A Forward Regression Algorithm based on M-estimators

XIA HONG[†], SHENG CHEN[§]
Department of Cybernetics
University of Reading
Reading, RG6 6AY, UK [†]

School of Electronics and Computer Science
University of Southampton
Southampton, SO17 1BJ, UK [§]

x.hong@reading.ac.uk

Abstract: This paper introduces an orthogonal forward regression (OFR) model structure selection algorithm based on the M-estimators. 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 is extended to bad data conditions with improved performance due to M-estimators' inherent robustness to outliers. An illustrative example is included to demonstrate the effectiveness of the proposed algorithm.

Key-Words: System identification, M-estimation, forward regression, parameter estimation, robustness

1 Introduction

The orthogonal forward regression (OFR) is an efficient algorithm to determine a parsimonious model structure [1]. Driven by requirements for improved model generalization, a few variants of OFR have been introduced in order to tackle ill-conditioning problem by modifying the selective criteria in forward regression [2]-[3]. 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, or are heavy tailed compared to normal distribution. As a generalization of maximum-likelihood estimation method for data with outliers, the general method of M-estimation [4] is well established to tackle outliers in observational data. 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 [5, 6, 7]. This paper presents a new model identification algorithm that combines the M-estimator with forward regression. Based on the modified Gram-Schmidt procedure for orthogonal forward regression (OFR), the proposed algorithm incorporates an IRLS inner loop into the modified Gram-Schmidt procedure to derive a M-estimator of model parameters. In combination with D-optimality for model structure selection[3], the proposed algorithm simultaneously derive robust model structure and parameter estimates for bad data conditions.

2 Preliminaries

A linear regression model (RBF neural network, B-spline neurofuzzy network) can be formulated as [8, 9]

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

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 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.

Eq.(1) can be written in the matrix form as

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

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

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

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

$$\mathbf{P} = \mathbf{WA} \tag{3}$$

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

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

with

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

so that (2) can be expressed as

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

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 [1], 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 [10], experimental design and leave-one-out (LOO) score respectively [11, 12].

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. The D-optimality criterion [13] maximizes the determinant of the design matrix defined as $\mathbf{W}_k^T\mathbf{W}_k$,

$$\max\{J_D = \det(\mathbf{W}_k^T\mathbf{W}_k) = \prod_{k=1}^{n_\theta} \kappa_k\} \tag{7}$$

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 the a subset of $\mathbf{W}_k$ from $\mathbf{W}$ is equivalent to the selection of the a subset of n_θ regressors from $\mathbf{P}$ [3].

In this study we are concerned about model construction from data exhibiting bad conditions such as outliers. The general method of tackling this problem is well established as M-estimation [4], which is a generalization of maximum-likelihood estimation method for data with outliers. The M-estimator [4] is described in the following section.

2.1 M-estimators

The M-estimators have been well studied [4]. Considering the linear regression model given by (1), M-estimator minimizes the cost function

$$V_M = \sum_{t=1}^{N} \rho(\xi(t)) \tag{8}$$

where the function $\rho(\xi(t))$ is some predetermined nonnegative 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 $\xi(t)$, assuming that if any outlier has yet been detected and removed in the estimation data set. 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 [4]

$$\rho_H(\xi) = \begin{cases} \frac{1}{2}\xi^2 & \text{for } |\xi| \le \tau \\ \tau|\xi| - \frac{1}{2}\tau^2 & \text{for } |\xi| > \tau \end{cases} \quad (9)$$

or the Turkey *bisquare* estimator, given by

$$\rho_B(\xi) = \begin{cases} \frac{\tau^2}{6}\{1 - [1 - (\frac{\xi}{\tau})^2]^3\} & \text{for } |\xi| \le \tau \\ \frac{1}{6}\tau^2 & \text{for } |\xi| > \tau \end{cases} \quad (10)$$

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. These values offer robustness against outliers, but yet produce 95% efficiency when the errors are normal [4].

The M-estimator can be derived by setting

$$\frac{\partial V_M}{\partial \Theta}|_{\Theta=\hat{\Theta}_M} = \mathbf{0} \quad (11)$$

to yield

$$\frac{\partial V_M}{\partial \Theta} = \mathbf{P}^T \boldsymbol{\psi} = \mathbf{0} \quad (12)$$

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

$$\begin{aligned} \boldsymbol{\psi} &= [\frac{\partial V_M}{\partial \xi(1)}, ..., \frac{\partial V_M}{\partial \xi(N)}]^T \\ &= [\psi(\xi(1)), ..., \psi(\xi(N))]^T \end{aligned} \quad (13)$$

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

$$\omega(t) = \frac{\psi(\xi(t))}{\xi(t)}, \quad \text{for} \quad t = 1, ..., N. \quad (14)$$

Equation (12) can be written as

$$\mathbf{P}^T \Omega \Xi = \mathbf{0} \quad (15)$$

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

$$\hat{\Theta}_M = \{\mathbf{P}^T \Omega \mathbf{P}\}^{-1} \mathbf{P}^T \Omega \mathbf{y} \quad (16)$$

