Orthogonal Forward Regression based on Directly Maximizing Model Generalization Capability

S. Chen† and X. Hong‡
† Department of Electronics and Computer Science
University of Southampton, Southampton SO17 1BJ, UK
‡ Department of Cybernetics
University of Reading, Reading, RG6 6AY, UK

Abstract

The paper introduces a construction algorithm for sparse kernel modelling using the leave-one-out test score also known as the PRESS (Predicted REsidual Sums of Squares) statistic. An efficient subset model selection procedure is developed in the orthogonal forward regression framework by incrementally maximizing the model generalization capability to construct sparse models with good generalization properties. The proposed algorithm achieves a fully automated model construction without resort to any other validation data set for costly model evaluation.

Index Terms — orthogonal forward regression, structure identification, cross validation, generalization.

1 Introduction

The least squares (LS) principle has been fundamental to data modelling and the training mean square error (MSE) has always played a central role in model structure construction and parameter estimation. It is well known that the model based on the pure LS estimate tend to be unsatisfactory for an ill conditioned design matrix, and may over-fit the noise in training data to produce an oversized ill-posed model with high parameter estimate variances. To produce a model with good generalization capabilities, model selection criteria such as the Akaike information criterion (AIC) [1], local regularization and optimal experimental design [2]–[4] incorporate some sorts of model structure regularization with the basic training MSE criterion. In forward regression setting [5], which is a practical way of constructing a kernel model from a large data set, local regularization and optimal experimental design criteria are known to offer better solutions [2]–[4], compared with the AIC.

In order to achieve a model structure with improved model generalization, it is natural that a model generalization 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 generalization capability is directly based on the concept of cross validation [6], it is highly desirable to develop model selective criteria based on the concept of cross validation that can distinguish model generalization 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 [7]–[9]. The leave-one-out test score is a measure of model generalization capability. Traditional model structure determination based on the leave-one-out test score or PRESS statistic is however inherently inefficient and computationally prohibitive.

The paper introduces an efficient automatic model construction algorithm that directly optimizes model generalization capability. The computational efficiency is achieved through incrementally minimizing the leave-one-out test score in an orthogonal forward regression framework, which minimizes the effort in the computation of the PRESS statistic. Further significant reduction in computation arises owing to a forward recursive formula to compute PRESS errors. In the proposed algorithm, the PRESS statistic, which is a measure of model generalization capability, is applied directly in the orthogonal forward regression model structure construction process as a cost function in order to optimize the model generalization capability. The proposed algorithm achieves 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 Kernel modelling

Consider a general discrete stochastic nonlinear system represented by [10]:

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

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 model structure. The system model (1) is to be identified from an N-sample system observational data set $D_N = \{\mathbf{x}(t), y(t)\}_{t=1}^N$.

Consider the modelling of the unknown dynamical process (1) by using a linear-in-the-parameters model of the form:

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

where M is the number of candidate regressors, $\mathbf{p}(t) = [p_1(\mathbf{x}(t))\cdots p_M(\mathbf{x}(t))]^T$, θ_k are the model 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

$$\mathbf{y} = \mathbf{P}\boldsymbol{\theta} + \boldsymbol{\xi} \qquad (3)$$

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 $\mathbf{p}_j = [p_j(\mathbf{x}(1))\cdots p_j(\mathbf{x}(N))]^T$, $1 \leq j \leq M$. An orthogonal decomposition of $\mathbf{P}$ can be expressed as

$$\mathbf{P} = \mathbf{W}\mathbf{A} \qquad (4)$$

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

$$\mathbf{W}^T\mathbf{W} = diag\{\kappa_1, \cdots, \kappa_M\} \qquad (5)$$

with $\kappa_k = \mathbf{w}_k^T\mathbf{w}_k$, $1 \leq k \leq M$. The model (3) can alternatively be expressed as

$$\mathbf{y} = (\mathbf{P}\mathbf{A}^{-1})(\mathbf{A}\boldsymbol{\theta}) + \boldsymbol{\xi} = \mathbf{W}\mathbf{g} + \boldsymbol{\xi} \qquad (6)$$

