Automatic Kernel Regression Modelling Using Combined Leave-One-Out Test Score and Regularised Orthogonal Least Squares

X. Hong[†], S. Chen[‡], P.M. Sharkey[†]

[†] Department of Cybernetics
University of Reading, Reading, RG6 6AY, UK
Email: x.hong@cyber.reading.ac.uk; WWW: http://www.cyber.rdg.ac.uk/CIRG/
Tel: +44 (0)118 9318222; Fax: +44 (0)118 9318220

[‡] School of Electronics and Computer Science
University of Southampton, Southampton SO17 1BJ, UK
Email: sqc@ecs.soton.ac.uk; WWW: http://www.ecs.soton.ac.uk
Tel: +44 (0)23 80596660; Fax: +44 (0)23 80594508

Abstract

This paper introduces an automatic robust nonlinear identification algorithm using the leave-one-out test score also known as the PRESS (Predicted REsidual Sums of Squares) statistic and regularised orthogonal least squares. The proposed algorithm aims to achieve maximised model robustness via two effective and complementary approaches, parameter regularisation via ridge regression and model optimal generalisation structure selection. The major contributions are to derive the PRESS error in a regularised orthogonal weight model, develop an efficient recursive computation formula for PRESS errors in the regularised orthogonal least squares forward regression framework and hence construct a model with a good generalisation property. Based on the properties of the PRESS statistic the proposed algorithm can achieve a fully automated model construction procedure without resort to any other validation data set for model evaluation.

Keywords — orthogonal forward regression, structure identification, cross validation, regularisation.

1 Introduction

A large class of nonlinear models and neural networks can be classified as a kernel regression model [1]. The orthogonal forward regression is an efficient model construction method [2] which selects regressors in a forward manner by virtue of their contribution to the maximisation of the model error reduction ratio. Regularisation techniques based on ridge regression [3] have been incorporated into the orthogonal least squares (OLS) algorithm to produce a regularised OLS (ROLS) algorithm that reduces the variance of parameter estimates [4, 5]. To produce a model with good generalisation capabilities, model selection criteria such as the Akaike information criterion (AIC) [6] are usually incorporated into the procedure to determine the model construction process. Yet the use of AIC or other information based criteria in forward regression only affects the stopping point of the model selection, but does not penalise the regressor that might cause poor model performance, e.g. too large parameter variance or ill-posedness of the regression matrix if such a regressor is selected. This is due to the fact that AIC or other information based criteria are usually simplified measures derived as an approximation formula that is particularly sensitive to model complexity.

In order to achieve a model structure with improved model generalisation, it is natural that a model generalisation capability cost function should be used in the overall model searching process, rather than only being applied as a measure of model complexity. Because the evaluation of the model generalisation capability is directly based on the concept of cross validation [7], it is highly desirable to develop new model selective criteria based on the fundamental concept of cross validation that can distinguish model generalisation capability during the model construction process. A fundamental concept in cross validation is that of delete-1 cross validation in statistics, and the associated concept of the leave-one-out test score also known as the PRESS (Predicted REsidual Sums of Squares) statistic [8, 9, 10, 11, 12]. The computation of the leave-one-out test score or PRESS statistic usually involves large computational expense. Recently an automatic nonlinear regression model construction algorithm has been introduced based on orthogonal forward regression and the PRESS statistic which can minimise this computational expense [13].

Because parameter regularisation and robust model structure selection are effective and complementary approaches for robust modelling, it is highly desirable to develop algorithms by combining parameter regularisation with model structure selection via a direct optimisation of model generalisation capability. Such a combined approach is capable of maximising model robustness. In this paper, an automatic nonlinear regression model construction algorithm is introduced based on the combined ROLS and PRESS statistic. In order to combine parameter regularisation with model structure selection based on the PRESS statistic, we initially derive the PRESS error in the regularised orthogonal weight model. Due to the inherent computation efficiency associated with forward regression based on the ROLS algorithm, the effort involved in the computation of the PRESS statistic is minimised. The key in improving computational efficiency is to utilise an inherent orthogonalisation process for avoiding a matrix inversion in the computation of the PRESS error. Further significant reduction in computation arises owing to the derivation of a forward recursive formula to compute PRESS errors. In the proposed algorithm, the PRESS statistic, which is a measure of model generalisation capability, is applied directly in the forward regression model structure construction process as a cost function in order to optimise the model generalisation capability. Based on the properties of the PRESS statistic, the proposed algorithm can achieve a fully automatic model selection procedure without resorting to another validation data set for model assessment. Two examples are included to demonstrate the effectiveness of the approach.

2 Regularised orthogonal least squares for kernel modelling

Consider a general discrete stochastic nonlinear system represented by

$$y(t) = f(y(t-1), \cdots, y(t-n_y), u(t-1), \cdots, u(t-n_u); \boldsymbol{\theta}) + \xi(t) = f(\mathbf{x}(t); \boldsymbol{\theta}) + \xi(t) \tag{1}$$

where $u(t)$ and $y(t)$ are the system input and output variables, respectively, n_u and n_y are positive integers representing the known lags in $u(t)$ and $y(t)$, respectively, the observation noise $\xi(t)$ is uncorrelated with zero mean and variance σ^2, $\mathbf{x}(t) = [y(t-1) \cdots y(t-n_y)\; u(t-1) \cdots u(t-n_u)]^T$ denotes the system input vector, $f(\bullet)$ is *a priori* unknown system mapping, and $\boldsymbol{\theta}$ is an unknown parameter vector associated with the appropriate, but yet to be determined, model structure. The system model (1) is to be identified from an N-sample system observational data set $D_N = \{\mathbf{x}(t), y(t)\}_{t=1}^N$, using some suitable functional which can approximate $f(\bullet)$ with arbitrary accuracy.

Consider the modelling of the unknown dynamical process (1) by using a linear-in-the-parameters model, e.g. the radial basis function (RBF) neural network and B-spline neurofuzzy network, formulated as [1]

$$y(t) = \sum_{k=1}^{M} p_k(\mathbf{x}(t))\theta_k + \xi(t) = \mathbf{p}^T(t)\boldsymbol{\theta} + \xi(t) \tag{2}$$

where $\mathbf{x}(t)$ is the system input vector with assumed known dimension $n = n_y + n_u$, $p_k(\bullet)$ is a known nonlinear basis function, such as RBF or B-spline fuzzy membership function, M is the number of regressors in an initial full model set, $\mathbf{p}(k) = [p_1(\mathbf{x}(t)) \cdots p_M(\mathbf{x}(t))]^T$, θ_k are the model parameters or weights and $\boldsymbol{\theta} = [\theta_1 \cdots \theta_M]^T$ the model parameter vector. The model (2) for $1 \leq t \leq N$ can be written in the matrix form as

$$\mathbf{y} = \mathbf{P}\boldsymbol{\theta} + \boldsymbol{\xi} \tag{3}$$

where $\mathbf{y} = [y(1) \cdots y(N)]^T$ is the desired output vector, $\boldsymbol{\xi} = [\xi(1) \cdots \xi(N)]^T$ is the residual vector, and $\mathbf{P} = [\mathbf{p}_1 \cdots \mathbf{p}_M]$ is the $N \times M$ regression matrix with $\mathbf{p}_j = [p_j(\mathbf{x}(1)) \cdots p_j(\mathbf{x}(N))]^T$, $1 \leq j \leq M$, being the regressor vectors. An orthogonal decomposition of $\mathbf{P}$ can be expressed as

