, B-spline neurofuzzy network) can be formulated as [17], [18]

$$y(t) = \sum_{k=1}^{M} p_k\left(\mathbf{x}(t)\right)\theta_k + \xi(t) \tag{1}$$

where $t = 1, 2, \cdots, N$, and N is the size of the estimation data set. $y(t)$ is system output variable, $\mathbf{x}(t) = [x_1(t), \cdots, x_n(t)]^T$ is system input vector with an assumed known dimension of n. $p_k(\bullet)$ is a known

Manuscript received March 11, 2004; revised June 14, 2004. This work was supported by the EPSRC UK. This paper was recommended by Associate Editor P. De Wilde..

X. Hong is with the Department of Cybernetics, University of Reading, RG6 6AY Reading, U.K.

S. Chen is with the School of Electronics, University of Southampton, SO17 1BJ Southhampton, U.K.

Digital Object Identifier 10.1109/TSMCB.2004.839910

nonlinear basis function, such as RBF, or B-spline fuzzy membership functions. $\xi(t)$ is an uncorrelated model residual sequence with zero mean and variance of σ^2. θ_k is model parameter, and M is the number of regressors.

Equation (1) can be written in the matrix form as

$$\mathbf{y} = \mathbf{P}\boldsymbol{\Theta} + \Xi \tag{2}$$

where $\mathbf{y} = [y(1), \cdots, y(N)]^T$ is the output vector. $\boldsymbol{\Theta} = [\theta_1, \cdots, \theta_M]^T$ is parameter vector, $\Xi = [\xi(1), \cdots, \xi(N)]^T$ is the residual vector, and $\mathbf{P}$ is the regression matrix

$$\mathbf{P} = \begin{bmatrix} p_1(1) & p_2(1) & \cdots p_k(1) \cdots & p_M(1) \\ p_1(2) & p_2(2) & \cdots p_k(2) \cdots & p_M(2) \\ & & & \\ p_1(N) & p_2(N) & \cdots p_k(N) \cdots & p_M(N) \end{bmatrix}$$

with $p_k(t) = p_k(\mathbf{x}(t))$. Denote the column vectors in $\mathbf{P}$ as $\mathbf{p}_k = [p_k(1), \cdots, p_k(N)]^T$, $k = 1, \cdots, M$. An orthogonal decomposition of $\mathbf{P}$ is

$$\mathbf{P} = \mathbf{W}\mathbf{A} \tag{3}$$

where $\mathbf{A} = \{\alpha_{ij}\}$ is an $M \times M$ unit upper triangular matrix and $\mathbf{W}$ is an $N \times M$ matrix with orthogonal columns that satisfy

$$\mathbf{W}^T\mathbf{W} = \text{diag}\{\kappa_1, \cdots, \kappa_M\} \tag{4}$$

with

$$\kappa_k = \mathbf{w}_k^T\mathbf{w}_k, \quad k = 1, \cdots, M \tag{5}$$

so that (2) can be expressed as

$$\mathbf{y} = (\mathbf{P}\mathbf{A}^{-1})(\mathbf{A}\Theta) + \Xi = \mathbf{W}\Gamma + \Xi \tag{6}$$

where $\Gamma = [\gamma_1, \cdots, \gamma_M]^T$ is an auxiliary vector. The above orthogonal decomposition can be realized by the modified Gram–Schmidt algorithm [6], in which least squares parameter estimates are usually used. Based on the modified Gram–Schmidt algorithm, a few variants of forward OLS algorithms have been introduced to improve model generalization capability based on the concepts from Bayesian regularization/basis pursuit [9], experimental design and leave-one-out (LOO) score, respectively, [10], [11]. Although these methods do not generally need normality error distribution assumption, the parameter estimator may not be statistically optimal if the data exhibit bad conditions such as outliers, or are heavy tailed. The general method of tackling this problem is well established as M-estimation [1], which is a generalization of maximum-likelihood estimation method for data with outliers. The M-estimator [1] is described in the following section.

A. M-Estimators

The M-estimators have been well studied [1]. Considering the linear regression model given by (1), the M-estimator minimizes the cost function

$$V_M = \sum_{t=1}^{N} \rho\left(\xi(t)\right) \tag{7}$$

where the function $\rho(\xi(t))$ is some predetermined non-negative functionals for different types of estimators, e.g., for least squares $\rho(\xi(t)) = \rho_L(\xi(t)) = \xi^2(t)$. Typically, $\rho(\xi(t))$ is an even function and nondecreasing with respect to the absolute value of $\xi(t)$. The problem of least squares estimator is that V_M will be influenced by any outlier typified by a large absolute value $\xi(t)$. The general M-estimator can tolerate undetected outliers by assigning a smaller weight to observations with residuals with large absolute values, so the parameter estimates are less vulnerable to unusual data. The most common types of M-estimators are the *Huber* estimator given by [1]

$$\rho_H(\xi) = \begin{cases} \frac{1}{2}\xi^2, & \text{for } |\xi| \le \tau \\ \tau|\xi| - \frac{1}{2}\tau^2, & \text{for } |\xi| > \tau \end{cases} \tag{8}$$

or the Turkey *bisquare* estimator, given by

$$\rho_B(\xi) = \begin{cases} \frac{\tau^2}{6}\left\{1 - \left[1 - \left(\frac{\xi}{\tau}\right)^2\right]^3\right\}, & \text{for } |\xi| \le \tau \\ \frac{1}{6}\tau^2 & \text{for } |\xi| > \tau \end{cases} \tag{9}$$

where the parameter τ is called a tuning constant, e.g. it is common to choose $\tau = 1.345\sigma$ for the *Huber* estimator and $\tau = 4.685\sigma$ for the Turkey *bisquare* estimator.[1]