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 $\mathbf{A}\boldsymbol{\theta} = \mathbf{g}$. The space spanned by the original model bases $p_k(t) = p_k(\mathbf{x}(t))$, $1 \leq k \leq M$, is identical to that spanned by the orthogonal bases $w_k(t)$, $1 \leq k \leq M$, and the model (2) is equivalently expressed by

$$y(t) = \mathbf{w}^T(t)\mathbf{g} + \xi(t) \qquad (7)$$

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

3 Orthogonal forward regression using PRESS statistic

Consider the model selection problem for modelling (1) by a set of K models, 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 optimize the model generalization capability, the model selection criteria are often based on cross-validation [6], and one commonly used version of cross validation is called delete-1 cross validation [8],[9]. The idea is that, for every model, 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:

$$\xi_k^{(-t)}(t|t-1) = y(t) - \hat{y}_k^{(-t)}(t|t-1). \qquad (8)$$

The leave-one-out test score or the mean square PRESS error [8],[9] for the k-th model $\hat{y}_k^{(-t)}(t|t-1)$ is obtained by averaging all these prediction errors:

$$E\left[\left(\xi_k^{(-t)}(t|t-1)\right)^2\right] = \frac{1}{N}\sum_{t=1}^{N}\left(\xi_k^{(-t)}(t|t-1)\right)^2. \qquad (9)$$

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

$$n_\theta = \arg\min_{1\leq k\leq K}\left[E\left[\left(\xi_k^{(-t)}(t|t-1)\right)^2\right]\right]. \qquad (10)$$

For linear-in-the-parameters models, the PRESS statistic can be generated without actually sequentially splitting the training data set and repeatedly estimating the associated models [8]. 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 [8],[9]:

$$\begin{aligned} \xi_M^{(-t)}(t|t-1) &= y(t) - \hat{y}_M^{(-t)}(t|t-1) \\ &= \frac{\xi_M(t)}{1 - \mathbf{p}^T(k)\left(\mathbf{P}^T\mathbf{P}\right)^{-1}\mathbf{p}(k)}, \end{aligned} \qquad (11)$$

where $\xi_M(t) = y(t) - \hat{y}_M(t|t-1)$. Obviously, choosing the best subset model that minimizes the PRESS statistic quickly becomes computationally prohibitive even for a modest M-term model set. Moreover, the PRESS error (11) itself is computational expensive because the matrix inversion involved. However, if we choose only to incrementally minimize 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.

It can readily be shown that the PRESS error $\xi_M^{(-t)}(t|t-1)$ for the M-term orthogonal weight model (7) is given by:

$$\xi_M^{(-t)}(t|t-1) = y(t) - \hat{y}_M^{(-t)}(t|t-1)$$

$$= \frac{\xi_M(t)}{1 - \mathbf{w}(t)^T\left(\mathbf{W}^T\mathbf{W} + \boldsymbol{\Lambda}\right)^{-1}\mathbf{w}(t)} = \frac{\xi_M(t)}{\beta_M(t)} \qquad (12)$$

assuming that regularization is applied with a regularization parameter λ, where $\mathbf{\Lambda} = diag\{\lambda, \cdots, \lambda\}$ is an $M \times M$ diagonal matrix and

$$\beta_M(t) = 1 - \sum_{i=1}^{M} \frac{w_i^2(t)}{\kappa_i + \lambda}. \tag{13}$$

Consider the orthogonal forward regression, 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 (12) and (13) can be written, by replacing M with a variable model size k, as

$$\xi_k^{(-t)}(t|t-1) = \frac{\xi_k(t)}{\beta_k(t)} \tag{14}$$

where

$$\beta_k(t) = 1 - \sum_{i=1}^{k} \frac{w_i^2(t)}{\kappa_i + \lambda} \tag{15}$$

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

$$\beta_k(t) = \beta_{k-1}(t) - \frac{w_k^2(t)}{\kappa_k + \lambda}. \tag{16}$$

As is in the conventional orthogonal LS algorithm [5], 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:

$$\begin{aligned} J_k &= E\left[\left(\xi_k^{(-t)}(t|t-1)\right)^2\right] \\ &= E\left[\frac{\xi_k^2(t)}{\beta_k^2(t)}\right] = \frac{1}{N}\sum_{t=1}^{N} \frac{\xi_k^2(t)}{\beta_k^2(t)} \end{aligned} \tag{17}$$

and this is used as the regressor selective criterion for the model construction which minimizes this mean square PRESS error. Note that 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 [11]. This property, i.e. $\Delta J = J_{k+1} - J_k$ changes the sign at certain model size k, can be applied to construct the automatic algorithm.