$$\mathbf{P} = \mathbf{W}\mathbf{A} \tag{4}$$

where $\mathbf{A} = \{a_{ij}\}$ is an $M \times M$ upper triangular matrix with unity diagonal elements and $\mathbf{W}$ is an $N \times M$ matrix having orthogonal columns that satisfies

$$\mathbf{W}^T\mathbf{W} = diag\{\kappa_1, \cdots, \kappa_M\} \tag{5}$$

with

$$\kappa_k = \mathbf{w}_k^T\mathbf{w}_k, \quad 1 \leq k \leq M. \tag{6}$$

The model (3) can alternatively be expressed as

$$\mathbf{y} = (\mathbf{P}\mathbf{A}^{-1})(\mathbf{A}\boldsymbol{\theta}) + \boldsymbol{\xi} = \mathbf{W}\mathbf{g} + \boldsymbol{\xi} \tag{7}$$

in which $\mathbf{g} = [g_1 \cdots g_M]^T$ is the orthogonal weight vector. Knowing $\mathbf{g}$, the original model weight vector $\boldsymbol{\theta}$ can be calculated from $\mathbf{A}\boldsymbol{\theta} = \mathbf{g}$ through backward substitution. The space spanned by the original model bases $p_k(t) = p_k(\mathbf{x}(t))$, $1 \leq k \leq M$, is identical to the space spanned by the orthogonal bases $w_k(t)$, $1 \leq k \leq M$, and the model (2) is equivalently expressed by

$$y(t) = \mathbf{w}^T(t)\mathbf{g} + \xi(t) \tag{8}$$

where $\mathbf{w}(t) = [w_1(t) \cdots w_M(t)]^T$.

The ROLS algorithm [4, 14] is based on the following regularised cost function

$$J_R(\boldsymbol{\xi}; \lambda) = \boldsymbol{\xi}^T \boldsymbol{\xi} + \lambda \mathbf{g}^T \mathbf{g} \tag{9}$$

where λ is a regularisation parameter. The parameter estimation for $\mathbf{g}$ is readily given by

$$g_i = \frac{\mathbf{w}_i^T \mathbf{y}}{\mathbf{w}_i^T \mathbf{w}_i + \lambda} \tag{10}$$

for $1 \leq i \leq M$. In the forward regression process, the model size is configured as a growing variable k, and a subset of k regressors ($k \ll M$) is selected from the full model set consisting of the M initial regressors given by (2) to approximate the system. The forward regression procedure constructs a parsimonious model by selecting a subset of $n_\theta \ll M$ regressors based on some model selective criterion. In the ROLS forward selection procedure [4, 14], this is based on the maximisation of an regularised error reduction ratio at each forward regression step to achieve a maximal model approximation capability to the estimation data set D_N.

Clearly, the model selection criterion adopted by the ROLS algorithm is the (regularised) training mean square error (MSE). Since the training MSE typically decreases as the model size n_θ increases, additional measure is often required to determine when to terminate the selection process in order to guarantee a parsimonious model that generalises well. This may be achieved with the aid of an additional validation data set and through monitoring the performance of the selected subset model on the validation data set. Such an approach obviously increases computational expense of the model construction process. In order to optimise model approximation and adequacy simultaneously, some composite model selection criterion based on experimental design criteria, including A-optimality and D-optimality, have recently been introduced [15, 16]. In this paper, an alternative model term search criterion is used based on the PRESS statistic, which is a measure of the model generalisation capability.

3 A robust model construction algorithm using PRESS statistic and regularised orthogonal least squares

Consider the general model selection problem for modelling the system (1) by a set of K models or predictors, indexed by $k = 1, 2, \cdots, K$, that are based on a variety of model structures. Denote these models as $\hat{y}_k(t|t-1)$ if they are identified using all the N data points in D_N. To optimise the model generalisation capability, the model selection criteria are often based on cross-validation [7, 17], and one commonly used version of cross validation is called delete-1 cross validation [9, 10, 11, 12]. The idea is that, for every predictor, each data point in the training data set D_N is sequentially set aside in turn, a model is estimated using the remaining $N-1$ data points, and the prediction error is derived using only the data point that was removed from the estimation data set. Specifically, let $D_N^{(-t)}$ be the resulting data set by removing the t-th data point from D_N, and denote the k-th model estimated using $D_N^{(-t)}$ as $\hat{y}_k^{(-t)}(t|t-1)$ and the related predicted model residual at t as:

$$\xi_k^{(-t)}(t|t-1) = y(t) - \hat{y}_k^{(-t)}(t|t-1). \tag{11}$$

The leave-one-out test score or the mean square PRESS error [9, 10] for the k-th model $\hat{y}_k^{(-t)}(t|t-1)$ is obtained by averaging all these prediction errors:

$$E\left[\left(\xi_k^{(-t)}(t|t-1)\right)^2\right] = \frac{1}{N}\sum_{t=1}^{N}\left(\xi_k^{(-t)}(t|t-1)\right)^2. \tag{12}$$

To select the best model from the K candidates $\hat{y}_k(t|t-1)$, $1 \leq k \leq K$, the same modelling process is applied to all the K models, and the predictor with the minimum PRESS statistic is selected, i.e. the n_θ-th model is selected if

$$n_\theta = \arg\min_{1\leq k\leq K}\left[E\left[\left(\xi_k^{(-t)}(t|t-1)\right)^2\right]\right]. \tag{13}$$

For linear-in-the-parameters models, there is an elegant way to generate the PRESS statistic, without actually sequentially splitting the training data set and repeatedly estimating the associated models, by using the Sherman-Morrison-Woodbury theorem [9]. Consider that an M-term model $\hat{y}_M(t|t-1)$ is identified using D_N based on the model form of (2). The PRESS errors $\xi_M^{(-t)}(t|t-1)$ can be calculated using [9, 10]:

$$\xi_M^{(-t)}(t|t-1) = y(t) - \hat{y}_M^{(-t)}(t|t-1) = \frac{\xi_M(t)}{1 - \mathbf{p}^T(k)\left(\mathbf{P}^T\mathbf{P}\right)^{-1}\mathbf{p}(k)}, \tag{14}$$

where $\xi_M(t) = y(t) - \hat{y}_M(t|t-1)$. Obviously, choosing the best subset model that minimises the PRESS statistic quickly becomes computationally prohibitive even for a modest M-term model set. Moreover, the PRESS error (14) itself is computational expensive because the matrix inversion involved[1]. However, if we choose only to incrementally minimise the PRESS statistic in an orthogonal forward regression manner with an efficient computation of the PRESS error, the model selection procedure based on the PRESS statistic becomes computationally affordable.