The M-estimator can be derived by setting

$$\left.\frac{\partial V_M}{\partial \Theta}\right|_{\Theta=\hat{\Theta}_M} = \mathbf{0} \tag{10}$$

to yield

$$\frac{\partial V_M}{\partial \Theta} = \mathbf{P}^T\boldsymbol{\psi} = \mathbf{0} \tag{11}$$

where $\mathbf{0}$ is zero vector.

$$\begin{aligned} \boldsymbol{\psi} &= \left[\frac{\partial V_M}{\partial \xi(1)}, \ldots, \frac{\partial V_M}{\partial \xi(N)}\right]^T \\ &= \left[\psi\left(\xi(1)\right), \ldots, \psi\left(\xi(N)\right)\right]^T \end{aligned} \tag{12}$$

where $\psi(\xi)$ is the derivative of $\rho(\xi)$ with respect to ξ. Define the weight function

$$\omega(t) = \frac{\psi\left(\xi(t)\right)}{\xi(t)}, \quad \text{for} \quad t = 1, \ldots, N. \tag{13}$$

Equation (11) can be written as

$$\mathbf{P}^T\Omega\Xi = \mathbf{0} \tag{14}$$

where $\Omega = \text{diag}\{\omega(1), \omega(2), \ldots \omega(N)\}$, whose solution is given as the weighted least squares

$$\hat{\Theta}_M = \left\{\mathbf{P}^T\Omega\mathbf{P}\right\}^{-1}\mathbf{P}^T\Omega\mathbf{y}. \tag{15}$$

Because $\omega(t)$s are primarily unknown, an iteratively reweighted least square (IRLS) is required. The M-estimator IRLS procedure is as follows:

Denote m as the iteration step. Initially set $m = 1$, $\Omega^{(1)} = \mathbf{I}$ (i.e. least squares) to derive an initial model residuals $\xi^{(1)}(t)$, then for $m = 2, \ldots, m_w$

$$\omega^{(m)}(t) = \frac{\psi\left(\xi^{(m-1)}(t)\right)}{\xi^{(m-1)}(t)}, \quad \text{for} \quad t = 1, \ldots, N. \tag{16}$$

From (8) and (9), the weight functions of *Huber* and the Turkey *bisquare* estimator can be explicitly given by

$$\omega_H^{(m)}(t) = \begin{cases} 1, & \text{for } \left|\xi^{(m-1)}(t)\right| \le \tau \\ \frac{\tau}{\left|\xi^{(m-1)}(t)\right|}, & \text{for } \left|\xi^{(m-1)}(t)\right| > \tau \end{cases} \tag{17}$$

[1]The theoretic foundation of choosing these values is due to [1] in that these values offer robustness against outliers, but yet produce 95% efficiency when the errors are normal. These values are default values in commercial software e.g. Matlab statistics toolbox by *The MathWorks* and S-PLUS of *Insightful*. Readers are referred to references within the documentations available at the web sites of these companies.

and

$$\omega_B^{(m)}(t) = \begin{cases} \left[1 - \left(\frac{\xi^{(m-1)}(t)}{\tau}\right)^2\right]^2, & \text{for } \left|\xi^{(m-1)}(t)\right| \le \tau \\ 0, & \text{for } \left|\xi^{(m-1)}(t)\right| > \tau \end{cases} \quad (18)$$

respectively. Let $\Omega^{(m)} = \text{diag}\{\omega^{(m)}(1), \omega^{(m)}(2), \ldots \omega^{(m)}(N)\}$; then

$$\hat{\Theta}_M^{(m)} = \left\{\mathbf{P}^T\Omega^{(m)}\mathbf{P}\right\}^{-1}\mathbf{P}^T\Omega^{(m)}\mathbf{y} \quad (19)$$

$$\Xi^{(m)} = \mathbf{y} - \mathbf{P}\hat{\Theta}_M^{(m)} \quad (20)$$

where $\Xi^{(m)} = [\xi^{(m)}(1), \cdots, \xi^{(m)}(N)]^T$ are ready for next iteration step. The above procedure iterates until the parameter estimator $\hat{\Theta}_M$ converges at $m = m_w$.

$$\hat{\Theta}_M = \left\{\mathbf{P}^T\Omega^{(m_w)}\mathbf{P}\right\}^{-1}\mathbf{P}^T\Omega^{(m_w)}\mathbf{y}. \quad (21)$$

The asymptotic covariance matrix of $\hat{\Theta}_M$ is given by [1]

$$\text{var}[\hat{\Theta}_M] = \frac{E(\psi^2)}{E^2\left[\frac{d\psi}{d\xi}\right]}(\mathbf{P}^T\mathbf{P})^{-1}. \quad (22)$$

From (22), it is seen that the efficiency of the M–estimator depends on the full rank of the $(\mathbf{P}^T\mathbf{P})^{-1}$. However this usually may not be true for an oversized $\mathbf{P}$, unless there is some robustness measure in place to select a parsimonious model structure. Robust model structure selection can be achieved via experimental design criteria that selects $\mathbf{P}_k^T\mathbf{P}_k$, where $\mathbf{P}_k$ is a subset of $\mathbf{P}$, if $\mathbf{P}^T\mathbf{P}$ is ill-conditioned. The basic OFR model structure detection algorithm using D-optimality [13] is initially given below, which will be incorporated in the proposed algorithm of Section III.