The proposed algorithm selects significant regressors that minimizes the PRESS statistic, with a growing model structure until $\Delta J > 0$ at a desired model size n_θ, where the contribution of the $(n_\theta+1)$th regressor in model approximation becomes insignificant. Thus the algorithm terminates at $J_{n_{\theta+1}} > J_{n_\theta}$, where the model is optimized based on the minimization 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. The proposed algorithm based on the standard Gram-Schmidt procedure is summarized in Appendix, in which the orthogonal basis $\mathbf{w}_i$ is constructed in a forward regression manner. In this algorithm a small fixed positive regularization 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.

4 Numerical examples

Two examples were used to demonstrate the effectiveness of the proposed model construction algorithm.

Example 1. Consider using a radial basis function (RBF) network to approximate an unknown scalar function

$$f(x) = \frac{\sin(x)}{x}, \ -10 \le x \le 10. \tag{18}$$

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

$$p_i(\mathbf{x}) = \exp\left(-\frac{\|\mathbf{x} - \mathbf{c}_i\|^2}{2\tau^2}\right) \tag{19}$$

was used, with a kernel width $\tau^2 = 10.0$. All the two hundred training data points were used as the candidate RBF center set for $\mathbf{c}_i$. 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 regularization parameter was fixed to $\lambda = 0.001$.

Fig. 1 depicts the evolution of the training MSE and PRESS statistic in log scale during the orthogonal forward regression with a typical set of noisy training data using the proposed algorithm. 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

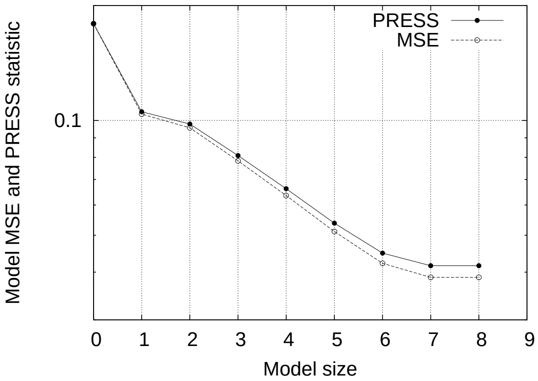


Figure 1: Evolution of training MSE and PRESS statistic versus model size for simple scalar function modelling.

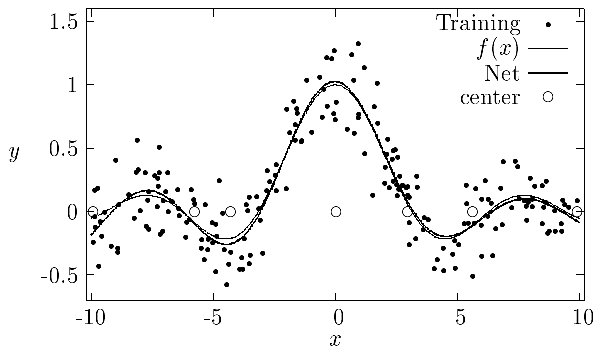


Figure 2: Simple scalar function modelling problem: a typical set of noisy training data y (dots), underlying function $f(x)$ (thin curve), model mapping (thick curve), and selected RBF centers (circles). The 7-term model was identified without the help of a validation set.

using this 7-term model identified. Table 1 summarizes the modelling accuracy (mean $\pm$ standard deviation) averaged over ten sets of different data realizations. It can be seen that the proposed algorithm was able to produce very sparse models with excellent generalization performance, without the need to use additional validation set 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 [12]. The data set, depicted in Fig. 3, 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

$$\mathbf{x}(t) = [y(t-1)\; u(t-1)\; u(t-2)]^T \tag{20}$$

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. 4 shows the evolution of the training MSE and PRESS statistic during the forward regression procedure, 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.