Thus it is necessary first to derive the PRESS error in a regularised orthogonal weight model, which is given in Appendix A. From (38) in Appendix A, the PRESS error $\xi_M^{(-t)}(t|t-1)$ for the M-term orthogonal weight model (7) is given by:

$$\begin{aligned}\xi_M^{(-t)}(t|t-1) &= y(t) - \hat{y}_M^{(-t)}(t|t-1) \\ &= \frac{\xi_M(t)}{1 - \mathbf{w}(t)^T\left(\mathbf{W}^T\mathbf{W} + \mathbf{\Lambda}\right)^{-1}\mathbf{w}(t)} = \frac{\xi_M(t)}{\beta_M(t)}\end{aligned} \tag{15}$$

where $\mathbf{\Lambda} = diag\{\lambda, \cdots, \lambda\}$ is an $M \times M$ diagonal matrix and

$$\beta_M(t) = 1 - \sum_{i=1}^{M}\frac{w_i^2(t)}{\kappa_i + \lambda}. \tag{16}$$

Clearly, the amount of computation is significantly reduced by using (15), in which no matrix inversion is involved. This is due to the fact that the calculation of the PRESS error is now based on an orthogonalised model with a diagonal Hessian matrix. It can further be shown that the computational expense can be significantly reduced by utilising the forward regression process via a recursive formula. Consider the model construction using the forward regression

[1] Even adopting the recursive least square approximation with the help of matrix inversion lemma [11], the computation of (14) is still very expensive, as it still involves M-dimensional matrix multiplications

process, in which a subset model of the k regressors ($k \ll M$) is selected from the full model set consisting of the M initial regressors given by (7). The PRESS errors (15) and (16) can be written, by replacing M with a variable model size k, as

$$\xi_k^{(-t)}(t|t-1) = \frac{\xi_k(t)}{\beta_k(t)} \tag{17}$$

where

$$\beta_k(t) = 1 - \sum_{i=1}^{k} \frac{w_i^2(t)}{\kappa_i + \lambda} \tag{18}$$

and $\xi_k(t)$ is the model residual associated with the subset model structure consisting of the k selected regressors. $\beta_k(t)$ can be written as a recursive formula, given by

$$\beta_k(t) = \beta_{k-1}(t) - \frac{w_k^2(t)}{\kappa_k + \lambda}. \tag{19}$$

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

As is in the conventional forward regression [2], a Gram-Schmidt procedure is used to construct the orthogonal basis $\mathbf{w}_i$ in a forward regression manner. At each regression step k, the PRESS statistic can be computed with:

$$\begin{aligned} J_k &= E\left[\left(\xi_k^{(-t)}(t|t-1)\right)^2\right] \\ &= E\left[\frac{\xi_k^2(t)}{\beta_k^2(t)}\right] = \frac{1}{N}\sum_{t=1}^{N} \frac{\xi_k^2(t)}{\beta_k^2(t)} \end{aligned} \tag{20}$$

and this is then used as the regressor selective criterion for the model construction which minimises this mean square PRESS error. Due to the properties associated with the minimisation of the PRESS statistic, a fully automatic model construction process can be achieved. This is because the function J_k is concave versus k, and there exists an "optimal" model size n_θ such that for $k < n_\theta$ J_k decreases as k increases, while for $k > n_\theta$ J_k increases as k increases. This point can be formally analyzed as follows. The model residual $\xi_k(t)$ for the model with size k is

$$\xi_k(t) = y(t) - \sum_{i=1}^{k} w_i(t) g_i \tag{21}$$

and clearly

$$\xi_k(t) = \xi_{k+1}(t) + w_{k+1}(t) g_{k+1}. \tag{22}$$

From (20) and(22), the PRESS statistic for the model of size k is given by

$$\begin{aligned} J_k &= E\left[\frac{\xi_k^2(t)}{\beta_k^2(t)}\right] = E\left[\frac{\left(\xi_{k+1}(t) + w_{k+1}(t) g_{k+1}\right)^2}{\beta_k^2(t)}\right] \\ &= E\left[\frac{\xi_{k+1}^2(t)}{\beta_k^2(t)}\right] + E\left[\frac{w_{k+1}^2(t) g_{k+1}^2}{\beta_k^2(t)}\right] \end{aligned} \tag{23}$$

by assuming that the model residual sequence is uncorrelated with the model regressors. The change in the PRESS statistic by increasing k to $(k+1)$ can be written as

$$\Delta J = J_{k+1} - J_k = E\left[\frac{\xi_{k+1}^2(t)}{\beta_{k+1}^2(t)}\right] - E\left[\frac{\xi_{k+1}^2(t)}{\beta_k^2(t)}\right] - E\left[\frac{w_{k+1}^2(t) g_{k+1}^2}{\beta_k^2(t)}\right]. \tag{24}$$

The difference between the first two terms in (24), $E\left[\xi_{k+1}^2(t)/\beta_{k+1}^2(t)\right] - E\left[\xi_{k+1}^2(t)/\beta_k^2(t)\right]$, represents the effects of the PRESS error inflation from a model with k regressors to that of $(k+1)$ regressors. Clearly

$$E\left[\frac{\xi_{k+1}^2(t)}{\beta_{k+1}^2(t)}\right] - E\left[\frac{\xi_{k+1}^2(t)}{\beta_k^2(t)}\right] > 0 \quad \text{for} \quad E\left[w_{k+1}^2(t)\right] = \frac{\mathbf{w}_{k+1}^T \mathbf{w}_{k+1}}{N} > 0 \tag{25}$$

due to $\beta_{k+1}^2(t) < \beta_k^2(t)$. The effect of this PRESS error inflation tends to increase J_k. On the other hand, the last term in (24), $E\left[w_{k+1}^2(t) g_{k+1}^2/\beta_k^2(t)\right]$, which represents the contribution of the $(k+1)$th regressor in modelling accuracy, tends to decrease J_k. However, since the training accuracy typically improves as k increases but at a gradually reduced rate, the last term becomes less significant and eventually it becomes less than the effects of the PRESS error inflation. That is, as the model achieves a sufficient approximation capability at a certain model size $k = n_\theta$, the last term in (24) becomes insignificant in comparison with the PRESS error inflation, resulting in $\Delta J > 0$ or $E\left[w_{k+1}^2(t) g_{k+1}^2/\beta_k^2(t)\right] < E\left[\xi_{k+1}^2(t)/\beta_{k+1}^2(t)\right] - E\left[\xi_{k+1}^2(t)/\beta_k^2(t)\right]$.

This property, i.e. ΔJ changes the sign at certain model size k, can be applied to construct the automatic algorithm. The proposed ROLS algorithm based on the PRESS statistic selects significant regressors that minimises the PRESS statistic, with a growing model structure until $\Delta J \geq 0$ at a desired model size n_θ, where the contribution of the $(n_\theta + 1)$th regressor in model approximation becomes insignificant. Thus the proposed algorithm terminates at $J_{n_{\theta+1}} \geq J_{n_\theta}$, where the model is optimised based on the minimisation of the PRESS statistics at J_{n_θ}. Note that neither a separate criterion to terminate the selection procedure nor any iteration of the procedure is needed (as the procedure does not use any controlling parameter to be adjusted via iterations)[2]. The proposed algorithm is based on the standard Gram-Schmidt procedure in which the orthogonal basis $\mathbf{w}_i$ is constructed in a forward regression manner. In this algorithm a small fixed positive regularisation parameter, e.g. $\lambda = 10^{-4}$, is used to improve parameter estimation variance. Note that the algorithm selects only those model terms which satisfy $E[w_{k+1}^2(t)] \neq 0$. Thus any numerical ill-conditioning problem is automatically avoided. The model selection procedure of this ROLS algorithm based on the PRESS statistic is summarized in Appendix B.

4 Numerical examples

Two examples were used to demonstrate the effectiveness of the proposed ROLS algorithm using the PRESS statistic and to compare it with the existing ROLS algorithm using the training MSE.