B. Model Structure Selection by D-Optimality

A significant advantage due to orthogonalization is that the contribution of model regressors to the model can be evaluated. The OFR estimator involves selecting a set of n_θ variables $\mathbf{p}_k = [p_k(1), \cdots, p_k(N)]^T$, $k = 1, \cdots, n_\theta$, from M regressors to form a set of orthogonal basis $\mathbf{w}_k$, $k = 1, \cdots, n_\theta$, in a forward regression manner. As the orthogonality property $\mathbf{w}_i^T\mathbf{w}_j = 0$ for $i \neq j$ holds, if (6) is multiplied by itself and then the time average is taken, the following equation is easily derived

$$\frac{1}{N}\mathbf{y}^T\mathbf{y} = \frac{1}{N}\sum_{k=1}^{M}\gamma_k^2\mathbf{w}_k^T\mathbf{w}_k + \frac{1}{N}\Xi^T\Xi. \quad (23)$$

The most relevant n_θ regressors can be forward selected according to the value of an error reduction ratio $[\text{ERR}]_k$ (defined as $\gamma_k^2\mathbf{w}_k^T\mathbf{w}_k/\mathbf{y}^T\mathbf{y}$, see [6]). At the kth selection, a candidate regressor is selected as the kth basis of the subset if it produces the largest value of $[\text{ERR}]_k$ from the remaining $(M - k + 1)$ candidates. This procedure can automatically select a subset of n_θ regressors to construct a parsimonious model. Equivalently, this procedure can be expressed as

$$J^{(k)} = J^{(k-1)} - \frac{1}{N}\gamma_k^2\kappa_k \quad (24)$$

where $J^{(0)} = \mathbf{y}^T\mathbf{y}$. At the kth forward regression stage, a candidate regressor is selected as the kth regressor if it produces the smallest $J^{(k)}$. Equation (24) can be modified to form an alternative model selective criterion to enhance model robustness. The D-optimality criterion [5] maximizes the determinant of the design matrix defined as $\mathbf{W}_k^T\mathbf{W}_k$

$$\max\left\{J_D = \det\left(\mathbf{W}_k^T\mathbf{W}_k\right) = \prod_{k=1}^{n_\theta}\kappa_k\right\} \quad (25)$$

where $\mathbf{W}_k \in \Re^{N\times n_\theta}$ denotes the resultant regression matrix, consisting of n_θ regressors selected from M regressors in $\mathbf{W}$. It can be easily verified that the selection of a subset of $\mathbf{W}_k$ from $\mathbf{W}$ is equivalent to the selection of the a subset of n_θ regressors from $\mathbf{P}$ [13]. In order to include D-optimality as a model selective criterion for improved model robustness, construct an augmented cost function as

$$\begin{aligned} J &= \frac{1}{N}\Xi^T\Xi + \alpha\log\left(\frac{1}{J_D}\right) \\ &= \frac{1}{N}\left(\mathbf{y}^T\mathbf{y} - \sum_{k=1}^{n_\theta}\gamma_k^2\kappa_k\right) + \alpha\sum_{k=1}^{n_\theta}\log\left[\frac{1}{\kappa_k}\right] \end{aligned} \quad (26)$$

where α is a positive small number. Note that this composite cost function simultaneously minimizes (24) and maximizes (25) [13]. Equation (26) can be directly incorporated into the OFR algorithm to select the most relevant kth regressor at the kth forward regression stage, via

$$J^{(k)} = J^{(k-1)} - \frac{1}{N}\gamma_k^2\kappa_k + \alpha\log\left[\frac{1}{\kappa_k}\right]. \quad (27)$$

At the kth forward regression stage, a candidate regressor is selected as the kth regressor if it produces the smallest $J^{(k)}$ and further reduction in $J^{(k-1)}$. Because $\log(1/J_D)$ is an increasing function if $\kappa_k < 1$, which is true for some $k > K$, the selection procedure will terminate if $J^{(k)} \geq J^{(k-1)}$ at the derived model size n_θ if an proper α is set. This is significant because this means that the proposed approach can detect a parsimonious model size in an automatic manner.

A method of orthgonalization as required in above procedure is the modified Gram–Schmidt procedure [6]. In [13], γ_k are derived as least squares estimates. In the following, we propose a new algorithm as how to incorporate D-optimality model structure selection with M-estimator (Section II-A). The basic idea is to extend the modified Gram–Schmidt algorithm to include, at every forward regression step, an IRLS procedure inner loop that derives the M-estimator for the auxiliary vector Γ.

III. Model Identification Algorithm Using Forward Regression With M-Estimation

The modified Gram–Schmidt procedure can be used to perform the orthogonalization and parameter estimation, usually with parameters derived as least squares parameters. In this section a new model identification algorithm that combines M-estimator with forward regression is introduced based on the modified Gram–Schmidt procedure. Geometrically the system output vector $\mathbf{y}$, is projected onto a set of orthogonal basis vectors, $\{\mathbf{w}_1, \ldots, \mathbf{w}_k, \ldots\}$. For the modified Gram-Schmidt algorithm, the model residual is decreased by projecting the system output vector $\mathbf{y}$ onto a new basis $\mathbf{w}_k$ at step k. Denote model residual vector as $\Xi_{(k)}$, where the subscript denotes forward regression step k. Initially model residuals $\Xi_{(0)}$ is $\mathbf{y}$. The procedure at forward regression step k, can be explicitly interpreted as fitting the previous model residual vector $\Xi_{(k-1)}$ (as derived from forward regression step $(k-1)$) using a single variable $\mathbf{w}_k$ to solve a new model residual vector $\Xi_{(k)}$. Because M-estimator can enhance model parameter robustness in bad data conditions such as outliers, the proposed algorithm in the following, as a variant of modified Gram–Schmidt procedure, include the IRLS inner loop so as to derive the M-estimators of the auxiliary vector Γ.

Starting from $k = 1$, the columns $\mathbf{p}_j$, $k+1 \le j \le M$ are made orthogonal to the kth column at the kth stage. The D-optimality criterion (27) for each of $\mathbf{p}_j$, $k+1 \le j \le M$ columns is evaluated, and the most relevant column is selected to be interchanged with the kth column. The M-estimator for the kth regressor (the selected regressor) is then derived, as shown below, via the proposed IRLS inner loop. The operation is repeated for $1 \le k \le n_\theta < (M-1)$.

