ROBUST IDENTIFICATION FOR LINEAR-IN-THE-PARAMETERS MODELS

X. Hong [*] **C. J. Harris** [**] **S. Chen** [**] **P. M. Sharkey** [*]

[*] *Dept of Cybernetics, University of Reading, Reading RG6 6AY, UK*

[**] *Dept of Electronics and Computer Science, University of Southampton, Southampton, SO17 1BJ, UK*

Abstract: In this paper new robust nonlinear model construction algorithms for a large class of linear-in-the-parameters models are introduced to enhance model robustness, including three algorithms using combined A- or D-optimality or PRESS statistic (Predicted REsidual Sum of Squares) with regularised orthogonal least squares algorithm respectively. A common characteristic of these algorithms is that the inherent computation efficiency associated with the orthogonalisation scheme in orthogonal least squares or regularised orthogonal least squares has been extended such that the new algorithms are computationally efficient. A numerical example is included to demonstrate effectiveness of the algorithms. *Copyright ©2003 IFAC*

Keywords: experimental design, structure identification, forward regression, cross validation, generalisation.

1. INTRODUCTION

A large class of nonlinear models and neural networks can be classified as a linear-in-the-parameters model (Harris *et al.*, 2002; Wang and Mendel, 1992). The forward regression approach is an efficient model construction method (Chen *et al.*, 1989) for these models. Regularisation techniques have been incorporated into the orthogonal least squares (OLS) algorithm to produce a regularised orthogonal least squares (ROLS) algorithm that reduces the variance of parameter estimates (Chen *et al.*, 1999; Orr, 1995). To produce a model with good generalisation capabilities, model selection criteria such as the Akaike information criterion (AIC) (Akaike, 1974) are usually incorporated into the procedure to determinate the model construction process. Yet the use of AIC or other information based criteria, if used in forward regression, only affects the stopping point of the model selection, but does not penalise regressors that might cause poor model performance, if this is selected at an earlier regression stage.

Parameter regularisation and robust model structure selection are effective and complementary approaches for robust modelling. This paper reviews some recent advances on robust modelling techniques based on forward regression developed by the authors (Hong and Harris, 2001*b*; Hong and Harris, 2001*a*; Chen, 2002; Chen *et al.*, 2002; Hong *et al.*, 2002). These algorithms aim to achieve maximum model robustness by combining parameter regularisation and model structure selection via the direct optimisation of model robustness.

2. PRELIMINARIES

A linear-in-the-parameters model (RBF neural network, B-spline neurofuzzy network) can be formulated as (Harris *et al.*, 2002)

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

where $t = 1, 2, \cdots, N$, and N is the size of the estimation data set. $y(t)$ is system output variable, $\mathbf{x}(t) = [y(t-1), \cdots, y(t-n_y), u(t-1), \cdots, u(t-n_u)]^T$ is system input vector of observables with assumed known dimension of (n_y+n_u). $u(t)$ is system input variable. $p_k(\bullet)$ 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. Eq.(1) can be written in the matrix form as

$$\mathbf{y} = \mathbf{P}\mathbf{\Theta} + \Xi \qquad (2)$$

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

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

By setting a cost function of $J_1 = \sum_{t=1}^{N}(y(t) - \sum_{k=1}^{M} p_k(\mathbf{x}(t))\theta_k)^2$, the least squares estimates of $\mathbf{\Theta}$ is given by (Soderström and Stoica, 1989)

$$\hat{\mathbf{\Theta}} = (\mathbf{P}^T\mathbf{P})^{-1}\mathbf{P}^T\mathbf{y} \qquad (3)$$