Example 1. Consider using a RBF network to approximate an unknown scalar function

$$f(x) = \frac{\sin(x)}{x}, \quad -10 \leq x \leq 10. \tag{26}$$

Four hundred training data were generated from $y = f(x) + \xi$, where the input x was uniformly distributed in $[-10,\ 10]$ and the noise ξ was Gaussian with zero mean and standard deviation 0.2. The first two hundred samples

[2]Regularization in the present algorithm is based on ridge regression with a single fixed small regularization parameter. In an alternative construction algorithm given in [18], a more powerful multiple-regularizer approach is adopted. However, it is then necessary to adapt the regularization parameters or hyperparameters within an iterative loop, which increases computational complexity.

were used for training and the last two hundred data points for possible model validation. The Gaussian function

$$p_i(\mathbf{x}) = \exp\left(-\frac{\|\mathbf{x} - \mathbf{c}_i\|^2}{2\tau^2}\right) \tag{27}$$

was used as the basis function to construct a RBF model, with a kernel width $\tau^2 = 10.0$. Appropriate value for the kernel width was found empirically in this study. In general, it can also be determined through cross-validation. All the two hundred training data points were used as the candidate RBF centre set for $\mathbf{c}_i$. The training data were very noisy. Two hundred noise-free data $f(x)$ with equally spaced x in $[-10,\ 10]$ were also generated as an additional testing data set for evaluating model performance. The regularisation parameter was fixed to $\lambda = 0.001$.

Fig. 1 depicts the evolution of the training MSE and PRESS statistic in log scale during the forward regression procedure with a typical set of noisy training data set using the proposed ROLS algorithm based on the PRESS statistic. It can be seen from Fig. 1 that the PRESS statistic continuously decreased until $J_8 = 0.041589 \geq J_7 = 0.041589$, and the algorithm terminated with a 7-term model. Fig. 2 shows the noisy training points y and the underlying function $f(x)$ together with the mapping generated using this 7-term model identified by the ROLS algorithm based on the PRESS statistic. The ROLS algorithm based on the training MSE [2, 14] was also used to fit the same training data set. Since the training MSE may continuously decrease as the model size k increases, the validation data set was employed to aid the determination of the model structure during the forward regression procedure. Fig. 3 depicts the training and testing MSE values over the training and validation data sets, respectively, versus the model size k, using the ROLS algorithm based on the training MSE. The test MSE over the validation set reached the minimum value of 0.041736 at $k = 9$, and this indicated a 9-term model. The corresponding model mapping generated by this 9-term model is illustrated in Fig. 4. Table 1 summarizes the modelling accuracies (mean $\pm$ standard deviation) of the two algorithms averaged over ten sets of different data realizations. It can be seen that the two algorithms had similarly good generalization performance, but the ROLS algorithm based on the PRESS statistic was able to produce sparser models and it had a further advantage that no additional validation set was needed for model evaluation during the model construction process.

Example 2. This example constructed a model representing the relationship between the fuel rack position (input $u(t)$) and the engine speed (output $y(t)$) for a Leyland TL11 turbocharged, direct injection diesel engine operated at low engine speed. Detailed system description and experimental setup can be found in [19]. The data set, depicted in Fig. 5, contained 410 samples. The first 210 data points were used in training and the last 200 points in possible model validation. A RBF model with the input vector

$$\mathbf{x}(t) = [y(t-1)\ u(t-1)\ u(t-2)]^T \tag{28}$$

and the Gaussian basis function of variance $\tau^2 = 1.69$ was used to model the data. All the 210 training data points were used as the candidate RBF centre set and the regularisation parameter was fixed to $\lambda = 10^{-7}$.

Fig. 6 shows the evolution of the training MSE and PRESS statistic during the forward regression procedure using the ROLS algorithm based on the PRESS statistic, where it can be seen that the PRESS statistic continuously decreased until $J_{24} = 0.000548 \geq J_{23} = 0.000548$. The algorithm thus automatically terminated with a 23-term model. Fig. 7 depicts the training and testing MSE values over the training and validation data sets, respectively, versus the model size k, using the ROLS algorithm based on the training MSE, where it can be seen that the training MSE continuously decreased as the model size increased. The test MSE over the validation set reached the minimum value of 0.000517 at $k = 25$, indicating a 25-term model. The two models constructed by the two algorithms are compared in Table 2. Again it can be seen that the two models had similarly excellent generalisation capabilities, but the model constructed by the ROLS algorithm based on the PRESS statistic was sparser and this algorithm did not need the validation set for model evaluation during the model selection procedure. The constructed RBF model $\hat{f}_{RBF}(\bullet)$ was used to generate the model prediction according to

$$\hat{y}(t) = \hat{f}_{\mathrm{RBF}}(\mathbf{x}(t)) \tag{29}$$

with the input vector $\mathbf{x}(t)$ given by (28). Fig. 8 depicts the model prediction $\hat{y}(t)$ and the prediction error $\xi(t) = y(t) - \hat{y}(t)$ for the 23-term model constructed by the ROLS algorithm based on the PRESS statistic. The other model have similar prediction performance to the results shown in Fig. 8.

5 Conclusions

This paper has introduced an automatic model construction algorithm for linear-in-the-parameters nonlinear models by combining parameter regularisation via ridge regression and model structure selection based directly on maximising model generalisation capability. It has been demonstrated that parameter regularisation and model optimal generalisation structure selection are two effective and complementary approaches for robust sparse modelling. The leave-one-out test score or PRESS statistic in the framework of regularised orthogonal least squares has been derived and, in particular, an efficient recursive computation formula for PRESS errors has been developed. The proposed algorithm based on forward regression combines parameter regularisation technique in orthogonal weight space and the PRESS statistic to optimise model structure in order to achieve improved generalisation capability. The proposed algorithm is applicable to a wide range of signal processing and model based controller design applications.

Appendix A: PRESS error in a regularised orthogonal weight model

Following (10), the parameter vector in the M-term regularised orthogonal weight model is

$$\mathbf{g} = \left(\mathbf{W}^T\mathbf{W} + \mathbf{\Lambda}\right)^{-1}\mathbf{W}^T\mathbf{y} = \mathbf{H}^{-1}\mathbf{W}^T\mathbf{y} \tag{30}$$

where $\mathbf{\Lambda} = \lambda\mathbf{I}_M$ with $\mathbf{I}_M$ being the $M \times M$ identity matrix. The model residual based on the M-term regularised orthogonal weight model is

$$\xi_M(t) = y(t) - \mathbf{g}^T\mathbf{w}(t) = y(t) - \mathbf{y}^T\mathbf{W}\mathbf{H}^{-1}\mathbf{w}(t). \tag{31}$$

If the data sample indexed at t is deleted from the estimation data set D_N, the delete-1 model parameter vector for the regularised orthogonal weight model is given by

$$\mathbf{g}^{(-t)} = \left([\mathbf{W}^{(-t)}]^T \mathbf{W}^{(-t)} + \mathbf{\Lambda} \right)^{-1} [\mathbf{W}^{(-t)}]^T \mathbf{y}^{(-t)} = [\mathbf{H}^{(-t)}]^{-1} [\mathbf{W}^{(-t)}]^T \mathbf{y}^{(-t)} \tag{32}$$

where $\mathbf{W}^{(-t)}$ and $\mathbf{y}^{(-t)}$ denote the resultant regression matrix and desired output vector, respectively, formed from $D_N^{(-t)}$. By definition, it can be shown that

$$\mathbf{H}^{(-t)} = \mathbf{H} - \mathbf{w}(t)\mathbf{w}^T(t) \tag{33}$$

and

$$[\mathbf{y}^{(-t)}]^T \mathbf{W}^{(-t)} = \mathbf{y}^T \mathbf{W} - y(t)\mathbf{w}^T(t). \tag{34}$$

The PRESS error evaluated at t in the associated regularised orthogonal weight model is given by

$$\xi_M^{(-t)}(t|t-1) = y(t) - [\mathbf{g}^{(-t)}]^T \mathbf{w}(t) = y(t) - [\mathbf{y}^{(-t)}]^T \mathbf{W}^{(-t)} [\mathbf{H}^{(-t)}]^{-1} \mathbf{w}(t). \tag{35}$$