Algorithm

1) Initially denote $\mathbf{p}_j^{(0)} = \mathbf{p}_j$, $1 \le j \le M$ and $\mathbf{P}^{(0)} = [\mathbf{p}_1^{(0)}, \cdots, \mathbf{p}_M^{(0)}]$, $\Xi_{(0)} = \mathbf{y}$, $J^{(0)} = \mathbf{y}^T\mathbf{y}$.

2) The kth stage of the forward regression selection procedure with D-optimality is given below. For $k \le j \le M$, compute

$$\gamma_{(k,j)} = \frac{\left(\mathbf{p}_j^{(k-1)}\right)^T \Xi_{(k-1)}}{\left(\mathbf{p}_j^{(k-1)}\right)^T \mathbf{p}_j^{(k-1)}}$$

$$\kappa_k^{(j)} = \left(\mathbf{p}_j^{(k-1)}\right)^T \mathbf{p}_j^{(k-1)} \tag{28}$$

$$J_j^{(k)} = J^{(k-1)} - \frac{1}{N}\left[\gamma_{(k,j)}\right]^2 \kappa_k^{(j)} + \alpha \log\left[\frac{1}{\kappa_k^{(j)}}\right] \tag{29}$$

3) Find

$$J^{(k)} = J_{j_k}^{(k)} = \min\left\{J_j^{(k)}, \quad k \le j \le M\right\} \tag{30}$$

The j_kth column of $\mathbf{P}^{(k-1)}$ is then interchanged with the kth column of $\mathbf{P}^{(k-1)}$, and the j_kth column of $\mathbf{A}$ up to the $(k-1)$th row is interchanged with the kth column of $\mathbf{A}$. For notational convenience, the resultant $\mathbf{P}^{(k-1)}$ is still be referred to as $\mathbf{P}^{(k-1)}$. This effectively selects the j_kth candidates as the kth regressor in the subset model. Then set $\gamma_k^{(1)} = \gamma_{(k,j_k)}$, and derive model residual vector as

$$\Xi_{(k)}^{(1)} = \Xi_{(k-1)} - \gamma_k^{(1)}\mathbf{w}_k \tag{31}$$

(*NB*. The objective of (28)-(30) is to realize the D-optimality selective criterion of (26).)

4)

$$\begin{aligned}
\mathbf{w}_k &= \mathbf{p}_k^{(k-1)} \\
\kappa_k &= \mathbf{w}_k^T\mathbf{w}_k \\
\alpha_{kj} &= \frac{\mathbf{w}_k^T\mathbf{p}_j^{(k-1)}}{\kappa_k}, \quad k+1 \le j \le M \\
\mathbf{p}_j^{(k)} &= \mathbf{p}_j^{(k-1)} - \alpha_{kj}\mathbf{w}_k, \quad k+1 \le j \le M
\end{aligned} \tag{32}$$

Denote $\mathbf{P}^{(k)} = [\mathbf{w}_1, \cdots, \mathbf{w}_k, \mathbf{p}_{k+1}^{(k)}, \cdots, \mathbf{p}_M^{(k)}]$.

5) The following IRLS algorithm inner loop which aims to derive either *Huber* or *bisquare* M-estimator for the kth element of the auxiliary vector Γ, which is initialized as $\gamma_k^{(1)} \ne 0$.

Iterated Re-weighted Least Squares (IRLS) Inner Loop

i. Initialize $m = 2$. Note that model residual vector is initialized as $\Xi_{(k)}^{(1)}$ from Step 2.

ii. For *Huber* M-estimator, set $\tau = \tau_{(k)}^H = 1.345\ \mathrm{std}(\Xi_{(k)}^{(m-1)})$, where $\mathrm{std}(\bullet)$ denotes standard deviation. Use (17) to construct

$$\Omega_H^{(m)} = \mathrm{diag}\left\{\omega_H^{(m)}\left(\xi_{(k)}^{(m-1)}(1)\right), \omega_H^{(m)}\left(\xi_{(k)}^{(m-1)}(2)\right) \ldots, \omega_H^{(m)}\left(\xi_{(k)}^{(m-1)}(N)\right)\right\}. \tag{33}$$

or for *bisquare* M-estimator, set $\tau = \tau_{(k)}^B = 4.685\ \mathrm{std}(\Xi_{(k)}^{(m-1)})$. Then use (18) to construct

$$\Omega_B^{(m)} = \mathrm{diag}\left\{\omega_B^{(m)}\left(\xi_{(k)}^{(m-1)}(1)\right), \omega_B^{(m)}\left(\xi_{(k)}^{(m-1)}(2)\right) \ldots, \omega_B^{(m)}\left(\xi_{(k)}^{(m-1)}(N)\right)\right\} \tag{34}$$

iii. Denote

$$\Omega^{(m)} = \begin{cases} \Omega_H^{(m)} & \text{for Huber M-estimator} \\ \Omega_B^{(m)} & \text{for } bisquare \text{ M-estimator} \end{cases} \tag{35}$$

and

$$\gamma_k^{(m)} = \frac{\mathbf{w}_k^T\Omega^{(m)}\Xi_{(k-1)}}{\mathbf{w}_k^T\Omega^{(m)}\mathbf{w}_k} \tag{36}$$

$$\Xi_{(k)}^{(m)} = \Xi_{(k-1)} - \gamma_k^{(m)}\mathbf{w}_k \tag{37}$$

where $\Xi_{(k)}^{(m)} = [\xi_{(k)}^{(m)}(1), \xi_{(k)}^{(m)}(2), \ldots, \xi_{(k)}^{(m)}(N)]^T$. (*NB*. The orthogonal forward regression can be explicitly interpreted as fitting the previous model residual vector $\Xi_{(k-1)}$ using the selected orthogonal basis $\mathbf{w}_k$. While $\gamma_k^{(1)}$ derived in Step 3 is associated with $\mathbf{w}_k$ as least squares parameter estimates, (36), (37) are the direct application of (19), (20) to derive Re-weighted least square parameter estimates for M-estimators.)

iv. If $\|\gamma_k^{(m)} - \gamma_k^{(m-1)}\| \ge \delta$, where δ is arbitrarily small number, then set $m = m+1$, and goto step ii. Otherwise, set $\Xi_{(k)} = \Xi_{(k)}^{(m)}$, $\gamma_k = \gamma_k^{(m)}$. Finish the IRLS inner loop.