Because $\omega(t)$'s are *a prior* 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 $m = 2, ..., m_w$,

$$\omega^{(m)}(t) = \frac{\psi(\xi^{(m-1)}(t))}{\xi^{(m-1)}(t)}, \quad \text{for} \quad t = 1, ..., N. \quad (17)$$

From (9) and (10), the weight functions of *Huber* and the Turkey *bisquare* estimator can be explicitly given by

$$\omega_H^{(m)}(t) = \begin{cases} 1 & \text{for } |\xi^{(m-1)}(t)| \le \tau \\ \frac{\tau}{|\xi^{(m-1)}(t)|} & \text{for } |\xi^{(m-1)}(t)| > \tau \end{cases} \quad (18)$$

and

$$\omega_B^{(m)}(t) = \begin{cases} [1 - (\frac{\xi^{(m-1)}(t)}{\tau})^2]^2 & \text{for } |\xi^{(m-1)}(t)| \le \tau \\ 0 & \text{for } |\xi^{(m-1)}(t)| > \tau \end{cases} \quad (19)$$

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

$$\hat{\Theta}_M^{(m)} = \{\mathbf{P}^T \Omega^{(m)} \mathbf{P}\}^{-1} \mathbf{P}^T \Omega^{(m)} \mathbf{y} \quad (20)$$

$$\Xi^{(m)} = \mathbf{y} - \mathbf{P}\hat{\Theta}_M^{(m)} \quad (21)$$

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 $m = m_w$.

$$\hat{\Theta}_M = \{\mathbf{P}^T \Omega^{(m_w)} \mathbf{P}\}^{-1} \mathbf{P}^T \Omega^{(m_w)} \mathbf{y} \quad (22)$$

3 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, $\{\mathbf{w}_1, ..., \mathbf{w}_k, ...\}$. 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 this work is a variant of modified Gram-Schmidt procedure that includes the IRLS inner loop so as to derive the M-estimators of the auxiliary vector Γ.

Starting from $k = 1$, the columns $\mathbf{p}_j$, $k+1 \leq j \leq M$ are made orthogonal to the kth column at the kth stage. The D-optimality criterion [3] for each of $\mathbf{p}_j$, $k+1 \leq j \leq M$ 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 Re-weighted least squares (IRLS) inner loop. The operation is repeated for $1 \leq k \leq n_\theta < (M-1)$.

The following IRLS algorithm inner loop aims to derive either *Huber* or *bisquare* M-estimator for the kth element of the auxiliary vector Γ, which is initialized as the ordinary least squares parameter estimator $\gamma_k^{(1)} = \frac{\mathbf{w}_k^T \Xi_{(k-1)}}{\mathbf{w}_k^T \mathbf{w}_k} \neq 0$.

Iterated Re-weighted least squares (IRLS) inner loop:

i. Initialize $m = 2$. Note that model residual vector is initialized as $\Xi_{(k)}^{(1)}$ based on the parameter $\gamma_k^{(1)}$.

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 (18) to construct

$$\begin{aligned}\Omega_H^{(m)} &= \mathrm{diag}\{\omega_H^{(m)}(\xi_{(k)}^{(m-1)}(1)), \omega_H^{(m)}(\xi_{(k)}^{(m-1)}(2)) \\ &\quad , ..., \omega_H^{(m)}(\xi_{(k)}^{(m-1)}(N))\}. \end{aligned} \tag{23}$$

or for *bisquare* M-estimator, set $\tau = \tau_{(k)}^B = 4.685\ \mathrm{std}(\Xi_{(k)}^{(m-1)})$. Then use (19) to construct

$$\begin{aligned}\Omega_B^{(m)} &= \mathrm{diag}\{\omega_B^{(m)}(\xi_{(k)}^{(m-1)}(1)), \omega_B^{(m)}(\xi_{(k)}^{(m-1)}(2)) \\ &\quad , ..., \omega_B^{(m)}(\xi_{(k)}^{(m-1)}(N))\} \end{aligned} \tag{24}$$

iii. Denote

$$\Omega^{(m)} = \begin{cases} \Omega_H^{(m)} & \text{for Huber M-estimator} \\ \Omega_B^{(m)} & \text{for } \textit{bisquare} \text{ M-estimator} \end{cases} \tag{25}$$

and

$$\gamma_k^{(m)} = \frac{\mathbf{w}_k^T \Omega^{(m)} \Xi_{(k-1)}}{\mathbf{w}_k^T \Omega^{(m)} \mathbf{w}_k} \tag{26}$$

$$\Xi_{(k)}^{(m)} = \Xi_{(k-1)} - \gamma_k^{(m)} \mathbf{w}_k \tag{27}$$

where $\Xi_{(k)}^{(m)} = [\xi_{(k)}^{(m)}(1), \xi_{(k)}^{(m)}(2), ..., \xi_{(k)}^{(m)}(N)]^T$.

(*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 $\gamma_k^{(1)}$ is derived as least squares parameter estimates associated with $\mathbf{w}_k$, (26)-(27) are the direct application of (20)-(21) to derive Re-weighted least square parameter estimates for M-estimators.)

iv. If $\|\gamma_k^{(m)} - \gamma_k^{(m-1)}\| \geq \delta$, 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.