From (33), using the matrix inversion lemma yields

$$\left(\mathbf{H}^{(-t)} \right)^{-1} = \left(\mathbf{H} - \mathbf{w}(t)\mathbf{w}^T(t) \right)^{-1} = \mathbf{H}^{-1} + \frac{\mathbf{H}^{-1}\mathbf{w}(t)\mathbf{w}^T(t)\mathbf{H}^{-1}}{1 - \mathbf{w}^T(t)\mathbf{H}^{-1}\mathbf{w}(t)} \tag{36}$$

and

$$\left(\mathbf{H}^{(-t)} \right)^{-1} \mathbf{w}(t) = \frac{\mathbf{H}^{-1}\mathbf{w}(t)}{1 - \mathbf{w}^T(t)\mathbf{H}^{-1}\mathbf{w}(t)}. \tag{37}$$

Substituting (34) and (37) into (35) yields

$$\begin{aligned} \xi_M^{(-t)}(t|t-1) &= y(t) - \left(\mathbf{y}^T \mathbf{W} - y(t)\mathbf{w}^T(t) \right) \times \frac{\mathbf{H}^{-1}\mathbf{w}(t)}{1 - \mathbf{w}^T(t)\mathbf{H}^{-1}\mathbf{w}(t)} \\ &= \frac{y(t) - \mathbf{y}^T \mathbf{W}\mathbf{H}^{-1}\mathbf{w}(t)}{1 - \mathbf{w}^T(t)\mathbf{H}^{-1}\mathbf{w}(t)} = \frac{\xi_M(t)}{1 - \mathbf{w}^T(t)\mathbf{H}^{-1}\mathbf{w}(t)} \\ &= \frac{\xi_M(t)}{1 - \mathbf{w}^T(t)\left(\mathbf{W}^T \mathbf{W} + \mathbf{\Lambda} \right)^{-1} \mathbf{w}(t)}. \end{aligned} \tag{38}$$

Appendix B: Combined PRESS statistic and regularised orthogonal least squares for subset model selection

1. At the first step, initialise $J_0 = \mathbf{y}^T \mathbf{y}$, $\xi_0(t) = y(t)$ and $\beta_0(t) = 1$ for $t = 1, \cdots, N$. For $1 \le i \le M$, compute

$$\begin{aligned} \mathbf{w}_1^{(i)} &= \mathbf{p}_i, \\ \kappa_1^{(i)} &= \left(\mathbf{w}_1^{(i)} \right)^T \mathbf{w}_1^{(i)}, \\ g_1^{(i)} &= \frac{\left(\mathbf{w}_1^{(i)} \right)^T \mathbf{y}}{\left(\mathbf{w}_1^{(i)} \right)^T \mathbf{w}_1^{(i)} + \lambda}, \\ \xi_1^{(i)}(t) &= \xi_0(t) - w_1^{(i)}(t) g_1^{(i)} \quad \text{for} \quad t = 1, \cdots, N, \end{aligned}$$

$$\beta_1^{(i)}(t) = \beta_0(t) - \frac{\left(w_1^{(i)}(t)\right)^2}{\kappa_1^{(i)} + \lambda} \quad \text{for} \quad t = 1, \cdots, N,$$

$$J_1^{(i)} = \frac{1}{N} \sum_{t=1}^{N} \frac{\left(\xi_1^{(i)}(t)\right)^2}{\left(\beta_1^{(i)}(t)\right)^2}.$$

Find

$$i_1 = \arg\min\{J_1^{(i)},\ 1 \le i \le M\}$$

and select

$$\mathbf{w}_1 = \mathbf{w}_1^{(i_1)} = \mathbf{p}_{i_1}$$

with $J_1 = J_1^{(i_1)}$ and

$$\xi_1(t) = \xi_0(t) - w_1(t) g_1 \quad \text{for} \quad t = 1, \cdots, N,$$

$$\beta_1(t) = \beta_0(t) - \frac{w_1^2(t)}{\kappa_1 + \lambda} \quad \text{for} \quad t = 1, \cdots, N.$$

2. At the kth step where $k \ge 2$, for $1 \le i \le M$ and $i \ne i_1, \cdots, i \ne i_{k-1}$, compute

$$a_{jk}^{(i)} = \frac{\mathbf{w}_j^T \mathbf{p}_i}{\mathbf{w}_j^T \mathbf{w}_j}, \ \ 1 \le j < k,$$

$$\mathbf{w}_k^{(i)} = \mathbf{p}_i - \sum_{j=1}^{k-1} a_{jk}^{(i)} \mathbf{w}_j,$$

$$\kappa_k^{(i)} = \left(\mathbf{w}_k^{(i)}\right)^T \mathbf{w}_k^{(i)},$$

$$g_k^{(i)} = \frac{\left(\mathbf{w}_k^{(i)}\right)^T \mathbf{y}}{\left(\mathbf{w}_k^{(i)}\right)^T \mathbf{w}_k^{(i)} + \lambda},$$

$$\xi_k^{(i)}(t) = \xi_{k-1}(t) - w_k^{(i)}(t) g_k^{(i)} \quad \text{for} \quad t = 1, \cdots, N,$$

$$\beta_k^{(i)}(t) = \beta_{k-1}(t) - \frac{\left(w_k^{(i)}(t)\right)^2}{\kappa_k^{(i)} + \lambda} \quad \text{for} \quad t = 1, \cdots, N,$$

$$J_k^{(i)} = \frac{1}{N} \sum_{t=1}^{N} \frac{\left(\xi_k^{(i)}(t)\right)^2}{\left(\beta_k^{(i)}(t)\right)^2}.$$

Find

$$i_k = \arg\min\{J_k^{(i)},\ 1 \le i \le M \ \text{ and } \ i \ne i_1, \cdots, i \ne i_{k-1}\}$$

and select

$$a_{jk} = a_{jk}^{(i_k)},$$

$$\mathbf{w}_k = \mathbf{w}_k^{(i_k)} = \mathbf{p}_{i_k} - \sum_{j=1}^{k-1} a_{jk} \mathbf{w}_j$$

with $J_k = J_k^{(i_k)}$ and

$$\xi_k(t) = \xi_{k-1}(t) - w_k(t) g_k \quad \text{for} \quad t = 1, \cdots, N,$$

$$\beta_k(t) = \beta_{k-1}(t) - \frac{w_k^2(t)}{\kappa_k + \lambda} \quad \text{for} \quad t = 1, \cdots, N.$$

3. The selection procedure is terminated with an n_θ-term model at the $k = n_\theta$ step, when $J_k \geq J_{k-1}$. Otherwise, set $k = k + 1$, and go to step 2.

References

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

[2] Chen, S., Billings, S.A., and Luo, W., 1989, "Orthogonal least squares methods and their applications to non-linear system identification," *Int. J. Control*, Vol.50, No.5, pp.1873–1896.

[3] Hoerl, A.E., and Kennard, R.W., 1970, "Ridge regression: biased estimation for non-orthogonal problems," *Technometrics*, Vol.12, pp.55–67.

[4] Chen, S., Wu, Y., and Luk, B.L., 1999, "Combined genetic algorithm optimization and regularized orthogonal least squares learning for radial basis function networks," *IEEE Trans. Neural Networks*, Vol.10, No.5, pp.1239–1243.

[5] Orr, M.J.L., 1993, "Regularisation in the selection of radial basis function centers," *Neural Computation*, Vol.7, No.3, pp.954–975.

[6] Akaike, H., 1974, "A new look at the statistical model identification," *IEEE Trans. Automatic Control*, Vol.AC-19, pp.716–723.