Assume that Eq.(2) represents the data generating process. If $\mathbf{P}^T\mathbf{P}$ is nonsingular, then

$$\begin{aligned} &(i)\ E\hat{\mathbf{\Theta}} = \mathbf{\Theta} \\ &(ii)\ \mathrm{cov}\hat{\mathbf{\Theta}} = \sigma^2(\mathbf{P}^T\mathbf{P})^{-1} \end{aligned} \qquad (4)$$

where the matrix $(\mathbf{P}^T\mathbf{P})$ is called the design matrix. It is well known that a model based on least squares estimates tends to be unsatisfactory for a near ill conditioned regression matrix (or design matrix). The condition number of the design matrix is given by $C = \frac{\max \lambda_k}{\min \lambda_k}$, where $\lambda_k, (k = 1, \cdots, M)$ are the eigenvalues of the design matrix. Too large a condition number of the design matrix will result in unstable parameter estimates if a least squares algorithm is used (Harris *et al.*, 2002), whilst a small condition number of the design matrix leads to model robustness. Experimental design criteria of A-optimality and D-optimality (Atkinson and Donev, 1992) are introduced in Section 2.1, which provides a background for Section 3.1 and Section 3.2 for two model identification algorithms.

Alternatively, parameter estimates can be derived based on a regularised cost function of $J_r = \sum_{t=1}^{N}(y(t) - \sum_{k=1}^{M} p_k(\mathbf{x}(t))\theta_k)^2 + \sum_{k=1}^{M} \gamma_k \theta_k^2$, where $\gamma_k > 0$, $k = 1, 2, \cdots, M$ are regularisation parameters. The regularised least squares estimates of $\hat{\mathbf{\Theta}}_r$ is given by (Marquardt, 1970)

$$\hat{\mathbf{\Theta}}_r = (\mathbf{P}^T\mathbf{P} + \Gamma)^{-1}\mathbf{P}^T\mathbf{y} \qquad (5)$$

where $\Gamma = diag\{\gamma_1, \gamma_2, \cdots, \gamma_M\}$. The concept of parameter regulasation may be incorporated into a forward orthogonal least squares algorithm as a locally regularised orthogonal least square estimator (see Appendix A for details), which forms the foundation for all the robust identification algorithms introduced in this paper (see Section 3).

2.1 Optimal experimental design criteria

Consider a subset model is constructed from the full model with regression matrix $\mathbf{P}$ by using n_θ regressors selected from M regressors in $\mathbf{P}$, $n_\theta \ll M$. Denote the resultant regression matrix $\mathbf{P}_k \in \Re^{N \times n_\theta}$, the resultant design matrix by $\mathbf{P}_k^T\mathbf{P}_k$, and with λ_k, $k = 1, ..., n_\theta$ as the eigenvalues of $\mathbf{P}_k^T\mathbf{P}_k$.

Definition 1: A-optimality criterion: The A-optimality design criterion, which can be applied as a model selection criterion is that which minimises the sum of the variance of a parameter estimate vector $\hat{\Theta} = [\theta_1, \cdots, \theta_{n_\theta}]^T$

$$\min\{J_2 = \mathrm{tr}\left[\mathrm{cov}\hat{\mathbf{\Theta}}\right] = \sigma^2 \sum_{k=1}^{n_\theta} \frac{1}{\lambda_k}\} \qquad (6)$$

Alterntively the D-optimality design criterion can be applied as a model selection criterion that maximises the determinant of the design matrix of $\mathbf{P}_k^T\mathbf{P}_k$.

Definition 2: The D-optimality criterion is that which

$$\max\{J_3 = \det(\mathbf{P}_k^T\mathbf{P}_k) = \prod_{k=1}^{n_\theta} \lambda_k\} \qquad (7)$$

Maximisation of the D-optimality criterion (Atkinson and Donev, 1992) for model selection criterion inherently improves model robustness. Robust identification algorithms using the combined A-optimality and D-optimality with regularised orthogonal least squares are introduced in Section 3.1 and 3.2 respectively.