4 An illustrative example

Only a simple illustrative example is provided in this section, more experimental studies can be found in [14]. Consider using an RBF network to approximate the 'sinc' function

$$z(x) = \frac{\sin(x)}{x}, \qquad -10 \leq x \leq 10 \tag{28}$$

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

Table 1: RMS errors and model size of derived models with respective to true function z

		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

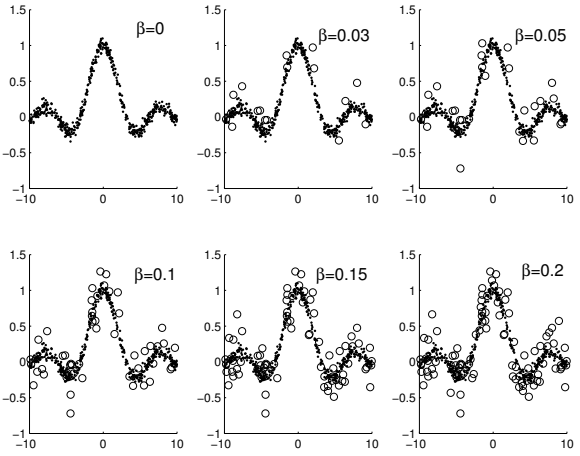

Figure 1: Data generated by 'sinc ' function with additive noise of various levels of outliers;(Dotted – $N(0, 0.05^2)$ (normal) and Circle – $N(0, 0.2^2)$ (outliers))

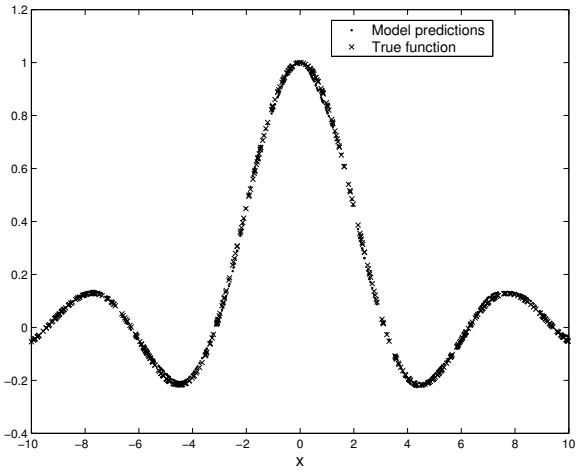

Figure 2: The Bisquare M-estimator model predictions with $\beta = 0.1$ and true functions.

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.1 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 centre set c_i's, 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 [3]. 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. All of the derived models based on OFR algorithm have the number of centers in the range of $n_\theta = 21 \sim 22$. The root of mean squares (RMS) errors of a range of data conditions are listed in Table 1. 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 with $\beta = 0.1$, the model predicted output by using the proposed algorithm with Turkey bisquare M-estimators is shown in Fig.2.

5 Conclusions

In this paper an orthogonal forward regression (OFR) model identification algorithm has been introduced. 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 proposed algorithm includes M-estimator by using an iterative re-weighted least squares (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.

Acknowledgement The authors gratefully acknowledge that part of this work was supported by EPSRC in the UK.

References

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

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

[3] Hong X. and Harris C.J. Nonlinear model structure design and construction using orthogonal least squares and D-optimality design. *IEEE Transactions on Neural Networks*, vol. 13(5), 2001, pp. 1245–1250.

[4] Huber P.J. *Robust Statistics*. J. Wiley, 1981.

[5] Connor J.T. and D.Martin R. Recurrent neural networks and robust time series prediction. *IEEE Transactions of Neural Networks*, vol. 5(2), 1994, pp. 240–253.

[6] Chen J.H., Chen C.S., and Chen Y.S. Fast algorithm for robust template matching with M-estimators. *IEEE Trans. on Signal Processing*, vol. 51(1), 2003, pp. 230–243.

[7] Hamza A.B., Krim H., and Unal G.B. Unifying probabilistic and variational estimation. *IEEE Signal Processing Magazine*, vol. 19(September), 2002, pp. 37–47.

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

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

[10] Hong X., Brown M., Chen S., and Harris C.J. Sparse model identification using orthogonal forward regression with basis pursuit and D-optimality. *IEE Proc. - Control Theory and Applications*, vol. 151(4), 2004, pp. 491–498.

[11] Chen S., Hong X., and Harris C.J. Sparse kernel regression modelling using combined locally regularised orthogonal least squares and D-optimality experimental design. *IEEE Trans. on Automatic Control*, vol. 48(6), 2003, pp. 1029–1036.

[12] Hong X., Harris C.J., Chen S., and Sharkey P.M. Robust nonlinear model identification methods using forward regresion. *IEEE Transactions on Systems, Man, and Cybernetics, Part A*, vol. 33(4), 2003, pp. 514–523.

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

[14] Hong X. and Chen S. M-estimator and D-optimality model construction using orthogonal forward regression. *IEEE Trans. Systems, Man and Cybernetics, Part B*, vol. 35(1), 2003, pp. 155–162.