[7] Stone, M., 1974, "Cross validatory choice and assessment of statistical predictions," *J. R. Statist. Soc. Ser. B.*, Vol.36, pp.117–147.

[8] Breiman, L., 1996, "Stacked regression," *Machine Learning*, Vol.5, pp.49–64.

[9] Myers, R.H., 1990, *Classical and Modern Regression with Applications*, 2nd Edition, Boston: PWS-KENT.

[10] Hansen, L.K., and Larsen, J., 1996, "Linear unlearning for cross-validation," *Advances in Computational Mathematics*, Vol.5, pp.269–280.

[11] Bontempi, G., Birattari, M. and Bersini, H., 1999, "Lazy learning for local modelling and control design," *Int. J. Control*, Vol.72, No.7/8, pp.643–658.

[12] Cawley, G.C., and Talbot, N.L.C., 2003, "Efficient leave-one-out cross-validation of kernel Fisher discriminant classifiers," *Pattern Recognition*, Vol.36, No.11, pp.2585–2592.

[13] Hong, X., Sharkey, P.M., and Warwick, K., 2003, "Automatic nonlinear predictive model construction algorithm using forward regression and the PRESS statistic," *IEE Proc. Control Theory and Applications*, Vol.150, No.3, pp.245–254.

[14] Chen, S., Chng, E.S., and Alkadhimi, K., 1996, "Regularised orthogonal least squares algorithm for constructing radial basis function networks," *Int. J. Control*, Vol.64, No.5, pp.829–837.

[15] Hong, X., and Harris, C.J., 2001, "Neurofuzzy design and model construction of nonlinear dynamical processes from data," *IEE Proc. Control Theory and Applications*, Vol.148, No.6, pp.530–538.

[16] Hong, X., and Harris, C.J., 2002, "Nonlinear model structure design and construction using orthogonal least squares and D-optimality design," *IEEE Trans. Neural Networks*, Vol.13, No.5, pp.1245–1250.

[17] Ljung, L., 1987, *System Identification: Theory for the User*, Prentice Hall.

[18] Chen, S., Hong, X., Harris, C.J., and Sharkey, P.M., 2004, "Sparse modelling using orthogonal forward regression with PRESS statistic and regularization," *IEEE Trans. Systems, Man and Cybernetics, Part B*, to appear, 2004.

[19] Billings, S.A., Chen, S., and Backhouse, R.J., 1989, "The identification of linear and non-linear models of a turbocharged automotive diesel engine," *Mechanical Systems and Signal Processing*, Vol.3, No.2, pp.123–142.

Table 1: Modelling accuracy (mean $\pm$ standard deviation) over ten sets of different data realizations for the simple scalar function modelling.

algorithm	ROLS with training MSE	ROLS with PRESS
validation set used	Yes	No
model terms	8.7 ± 1.6	7.8 ± 0.6
MSE over training set	0.037938 ± 0.003568	0.037703 ± 0.003708
PRESS statistic	0.040852 ± 0.003765	0.040725 ± 0.003893
MSE over noisy test set	0.041533 ± 0.002519	0.041692 ± 0.002458
MSE over noise-free test set	0.001701 ± 0.000660	0.001749 ± 0.000630

Table 2: Modelling accuracy for the engine data set modelling.

algorithm	ROLS with training MSE	ROLS with PRESS
validation set used	Yes	No
model terms	25	23
MSE over training set	0.000450	0.000449
PRESS statistic	0.000571	0.000548
MSE over test set	0.000517	0.000487

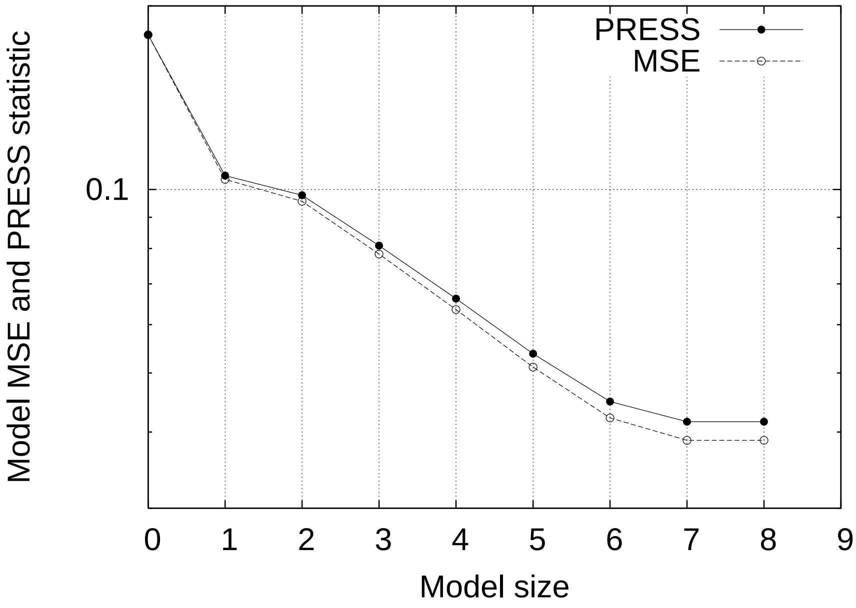


Figure 1: The evolution of training MSE and PRESS statistic versus model size for simple scalar function modelling problem using the ROLS algorithm based on PRESS statistic without the help of a validation set.

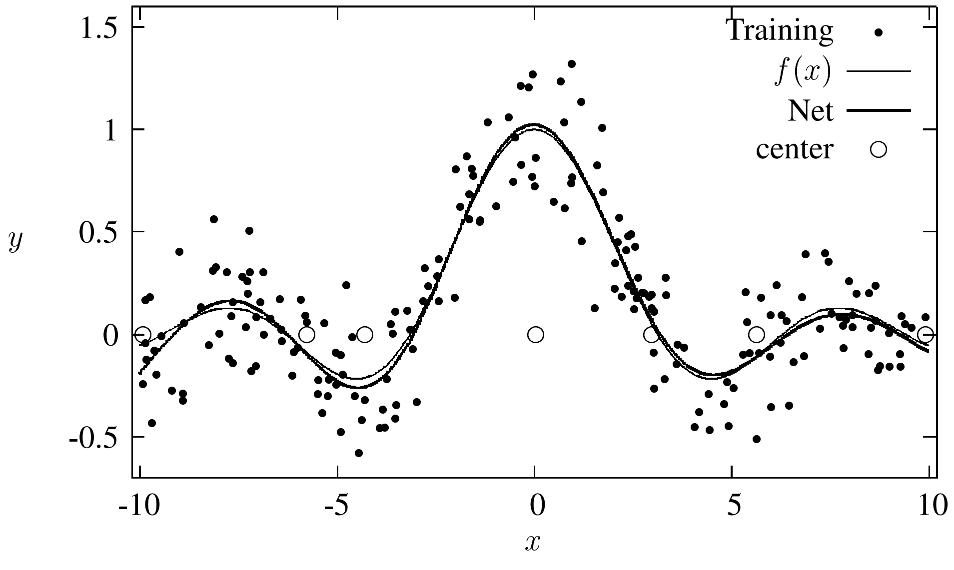


Figure 2: Simple scalar function modelling problem: a typical set of noisy training data y (dots), underlying function $f(x)$ (thin curve), model mapping (thick curve), and selected RBF centres (circles). The 7-term model was identified by the ROLS algorithm based on PRESS statistic without the help of a validation set.