2.2 PRESS statistic

Cross validation criteria are metrics that measures a model's generalisation capability, which can alternatively be used as a model selection criterion for robustness. One commonly used version of cross-validation is the so called delete-1 cross-validation. The idea is that, for any model, each data point in the estimation data set $D_N = \{\mathbf{x}(t), y(t)\}_{t=1}^{N}$ is sequentially set aside in turn, a model is then estimated using the remaining $(N-1)$ data, and the prediction error is derived using only the data point that was removed. To select a model by using the delete-1 cross-validation as the model selective criterion, the model with a minimal mean squares of the prediction errors is selected. The prediction error known as the Predicted REsidual Sums of Squares (PRESS) statistic (Myers, 1990) for linear-in-the-parameters models, can be generated without actually sequentially splitting the estimation data set by using the Sherman-Morrison-Woodbury theorem (Myers, 1990). Consider a predictor that is identified based on (1), the PRESS errors $\xi^{(-t)}(t|t-1)$ can be calculated using (Myers, 1990) as

$$\begin{aligned}\xi^{(-t)}(t|t-1) &= y(t) - \hat{y}^{(-t)}(t|t-1)\\ &= \frac{\xi(t)}{1-\mathbf{p}(t)^T[\mathbf{P}^T\mathbf{P}]^{-1}\mathbf{p}(t)}\end{aligned} \quad (8)$$

and the PRESS statistic is computed by

$$J_p = E\left[[\xi^{(-t)}(t|t-1)]^2\right] \quad (9)$$

A robust identification algorithm using the PRESS statistic and regularised orthogonal least squares is introduced in Section 3.3.

3. ROBUST IDENTIFICATION FOR LINEAR-IN-THE-PARAMETERS MODELS

For simplicity of notation, as a function of forward regression step k, the resultant model selection criteria for all the proposed algorithms are denoted as $J^{(k)}$.

3.1 Combined A-optimality and ROLS

Consider the A-optimality design criterion given in Definition 1, but based on model (26) (Appendix A) with orthogonal basis $\mathbf{w}_k$. The A-optimaility cost function that minimises the sum of the variance of the auxiliary parameter estimate vector $\mathbf{g} = [g_1, \cdots, g_{n_\theta}]^T$ for a subset model with n_θ regressors is given by

$$\min\{J_A = \mathrm{tr}\,[\mathrm{cov}\hat{\mathbf{g}}] = \sigma^2 \sum_{k=1}^{n_\theta} \frac{1}{\kappa_k}\} \quad (10)$$

Due to $\mathbf{A}\Theta = \mathbf{g}$, it can be assumed that to penalize the large variance of the auxiliary parameter vector $\mathbf{g}$ will also consequently penalize large variance of parameter vector Θ.

A composite cost function is defined as

$$\begin{aligned}J &= J_1 + \alpha_1 J_A\\ &= \frac{1}{N}(\mathbf{y}^T\mathbf{y} - \sum_{k=1}^{n_\theta} g_k^2\kappa_k) + \alpha \sum_{k=1}^{n_\theta}\frac{1}{\kappa_k}\end{aligned} \quad (11)$$

where, for the sake of simplicity, $\alpha = \sigma^2\alpha_1$, is a positive small number. Eq.(11) can be directly incorporated into the conventional forward OLS algorithm to select the most relevant kth regressor at the kth forward regression stage, via

$$J^{(k)} = J^{(k-1)} - \frac{1}{N} g_k^2 \kappa_k + \frac{\alpha}{\kappa_k} \quad (12)$$

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 on $J^{(k-1)}$. The selection procedure will terminate if $J^{(k)} \geq J^{(k-1)}$ at the derived model size n_θ. This is significant because this means that the proposed approach can automatically detect a parsimonious model size.

The above A-optimality based design model construction algorithm was firstly introduced by the authors in outline in (Hong and Harris, 2001*b*) and applied as part of the B-spline based neurofuzzy model (NeuDec) (Hong and Harris, 2001*a*). It was shown in (Hong and Harris, 2001*b*; Hong and Harris, 2001*a*) that the resultant models can be improved based on the reduction of model parameter variance.