6) update

$$J^{(k)} = J^{(k-1)} - \frac{1}{N}\gamma_k^2\kappa_k + \alpha \log\left[\frac{1}{\kappa_k}\right] \tag{38}$$

7) The procedure is monitored and terminated at the derived $k = n_\theta$ step, when $J^{(k)} \ge J^{(k-1)}$, for a predetermined $\alpha > 0$. Otherwise, set $k = k+1$, go to step 2.

8) The original model coefficient vector $\Theta = [\theta_1, \cdots, \theta_{n_\theta}]^T$ can then be calculated from $\mathbf{A}\Theta = \Gamma$ through back substitution.

Note that in OFR-based algorithms it is important to make a clear distinction between model selective criteria and the parameter estimation cost function. The model selective criteria is used to decide which term to be included into the model, and parameter estimation cost function is used to derive parameters for a given model. It is possible to incorporate M-estimator in model selective criterion by using IRLS loop at step 2, but this will be computationally expensive. The proposed algorithm can still detect ill-conditioning as a model term with deterioration in model conditioning will not be selected.

A. Relations to SVM Regression

The support vector machine (SVM) regression is an alternative robust modeling approach [14], [15] based on the following model structure:

$$y(t) = \phi\left(\mathbf{x}(t)\right)^T \varpi + b + \xi(t) \tag{39}$$

where ϖ, b are parameters, and $\phi(\mathbf{x}) \in \mathcal{F} \in \Re^{N_F}$ denotes a mapping from data $\mathbf{x}$ into a feature space $\mathcal{F}$. By *Mercers* condition [14], it is possible that the inner product in *some* feature space $\mathcal{F}$ can be efficiently represented by some kernel functions, given by

$$k\left(\mathbf{x}(i), \mathbf{x}(j)\right) = \phi(\mathbf{x}(i))^T \phi\left(\mathbf{x}(j)\right) \in \Re \tag{40}$$

in which i, j are data labels. Note that some RBF functions, e.g. Gaussian, belong to kernel functions family [15].

The ε-insensitive function is defined as

$$\rho_\varepsilon(\xi) = \begin{cases} 0, & \text{for } |\xi| \le \varepsilon \\ |\xi| - \varepsilon, & \text{for } |\xi| > \varepsilon \end{cases} \tag{41}$$

where ε is a predetermined nonnegative parameter that is effectively used in controlling model complexity. Based on the SRM principle the SVM regression modeling usually uses a composite functional

$$\|\varpi\|^2 + C \sum_{i=1}^{N} \rho_\varepsilon\left(\xi(t)\right) \tag{42}$$

as the objective function, where C is a predetermined smoothing parameter. The minimization of (42) based on (39) can be reformulated as a constrained convex quadratic programming problem to derive a "global" parameter optimal solution under the condition that both C and ε are appropriately chosen [14], [15]. Increasing ε will reduce final model size, but taking $\varepsilon = 0$, results in model size to be equal to number of data points N. To formulate a constrained convex quadratic programming problem [14], [15] some slack variables associated with the bounds of ρ_ε are initially introduced, followed by the reformulation of the functional (42) into a Langrangian form including some additional Langrangian parameters α_i, α_i^*, $i = 1, \ldots, N$ [14]. The derived SVM regression model is the optimal solution given by

$$y(t) = \sum_{i=1}^{N} \left(\alpha_i - \alpha_i^*\right) k\left(\mathbf{x}(t), \mathbf{x}(i)\right) + b + \xi(t) \tag{43}$$

in which α_i, α_i^* are the derived Langrangian parameters. The simultaneous parameter and structure identification can be achieved because some of $(\alpha_i - \alpha_i^*)'s$ are derived as zeros, by controlling the size of ε. The SVM usually generates excellent model, but the computation expense is much higher than that of OFR algorithm [16]. By comparing (43) with (1), it is seen that the SVM is a linear-in-the-parameters model with the parameter associated with each kernel as $(\alpha_i - \alpha_i^*)$. However in (1) the basis functions $p_k(\bullet)$ are not restricted to be kernel functions.

Various cost functions described in this paper are shown in Fig. 1. It is seen from Fig. 1(b) and (c) that the ε-insensitive function is very similar to *Huber*'s cost function, because both are based on l_1 norm of errors for larger errors. Note that the ε-insensitive function is not a smooth function, and cannot be fitted into M-estimator family, in which the derivative information of the loss function is basic information in evaluating the robustness of M-estimator. It is also not suitable for many optimization procedures, which require derivative information, to be applied. Instead of using ε-insensitive cost function, the Huber cost function has been used in support vector machine (SVM) regression that is solved by using a recurrent neural network [19].