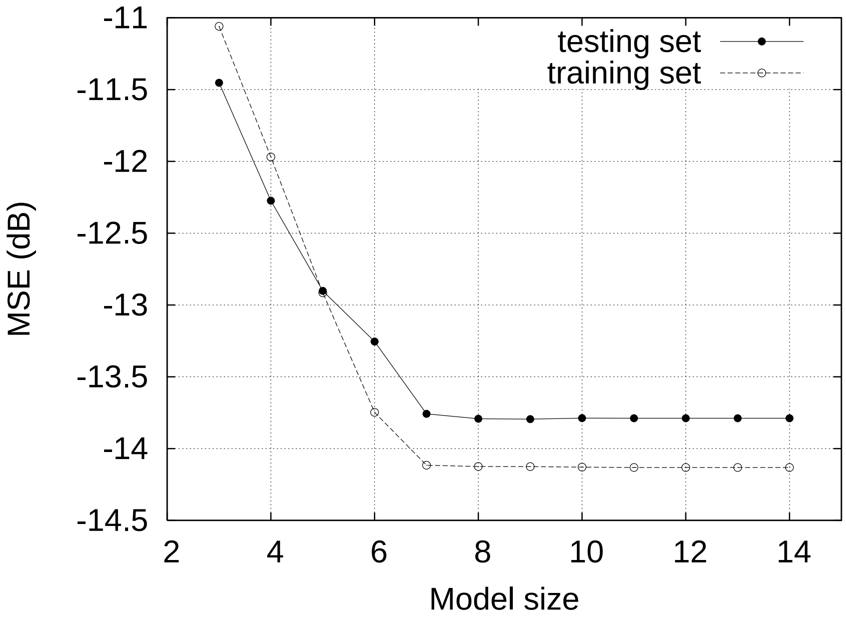


Figure 3: Training and testing MSE values over the training and validation sets, respectively, versus model size for simple scalar function modelling problem using the ROLS algorithm based on training MSE with the aid of a validation set.

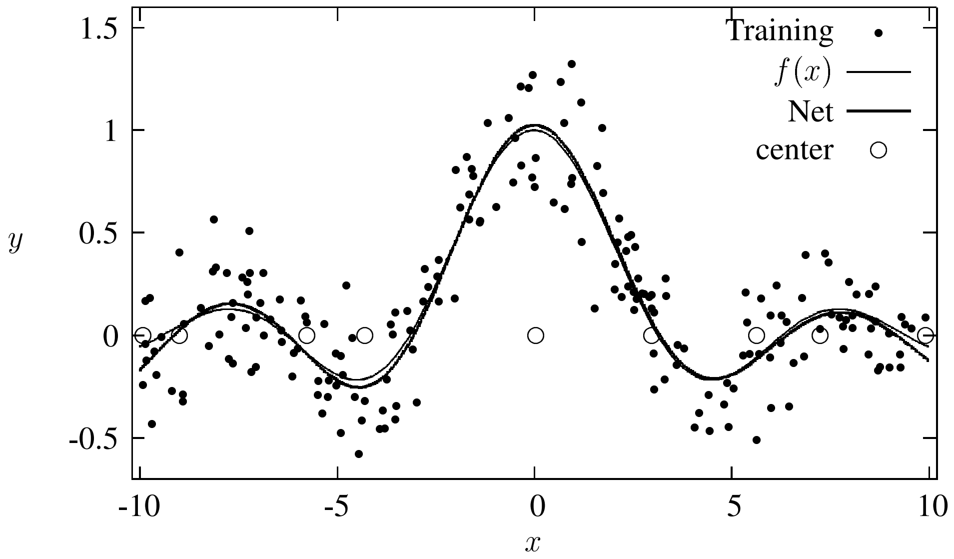
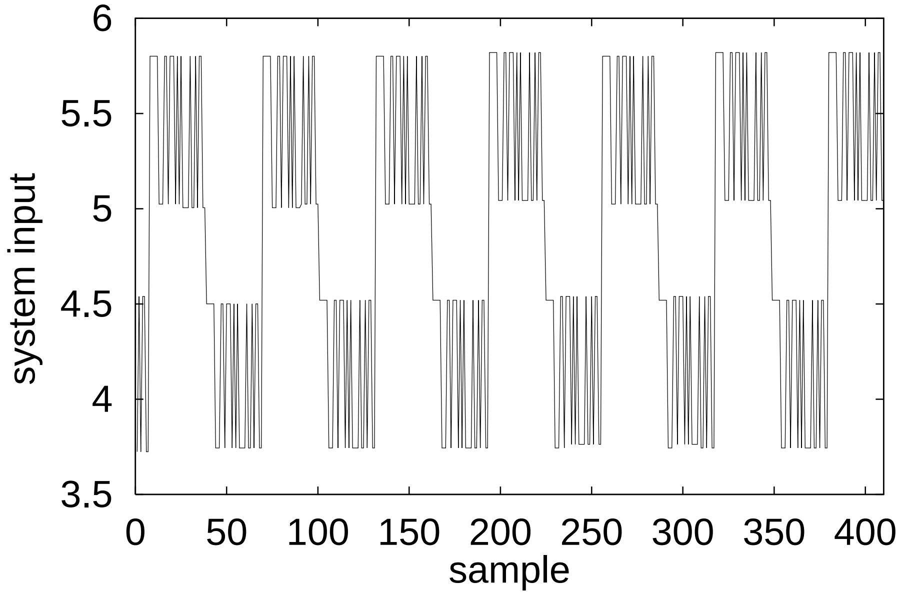
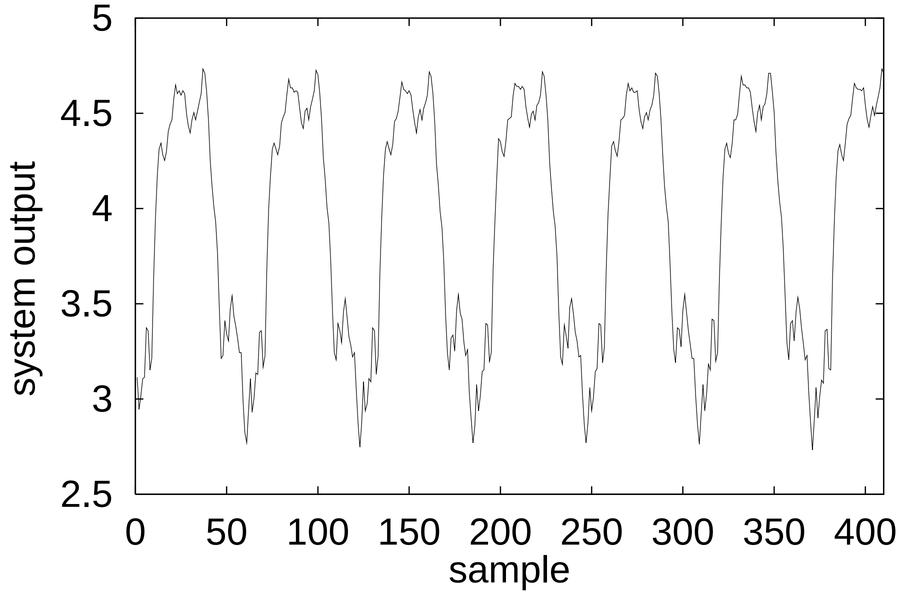


Figure 4: Simple scalar function modelling problem: a typical set of noisy training data y (dots), underlying function $f(x)$ (thin curve), model mapping (thick curve), and selected RBF centres (circles). The 9-term model was identified by the ROLS algorithm based on training MSE with the aid of a validation set.



(a)



(b)

Figure 5: Engine data set (a) input $u(t)$ and (b) output $y(t)$.