3.2 Combined D-optimality and ROLS

Consider the D-optimality design criterion given in Definition 2, but based on model (26) with orthogonal basis $\mathbf{w}_k$. The D-optimality design criterion that maximises the determinant of the design matrix of $\mathbf{W}_k^T\mathbf{W}_k$ is given by

$$\max\{J_{D_0} = \det(\mathbf{W}_k^T\mathbf{W}_k) = \prod_{k=1}^{n_\theta}\kappa_k\}. \quad (13)$$

The equivalence of (7) and (13) can be easily verified (Hong and Harris, 2002), and this implies that the selection of the a subset of $\mathbf{P}_k$ from $\mathbf{P}$ is equivalent to the selection of a subset of $\mathbf{W}_k$ from $\mathbf{W}$, or that a better conditioned $\mathbf{P}_k$ can be achieved via a better conditioned $\mathbf{W}_k$.

Construct the following cost function

$$J_D = \psi(J_{D_0}) = -\log(J_{D_0}) = \sum_{k=1}^{n_\theta}\log[\frac{1}{\kappa_k}] (14)$$

Clearly the maximisation of J_{D_0} is equivalent to the minimisation of $\psi(J_{D_0})$, due to the fact that the solution of $\partial\psi(J_{D_0}) = -\frac{1}{J_{D_0}}\partial J_{D_0} = 0$, is equivalent to that of $\partial J_{D_0} = 0$ for $J_{D_0} > 0$.

The new augumented cost function is defined as

$$\begin{aligned} J &= J_1 + \beta J_D \\ &= \frac{1}{N}(\mathbf{y}^T\mathbf{y} - \sum_{k=1}^{n_\theta} g_k^2 \kappa_k) + \beta \sum_{k=1}^{n_\theta} \log[\frac{1}{\kappa_k}] \quad (15) \end{aligned}$$

where β is a small positive number. Eq.(15) can be incorporated into the forward OLS algorithm to select the most relevant kth regressor at the kth forward regression stage, via

$$J^{(k)} = J^{(k-1)} - \frac{1}{N} g_k^2 \kappa_k + \beta \log[\frac{1}{\kappa_k}] \quad (16)$$

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 on $J^{(k-1)}$. Because 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 β is set.

The complete robust identification procedure using combined D-optimality and regularised orthogonal least squares based on the forward Gram-Schmidt procedure, including optimisation of regularisation parameters, can be found (Chen *et al.*, 2002), in which an effective Bayesian evidence method (MacKay, 1992) has been introduced to optimise local regularisation parameters.

3.3 Combined PRESS statistic and ROLS

Alternatively the PRESS statistic of (9) that optimises model generalation capability can be used as a robust model selective criterion. Note that (8) does not incorporate parameter regularisation. In order to combine the PRESS statistic into a model with regularisation and forward regression learning algorithm, initially it is necessary to derive the PRESS error in an orthogonal weight regularised model. It can be shown (Hong *et al.*, 2002) that the PRESS error, based on the system in the orthogonalised form (given by (26)), is given

$$\begin{aligned} \xi^{(-t)}(t|t-1) &= y(t) - \hat{y}^{(-t)}(t|t-1) \\ &= \frac{\xi(t)}{1 - \mathbf{w}(t)^T[\mathbf{W}^T\mathbf{W} + \Gamma]^{-1}\mathbf{w}(t)} \\ &= \frac{\xi(t)}{\eta_M(t)} \quad (17) \end{aligned}$$

where

$$\eta_M(t) = 1 - \sum_{i=1}^{M} \frac{w_i^2(t)}{\kappa_i + \gamma_i} \quad (18)$$