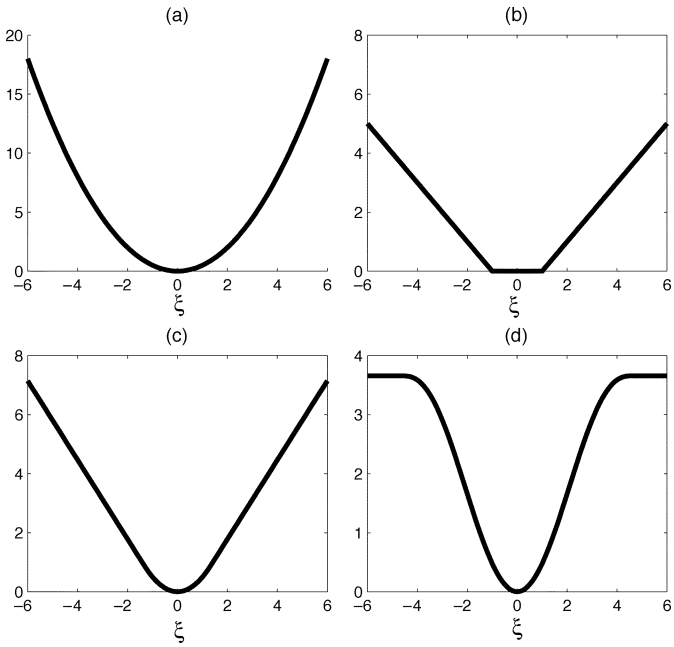

Fig. 1. Cost functions for different estimators. (a) Least squares. (b) Vapnik's ε-insensitive. (c) Huber and (d) Turkey Bisquare.

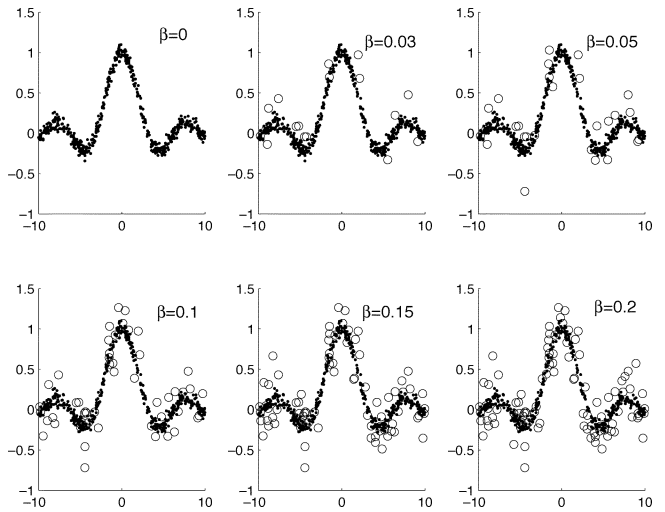

Fig. 2. Data generated by "sinc" function with additive noise of various levels of outliers in Example 1; (dotted-$N(0, 0.05^2)$, (normal), and circle-$N(0, 0.2^2)$ (outliers).

IV. Numerical Examples

Example 1: Consider using an RBF network to approximate the "sinc" function

$$z(x) = \frac{\sin(x)}{x}, \quad -10 \le x \le 10. \tag{44}$$

TABLE I
RMS ERRORS AND MODEL SIZE OF DERIVED MODELS WITH RESPECTIVE TO TRUE FUNCTION z (EXAMPLE 1)

		β					
		0	0.03	0.05	0.10	0.15	0.20
OFR with D-optimality and least squares	Training set	0.0102	0.0138	0.0143	0.0157	0.0175	0.0249
	Test set	0.0102	0.0135	0.0139	0.0158	0.0175	0.0254
	Model size	22	22	22	22	22	21
OFR with D-optimality and Huber M-estimator	Training set	0.0131	0.0139	0.0141	0.0129	0.0140	0.0219
	Test set	0.0131	0.0135	0.0136	0.0126	0.0137	0.0219
	Model size	22	22	22	22	22	21
OFR with D-optimality and Bisquare M-estimator	Training set	0.0128	0.0131	0.0137	0.0124	0.0135	0.0218
	Test set	0.0128	0.0128	0.0132	0.0121	0.0133	0.0217
	Model size	22	22	22	22	22	21
Support vector regression (SVR)	Training set	0.0144	0.0151	0.0156	0.0165	0.0176	0.0190
	Test set	0.0158	0.0156	0.0162	0.0169	0.0181	0.0192
	Model size	123	134	142	150	161	176

1000 training data $y(x)$ were generated from $y(x) = z(x) + \xi$, using uniformly distributed random $x \in [-10, 10]$. The additive noise ξ is a Gaussian mixture that mixes two types of noises, a larger portion of normal noise with smaller variance and a smaller portion of noise with higher variance. i.e. $\xi \sim \beta N(0, 0.2^2) + (1-\beta)N(0, 0.05^2)$, where $0 < \beta < 0.2$ as a small number to denote the contamination ratio, such that ξ has the probability $(1-\beta)$ of being drawn from $N(0, 0.05^2)$ (as "normal"), and a probability β of $N(0, 0.2^2)$ (as "outliers").

For various levels of contamination ratio β, 1000 noisy observations were generated and divided into a training data set of 500 data points and a test data set of 500 data points. The 500 training data points is shown in Fig. 2 for different β. For each case, the proposed algorithm is applied based on the RBF network. All the training data points are used as the candidate center set c_is, with $p_k(\mathbf{x}(t))$ constructed using Gaussian function $p_k = \phi(x, c_k) = \exp\{-\|x - c_k\|^2/h^2\}$. The width $h = 1$ is fixed for simplicity. Note that by removing the IRLS inner loop of the algorithm, the procedure simply reduces to OFR with D-optimality algorithm [13]. For comparison the SVM regression approach was applied, with the same Gaussian function $k(\mathbf{x}, c_k) = \exp\{-\|x - c_k\|^2/h^2\}$, and $h = 1$, as kernels. The parameters in SVM regression was set as $\varepsilon = 0.06$ and $C = 1$, as these values give the best tradeoff between model sparseness and generalization by trial and error.