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

model terms	7.8 ± 0.6
MSE over training set	0.037703 ± 0.003708
PRESS statistic	0.040725 ± 0.003893
MSE over noisy test set	0.041692 ± 0.002458
MSE over noise-free test set	0.001749 ± 0.000630

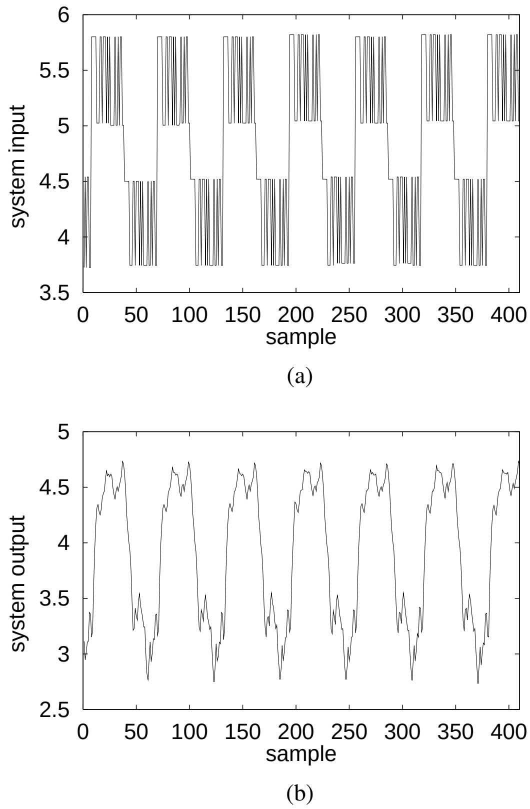


Figure 3: Engine data set (a) input $u(t)$ and (b) output $y(t)$.

The modelling accuracy is summarized in Table 2. The constructed RBF model $\hat{f}_{RBF}(\bullet)$ was used to generate the model prediction according to

$$\hat{y}(t) = \hat{f}_{\mathrm{RBF}}(\mathbf{x}(t)) \tag{21}$$

with the input vector $\mathbf{x}(t)$ given by (20). Fig. 5 depicts the model prediction $\hat{y}(t)$ and the prediction error $\xi(t) = y(t) - \hat{y}(t)$ for the 23-term model constructed. Again, it is seen that the proposed algorithm was able to produce very sparse models with excellent generalization performance, without the need to use additional validation set for model evaluation during the model construction process.

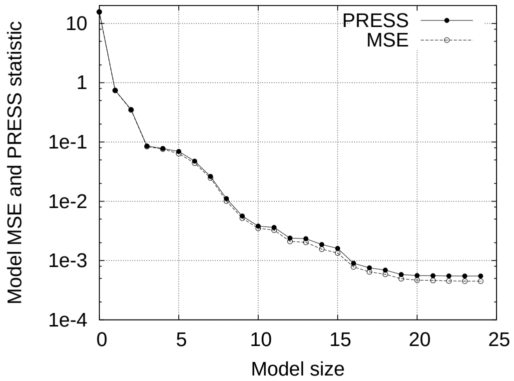


Figure 4: Evolution of training MSE and PRESS statistic versus model size for engine data set modelling.

Table 2: Modelling accuracy for engine data set modelling.

model terms	23
MSE over training set	0.000449
PRESS statistic	0.000548
MSE over test set	0.000487

5 Conclusions

This paper has introduced an automatic model construction algorithm for linear-in-the-parameters nonlinear models based directly on maximizing model generalization capability. The leave-one-out test score or PRESS statistic in the framework of regularized 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 orthogonal forward regression combines parameter regularization technique in orthogonal weight space and the PRESS statistic to optimize model structure in order to achieve improved generalization capability, without resorting to another validation data set for model assessment.

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

1. Initialization: initialize $J_0 = \mathbf{y}^T\mathbf{y}$, $\xi_0(t) = y(t)$ and $\beta_0(t) = 1$ for $t = 1, \cdots, N$. For $1 \le i \le M$, compute

$$\begin{aligned}
\mathbf{w}_1^{(i)} &= \mathbf{p}_i, \\
\kappa_1^{(i)} &= \left(\mathbf{w}_1^{(i)}\right)^T \mathbf{w}_1^{(i)}, \\
g_1^{(i)} &= \frac{\left(\mathbf{w}_1^{(i)}\right)^T \mathbf{y}}{\left(\mathbf{w}_1^{(i)}\right)^T \mathbf{w}_1^{(i)} + \lambda}, \\
\xi_1^{(i)}(t) &= \xi_0(t) - w_1^{(i)}(t) g_1^{(i)}, \ t = 1, \cdots, N, \\
\beta_1^{(i)}(t) &= \beta_0(t) - \frac{\left(w_1^{(i)}(t)\right)^2}{\kappa_1^{(i)} + \lambda}, \ t = 1, \cdots, N, \\
J_1^{(i)} &= \frac{1}{N} \sum_{t=1}^{N} \frac{\left(\xi_1^{(i)}(t)\right)^2}{\left(\beta_1^{(i)}(t)\right)^2}.
\end{aligned}$$