The computational expense can be further significantly reduced by utilising the forward regression process via a recursive formula. In the forward regression process, the model size is configured as a growing variable k. Consider the model construction by using a subset of k regressors ($k \ll M$), that is a subset selected from the full model set consisting of M initial regressors (given by (2)) to approximate the system. The PRESS errors (17)–(18) can be written, by replacing M with a variable model size k, as

$$\xi^{(-t)}(t|t-1) = \frac{\xi_k(t)}{\eta_k(t)} \quad (19)$$

where $\eta_k(t) = 1 - \sum_{i=1}^{k} \frac{w_i^2(t)}{\kappa_i + \gamma_i}$, and $\xi_k(t)$ is the model residual associated with a subset model structure with k regressors. $\eta_k(t)$ can be written as a recursive formula, given by

$$\eta_k(t) = \eta_{k-1}(t) - \frac{w_k^2(t)}{\kappa_k + \gamma_k} \quad (20)$$

This is advantageous in that, for a new model with size increased from $(k-1)$ to k, the PRESS error coefficient $\eta_k(t)$ needs only to be adjusted based on that of a model of size $(k-1)$, with a minimal computational effort.

As in conventional forward regression (Chen *et al.*, 1989), a Gram-Schmidt procedure is used to construct the orthogonal basis $\mathbf{w}_k$ in a forward regression manner. At each regression step, the PRESS statistic can be formed using the algorithm and this is then used as a regressor selective criteria for model construction that minimises the mean square PRESS errors

$$\begin{aligned} J^{(k)} &= E\left[[\xi^{(-t)}(t|t-1)]^2\right] = E[\frac{[\xi_k(t)]^2}{\eta_k^2(t)}] \\ &= \frac{1}{N}\sum_{t=1}^{N} \frac{[\xi_k(t)]^2}{\eta_k^2(t)} \quad (21) \end{aligned}$$

It can be analysed that due to the properties associated with the minimisation of the PRESS statistic, a fully automatic nonlinear predictive model contruction algorithm can be achieved (Analysis of the function $J^{(k)}$ shows that it is concave with respect to k (Hong *et al.*, 2002)). The complete robust identification procedure using combined PRESS statistic and regularised orthogonal least squares can be found in (Hong *et al.*, 2002).

4. ILLUSTRATIVE EXAMPLE

The robust algorithm introduced in Sec.3.3 is used only as illustration. For more examples on simulated data and practical implementation of these algorithms can be found in (Hong and Harris, 2001*b*; Hong and Harris, 2001*a*; Chen, 2002; Chen *et al.*, 2002; Hong *et al.*, 2002). Consider the following benchmark dynamic system given by (Narendra and Parthasarathy, 1990)

$$z(t) = \frac{z(t-1)z(t-2)z(t-3)u(t-2)[z(t-3)-1]+u(t-1)}{1+z^2(t-2)+z^2(t-3)} \tag{22}$$

where the system input $u(t)$ is given as a uniformly distributed random signal in the range $[-1,1]$. $y(x) = z(x) + \xi$, in which the noise $\xi \sim N(0, 0.05^2)$. 200 data points were generated. The input vector is predetermined as a 5-input vector as $\mathbf{x}(t) = [y(t-1), y(t-2), y(t-3), u(t-1), u(t-2)]^T$. The Gaussian function $\phi(x, c_i) = \exp\{-\|x-c_i\|^2/\tau^2\}$ is used as basis functions to construct an RBF model, with a width $\tau = 1$. All 200 training data points are used as the candidate centre set. The proposed combined PRESS statistic and ROLS of Section 3.3 was applied for automatic model structure detection, in which the regularisation parameter was set as $\gamma_i = 10^{-6}$, $\forall i$. A parsimonious model structure can be detected at a derived model size when the PRESS statistic achieves at a minimum. During the forward regression model construction process, the PRESS statistic gradually decreases until $n_\theta = 37$, with an increment of $\Delta J = 1.97 \times 10^{-7} > 0$, such that the model with 37 centres is automatically derived as the final model. The results of the derived RBF model with 37 centres, are shown in Fig.1. The model MSE and PRESS at $n_\theta = 37$, is 0.0995^2, and 0.11^2 respectively, demonstrating that the model is appropriate.