With various values of β as different level of bad data conditions, the proposed algorithm is compared with OFR with D-optimality algorithm using only least squares estimates and SVM regression. With a predetermined small number $\alpha = 0.001$, all of the derived models based on OFR algorithm have the number of centers in the range of $n_\theta = 21 \sim 22$, but the model size of SVM regression range from $n_\theta = 123 \sim 176$. The root of mean squares (RMS) errors of a range of data conditions are listed in Table I. It is seen that the proposed algorithm is most robust to outliers when the data contains approximately 10% outliers. To achieve better performance for M-estimators, it is useful to slightly adjust tuning constants because these are set for 95% efficiency when data is normal. As data distribution is unknown these values can be adjusted via iterations and cross-validation. For the training data set as shown in Fig. 2 with $\beta = 0.1$, the model predicted output by using the proposed algorithm with Turkey bisquare M-estimators is shown in Fig. 3. For the best results in term of sparseness and efficiency, OFR based algorithms are better than SVM regression. The SVM regression is very robust even with worst data condition, and gives consistent results for all data conditions.

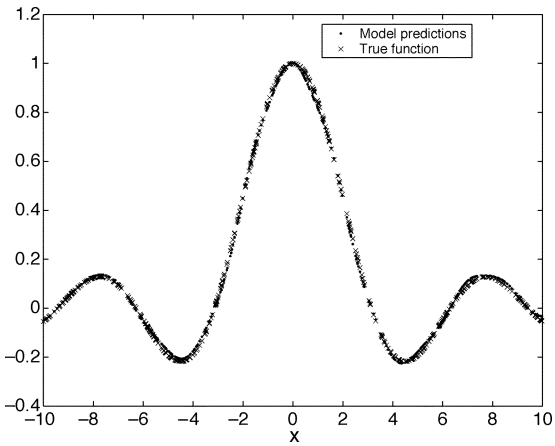

Fig. 3. Bisquare M-estimator model predictions with $\beta = 0.1$ and true functions (Example 1).

Example 2: Automobile MPG data. This data concerns city cycle fuel consumption in miles per gallon (MPG) (ftp.ics.uci.edu/pub/machine-learning-databases) 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, horse power, weight, acceleration, and model year. In a previous study [20], it has been shown that three inputs (horse power, weight and model year) are significant in modeling MPG. These three inputs are used in this study. In order to test the robustness of the proposed algorithm, a comparison study was performed based on modeling the original data, and data with some added outliers respectively. For each data point, with a probability 10%, a Gaussian noise with zero mean and standard deviation of 15, was randomly generated and added to the data to form the contaminated data, if the contaminated data is greater than 10. (any outlier below the minimum MPG are removed to generate more feasible outliers). The data was plotted in Fig. 4. With the input vector $\mathbf{x} = [\text{horse power},\ \text{weight},\ \text{model year}]$, all the training data points were used as the candidate center set $\mathbf{c}_i$s. The standard deviations of three inputs: {horse power, weight and model year} are $\{38.4912, 849.4026, 3.6837\}$, respectively. To achieve a balanced scale for each input, $p_k(\mathbf{x}(t))$ was constructed using Gaussian function $p_k = \phi(\mathbf{x}, c_k) = \exp\{-(\mathbf{x} - \mathbf{c}_k)^T \text{diag}\{50, 500, 5\}(\mathbf{x} - \mathbf{c}_k)\}$,

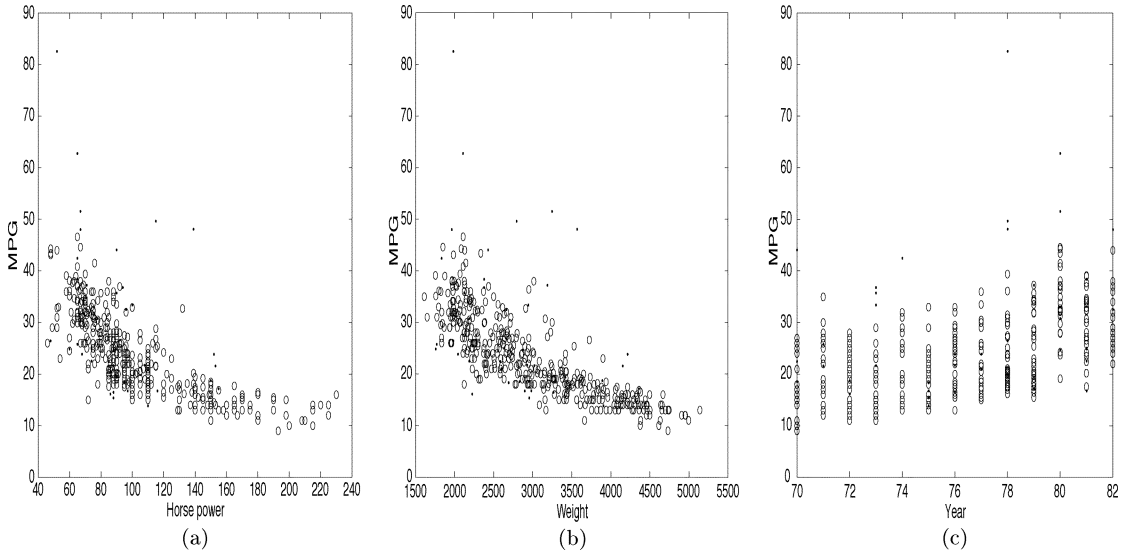

Fig. 4. Automobile MPG data (circle: original data; and dotted: synthetical outliers).