Find

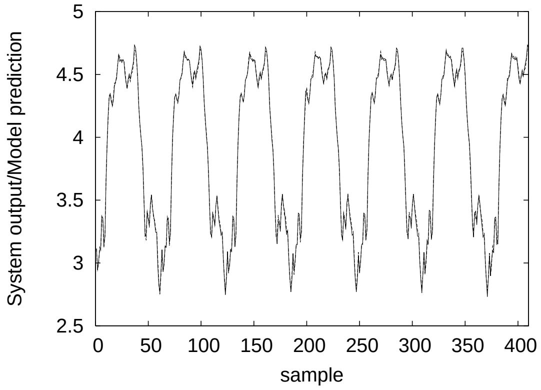
$$i_1 = \arg\min\{J_1^{(i)}, \ 1 \le i \le M\}$$

and select

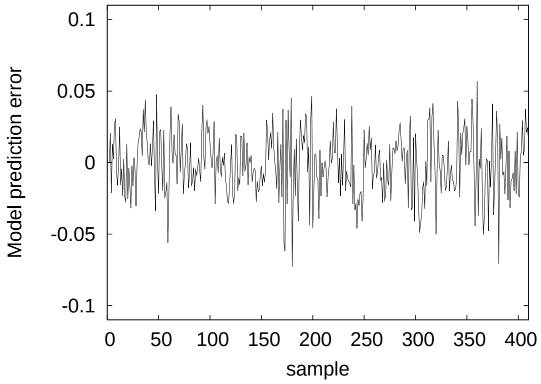
$$\mathbf{w}_1 = \mathbf{w}_1^{(i_1)} = \mathbf{p}_{i_1}$$

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

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



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



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

Figure 5: Modelling performance for engine data set modelling problem. The 23-term model was constructed without the help of a validation set.

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

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

$$\begin{aligned}
a_{jk}^{(i)} &= \frac{\mathbf{w}_j^T \mathbf{p}_i}{\mathbf{w}_j^T \mathbf{w}_j}, \ 1 \le j < k, \\
\mathbf{w}_k^{(i)} &= \mathbf{p}_i - \sum_{j=1}^{k-1} a_{jk}^{(i)} \mathbf{w}_j, \\
\kappa_k^{(i)} &= \left(\mathbf{w}_k^{(i)}\right)^T \mathbf{w}_k^{(i)}, \\
g_k^{(i)} &= \frac{\left(\mathbf{w}_k^{(i)}\right)^T \mathbf{y}}{\left(\mathbf{w}_k^{(i)}\right)^T \mathbf{w}_k^{(i)} + \lambda}, \\
\xi_k^{(i)}(t) &= \xi_{k-1}(t) - w_k^{(i)}(t) g_k^{(i)}, \ t = 1, \cdots, N, \\
\beta_k^{(i)}(t) &= \beta_{k-1}(t) - \frac{\left(w_k^{(i)}(t)\right)^2}{\kappa_k^{(i)} + \lambda}, \ t = 1, \cdots, N,
\end{aligned}$$

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

Find

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

and select

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

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 $J_k \ge J_{k-1}$. Otherwise, set $k = k + 1$, and go to step 2.

References

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

[2] Chen, S., 2002, "Locally regularised orthogonal least squares algorithm for the construction of sparse kernel regression models," in *Proc. 6th Int. Conf. Signal Processing* (Beijing, China), Aug.26-30, 2002, pp.1229–1232.

[3] Chen, S., Hong, X., and Harris, C.J., 2002, "Sparse data modelling using combined locally regularized orthogonal least squares and D-optimality design," in: *Proc. Combined Annual Conf. Institute of Automation, the Chinese Academy of Sciences, and Annual Conf. Chinese Automation and Computer Science Society in U.K.* (Beijing, China), Sept.20-21, 2002, pp.112-117.

[4] Chen, S., Hong, X., and Harris, C.J., 2003, "Sparse kernel regression modelling using combined locally regularized orthogonal least squares and D-optimality experimental design," *IEEE Trans. Automatic Control*, to appear, June 2003.

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

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

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

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

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

[10] Chen, S., and Billings, S.A., 1989, "Representation of non-linear systems: the NARMAX model," *Int. J. Control*, Vol.49, No.3, pp.1013–1032.

[11] 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*, to appear.

[12] 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.