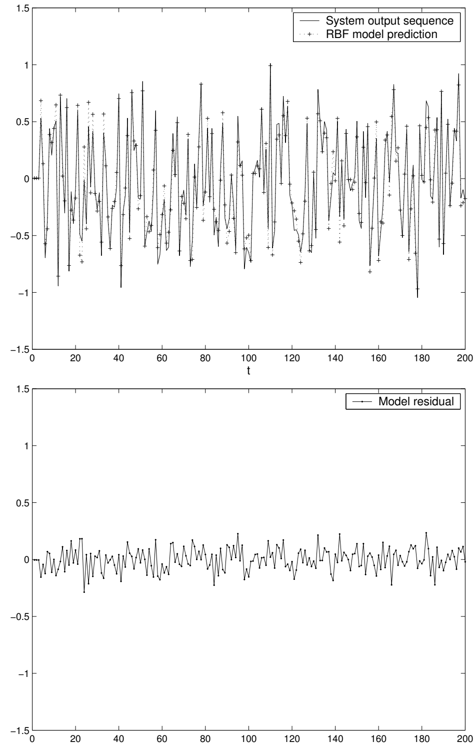

Fig. 1. Modelling results using RBF network with 37 centres.

5. CONCLUSIONS

In this paper, we have reviewed some recent advances in robust nonlinear modelling techniques in the framework of forward regression, that greatly enhance the well known forward orthogonal least squares (OLS) algorithm for model selection based on various robustness objectives.

REFERENCES

Akaike, H. (1974). A new look at statistical model identification. *IEEE Trans. on Automatic Control* **AC-19**, 716–723.

Atkinson, A. C. and A. N. Donev (1992). *Optimum Experimental Designs.* Clarendon Press, Oxford.

Chen, S. (2002). Locally regularization assisted orthogonal least squares regression. *IEEE Trans. on Neural Networks* p. Submitted.

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

Chen, S., X. Hong and C. J. Harris (2002). Sparse kernel regression modelling using combined locally regularised orthogonal least squares and d-optimality experimental design. *IEEE Trans. on Automatic Control* p. Submitted.

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

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

Hong, X. and C. J Harris (2001*a*). Neurofuzzy design and model construction of nonlinear dynamical processes from data. *IEE Proc. - Control Theory and Applications* **148**(6), 530–538.

Hong, X. and C. J Harris (2001*b*). nonlinear model structure detection using optimum experimental design and orthogonal least squares. *IEEE Transactions on Neural Networks* **12**(2), 435–439.

Hong, X. and C. J. Harris (2002). Nonlinear model structure design and construction using orthogonal least squares and d-optimality design. *IEEE Trans. on Neural Networks* p. Accepted.

Hong, X., S. Chen and P.M. Sharkey (2002). Automatic kernel regression modelling using combined press statistic and regularised orthogonal least squares. *IEE Proceedings - Vision, Image and Signal Processing* p. Submitted.

MacKay, D. J. C. (1992). Bayesian interpolation. *Neural Computation* **4**(3), 415–447.

Marquardt, D. W. (1970). Generalised inverse, ridge regression, biased linear estimation and nonlinear estimation. *Technometrics* **12**(3), 591–612.

Myers, R. H. (1990). *Classical and modern regression with applications.* 2nd Edition, PWS-KENT, Boston.

Narendra, K. S. and K. Parthasarathy (1990). Identification and control of dynamic systems using neural networks. *IEEE Trans. on Neural Networks* **1**(1), 4–27.

Orr, M. J. L. (1995). Regularisation in the selection of radial basis function centers. *Neural Computation* **7**(3), 954–975.

Soderström, T. and P. Stoica (1989). *System Identification.* Prentice Hall.

Wang, L. X. and J. M. Mendel (1992). Fuzzy basis functions, universal approximation, and orthogonal least squares learning. *IEEE Trans. on Neural Networks* **3**, 807–814.