TABLE II
RMS ERRORS AND MODEL SIZE OF DERIVED MODELS WITH RESPECTIVE TO MPG DATA OUTPUT (EXAMPLE 2)

		Original data	Data with outliers
OFR with D-optimality	RMS	2.2982	2.8781
and least squares	Model size	56	65
OFR with D-optimality	RMS	2.4761	2.5818
and Huber M-estimator	Model size	45	69
OFR with D-optimality	RMS	2.5664	2.5115
and Bisquare M-estimator	Model size	45	71
Support vector	RMS	2.5639	2.6613
machine regression	Model size	94	116

i.e. each input has a width with a similar scale of its standard deviation. For both the original data and modified data, four types of algorithms were applied and the modeling results were listed in Table II. In the modeling original data without outliers, a predetermined small number $\alpha = 0.01$ was used for all OFR based algorithms. In the modeling data with outliers, it was found that the setting of $\alpha = 0.01$ would terminate at a model with too small model size with insufficient approximation accuracy, so $\alpha = 0.0001$ was used to allow larger models and better approximation accuracy. In the SVM regression approach, the same Gaussian function $k(\mathbf{x}, \mathbf{c}_k) = \exp\{-(\mathbf{x} - c_k)^T diag\{50, 500, 5\}(\mathbf{x} - \mathbf{c}_k)\}$ was used, and the parameters in SVM regression was set as $\varepsilon = 3$ and $C = 10$. It is shown that the proposed algorithm and SVM are robust to outliers than least square parameter estimates, in the sense that they are less vulnerable to the change in data's deviation from its original data. The OFR based algorithms produce more sparse models than that of SVM regression for all cases.

V. CONCLUSIONS

In this correspondence, a new OFR model identification algorithm is introduced. The OFR, often based on the modified Gram–Schmidt procedure, is an efficient method incorporating structure selection and parameter estimation simultaneously. The proposed algorithm includes M-estimator by using an IRLS algorithm inner loop based on the modified Gram–Schmidt procedure. D-optimality as a model structure robustness criterion is used in model selection. In this manner the proposed approach extends the use of the OFR algorithm for parsimonious model structure determination even in bad data conditions via the derivation of parameter M–estimators with inherent robustness to outliers. Numerical examples have shown that the proposed algorithm has improved performance than OFR with least squares parameters as data condition deteriorates.

ACKNOWLEDGMENT

The authors wish to thank the referees for their suggestions which greatly improved the quality of this paper.

REFERENCES

[1] P. J. Huber, *Robust Statistics*. New York: Wiley, 1981.
[2] J. T. Connor and R. D. Martin, "Recurrent neural networks and robust time series prediction," *IEEE Trans. Neural Netw.*, vol. 5, no. 2, pp. 240–253, Mar. 1994.
[3] J. H. Chen, C. S. Chen, and Y. S. Chen, "Fast algorithm for robust template matching with m-estimators," *IEEE Trans. Signal Process.*, vol. 51, no. 1, pp. 230–243, Jan. 2003.
[4] A. B. Hamza, H. Krim, and G. B. Unal, "Unifying probabilistic and variational estimation," *IEEE Signal Process. Mag.*, vol. 19, pp. 37–47, Sep. 2002.
[5] A. C. Atkinson and A. N. Donev, *Optimum Experimental Designs*. Oxford, UK: Clarendon Press, 1992.
[6] S. Chen, S. A. Billings, and W. Luo, "Orthogonal least squares methods and their applications to nonlinear system identification," *Int. J. Contr.*, vol. 50, pp. 1873–1896, 1989.
[7] S. Chen, Y. Wu, and B. L. Luk, "Combined genetic algorithm optimization and regularized orthogonal least squares learning for radial basis function networks," *IEEE Trans. Neural Netw.*, vol. 10, no. 6, pp. 1239–1243, Nov. 1999.
[8] M. J. L. Orr, "Regularization in the selection of radial basis function centers," *Neural Computat.*, vol. 7, no. 3, pp. 954–975, 1995.
[9] X. Hong, M. Brown, S. Chen, and C. J. Harris, "Sparse model identification using orthogonal forward regression with basis pursuit and [d]-optimality," IEE Proc. Contr. Theory Applicat., vol. 151, no. 4, pp. 491–498, 2004, submitted for publication.
[10] S. Chen, X. Hong, and C. J. Harris, "Sparse kernel regression modeling using combined locally regularised orthogonal least squares and d-optimality experimental design," *IEEE Trans. Autom. Contr.*, vol. 48, no. 6, pp. 1029–1036, Jun. 2003.
[11] X. Hong, C. J. Harris, S. Chen, and P. M. Sharkey, "Robust nonlinear model identification methods using forward regression," *IEEE Trans. Syst., Man, Cybern., A, Syst. Humans*, vol. 33, no. 4, pp. 514–523, Jul. 2003.

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

[13] ——, "Nonlinear model structure design and construction using orthogonal least squares and d-optimality design," *IEEE Trans. Neural Netw.*, vol. 13, no. 5, pp. 1245–1250, Sep. 2001.

[14] V. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer-Verlag, 1995.

[15] K. R. Müller, S. Mika, G. Rätsch, K. Tsuda, and B. Schölkopf, "An introduction to kernel-based learning algorithms," *IEEE Trans. Neural Netw.*, vol. 12, no. 2, pp. 181–201, Mar. 2001.

[16] K. L. Lee and S. A. Billings, "Time series prediction using support vector machines, the orthogonal and the regularized orthogonal least squares algorithms," *Int. J. Syst. Sci.*, vol. 33, no. 10, pp. 811–821, 2002.

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

[18] M. Brown and C. J. Harris, *Neurofuzzy Adaptive Modeling and Control*. Englewood Cliffs, NJ: Prentice-Hall, 1994.

[19] Y. Xia and J. Wang, "A one-layer recurrent neural network for support vector machine learning," *IEEE Trans. Syst., Man, Cybern., B, Cybern.*, vol. 34, no. 2, pp. 1261–1269, Apr. 2004.

[20] S. R. Gunn, M. Brown, and K. Bossley, "Network performance assessment for neurofuzzy data modeling," in *Intelligent Data Analysis*. Berlin, Germany: Springer-Verlag, 1997, pp. 313–323.