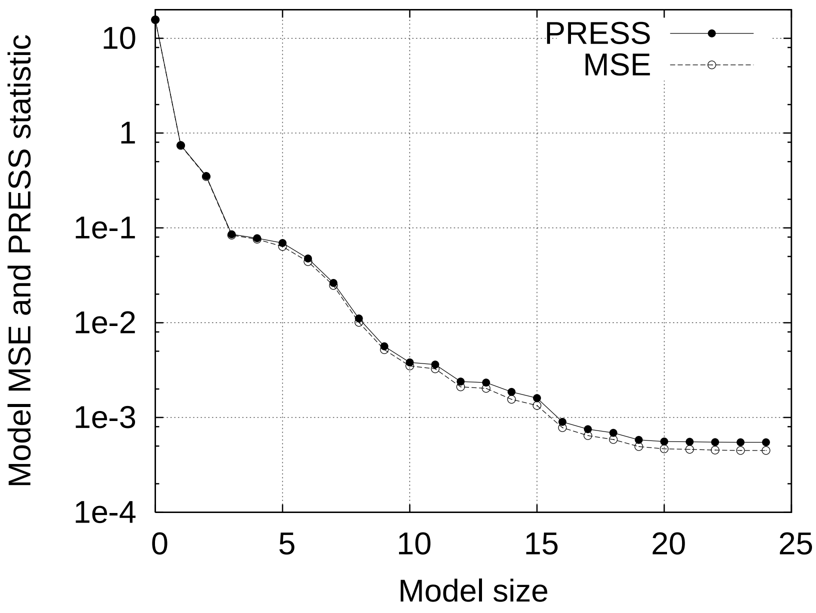


Figure 6: The evolution of training MSE and PRESS statistic versus model size for engine data set modelling problem using the ROLS algorithm based on PRESS statistic without the help of a validation set.

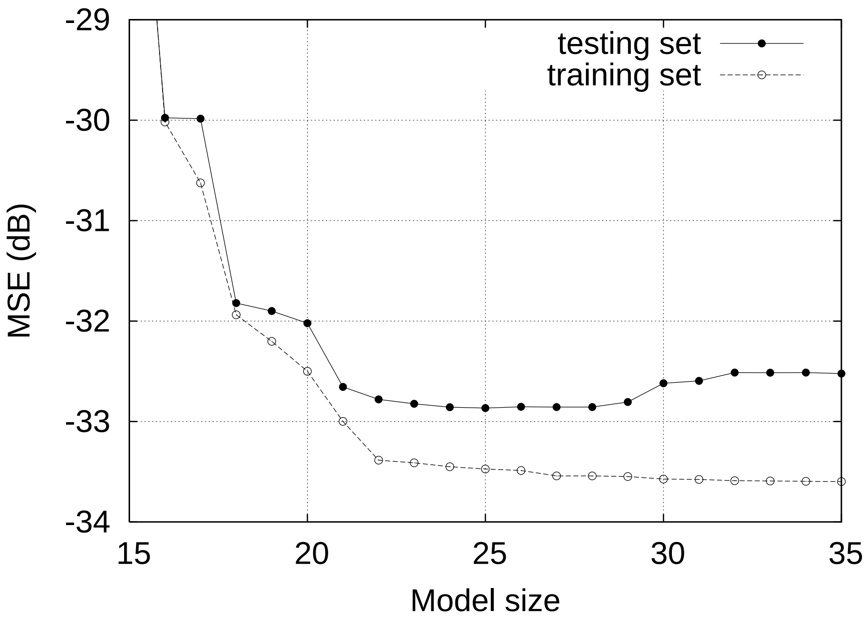
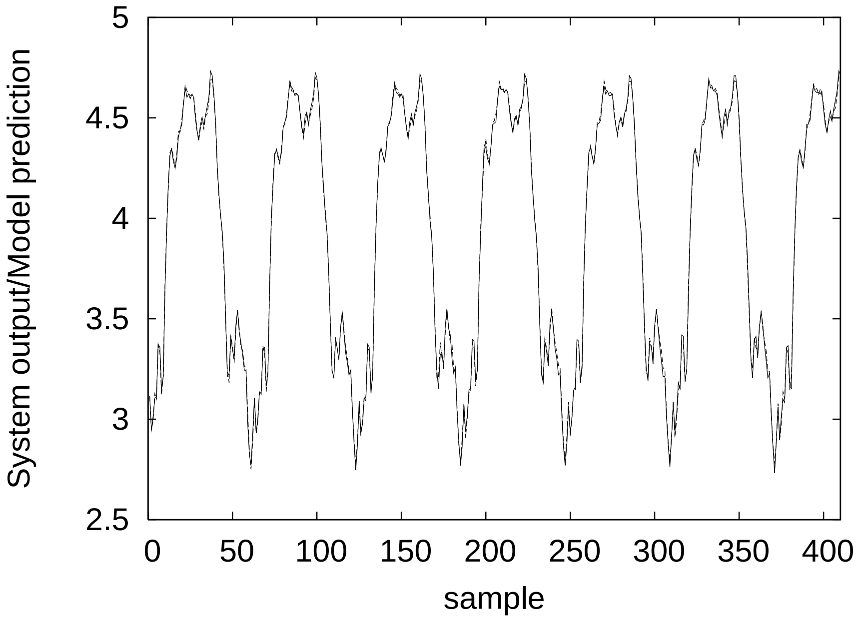
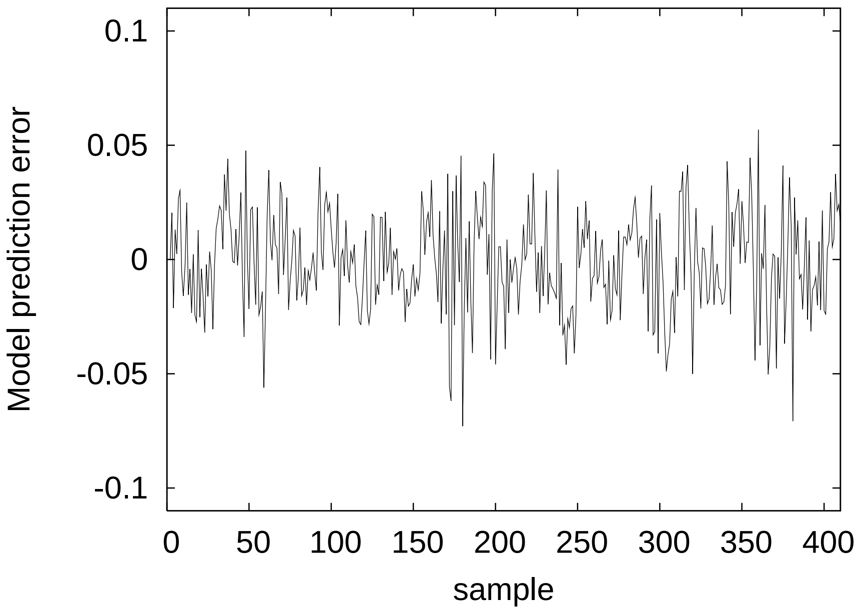


Figure 7: Training and testing MSE values over the training and validation sets, respectively, versus model size for engine data set modelling problem using the ROLS algorithm based on training MSE with the aid of a validation set.



(a) Model prediction $\hat{y}(t)$ (dashed) superimposed on system output $y(t)$ (solid)



(b) Model prediction error $\xi(t)$

Figure 8: Modelling performance for engine data set modelling problem. The 23-term model was constructed by the ROLS algorithm based on PRESS statistic without the help of a validation set.