APPENDIX A: LOCALLY REGULARISED ORTHOGONAL LEAST SQUARES

An orthogonal decomposition of $\mathbf{P}$ is

$$\mathbf{P} = \mathbf{W}\mathbf{A} \tag{23}$$

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

$$\mathbf{W}^T\mathbf{W} = diag\{\kappa_1, \cdots, \kappa_M\} \tag{24}$$

with

$$\kappa_k = \mathbf{w}_k^T\mathbf{w}_k, \qquad k = 1, \cdots, M \tag{25}$$

so that Eq.(2) can be expressed as

$$\mathbf{y} = (\mathbf{P}\mathbf{A}^{-1})(\mathbf{A}\Theta) + \Xi = \mathbf{W}\mathbf{g} + \Xi \tag{26}$$

where $\mathbf{g} = [g_1, \cdots, g_M]^T$ is an auxiliary vector. The LROLS algorithm uses the following error criterion for parameter estimation:

$$J_r = \Xi^T\Xi + \mathbf{g}^T\Gamma\mathbf{g} \tag{27}$$

Because $\xi(t)$ is uncorrelated with past output signals, it may be shown (Chen *et al.*, 1989) that

$$g_k = \frac{\mathbf{w}_k^T\mathbf{y}}{\mathbf{w}_k^T\mathbf{w}_k + \gamma_k}, \qquad k = 1, \cdots, M \tag{28}$$

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

The ROLS procedure can use the conventional OLS procedure for model term selection which maximises model approximation capability in a forward regression manner. The principle of the method is shown below. The number of all possible regressors M can be much larger than n_θ, but n_θ significant regressors can be identified using the forward OLS procedure. As the orthogonality property $\mathbf{w}_i^T\mathbf{w}_j = 0$ for $i \neq j$ holds, Eq.(26) is multiplied by itself and the time average is then taken, the following equation is easily derived

$$\frac{1}{N}\mathbf{y}^T\mathbf{y} = \frac{1}{N}\sum_{k=1}^{M} g_k^2\mathbf{w}_k^T\mathbf{w}_k + \frac{1}{N}\Xi^T\Xi \tag{29}$$

The output variance $E[y^2(t)] = \frac{1}{N}\mathbf{y}^T\mathbf{y}$ consists of two parts, $\frac{1}{N}\sum_{k=1}^{M} g_k^2\mathbf{w}_k^T\mathbf{w}_k$, the output variance explained by the regressors and $\frac{1}{N}\Xi^T\Xi$, the part of unexplained variance. The Error Reduction Ratio $[ERR]_k$, which is defined as the increment towards the overall output variance $E[y^2(t)]$ due to each regressor or input variable $p_k(t)$ divided by the overall output variance is computed through

$$[ERR]_k = \frac{g_k^2\mathbf{w}_k^T\mathbf{w}_k}{\mathbf{y}^T\mathbf{y}}, \qquad k = 1, \cdots, M \tag{30}$$

The most relevant n_θ regressors can be forward selected according to the value of the error reduction ratio $[ERR]_k$. At the kth selection, a candidate regressor is selected as the kth basis of the subset if it produces the largest value of $[ERR]_k$ from the remaining $(M - k + 1)$ candidates. By setting an appropriate tolerance ρ, which can be found by trial and error or via some statistical information criterion such as Akaike's information criterion(AIC) (Akaike, 1974) that forms a compromise between the model performance and model complexity, the variable selection is terminated when

$$1 - \sum_{k=1}^{n_\theta}[ERR]_k < \rho \tag{31}$$

This procedure can automatically select a subset of n_θ regressors to construct a parsimonious model. Equivalently, this procedure can be expressed as

$$J^{(k)} = J^{(k-1)} - \frac{1}{N}g_k^2\kappa_k \tag{32}$$

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 (32) is then used in the derivation of experimental design criteria based algorithms in Section 3.1 and Section 3.2.