Orthogonal Forward Regression based on Directly Maximizing Model Generalization Capability

S. Chen† and X. Hong‡

† School of Electronics and Computer Science
University of Southampton, Southampton SO17 1BJ
E-mail: sqc@ecs.soton.ac.uk

‡ Department of Cybernetics
University of Reading, Reading RG6 6AY
E-mail: x.hong@reading.ac.uk

Presented at CACSCUK'2003, Luton, U.K., 20 September, 2002

Motivation

Modelling from data: *generalization*, *interpretability*, *knowledge extraction* $\Longrightarrow$ all depend on ability to construct **appropriate** sparse models

◯ Main engine or criterion in most of subset model selection algorithms:

minimizing **training mean square error**

◯ It is highly desired to be able to construct sparse models by:

directly maximizing **model generalization capability**

◯ Cross validation via delete-one approach:

leave-one-out (LOO) test score, a measure of generalization

Delete-1 Approach with Leave-One-Out Score

◯ Concept of delete-1 with associated leave-one-out test score

◯ For linear-in-the-parameter models, no need to sequentially splitting training data set and repeatedly estimating associated models

Even so and even with only incrementally minimizing LOO test score, complexity becomes prohibitive for a modest model set

◯ Adopting orthogonal forward regression, model construction using LOO test score becomes computationally affordable

◯ Proposed OLS: incrementally minimizing LOO test score (generalization error) using just one training data set

Original OLS: incrementally minimizing training error

Regression Model

$$y(t) = \sum_{i=1}^{n_M} \theta_i \phi_i(t) + e(t) = \boldsymbol{\phi}^T(t)\boldsymbol{\theta} + e(t), \;\; 1 \le t \le N$$

$y(t)$: target or desired output, $e(t)$: model error, θ_i: model weights and $\boldsymbol{\theta} = [\theta_1 \cdots \theta_{n_M}]^T$, $\phi_i(t)$: regressors and $\boldsymbol{\phi}(t) = [\phi_1(t) \cdots \phi_{n_M}(t)]^T$, n_M: number of candidate regressors, and N: number of training samples.

Defining

$$\mathbf{y} = [y(1) \cdots y(N)]^T, \quad \mathbf{e} = [e(1) \cdots e(N)]^T, \quad \boldsymbol{\Phi} = [\boldsymbol{\phi}_1 \cdots \boldsymbol{\phi}_{n_M}]$$

with $\boldsymbol{\phi}_i = [\phi_i(1) \cdots \phi_i(N)]^T$, leads to matrix form

$$\mathbf{y} = \boldsymbol{\Phi}\boldsymbol{\theta} + \mathbf{e}$$

Note that $\boldsymbol{\phi}(t)$ is t-th row of $\boldsymbol{\Phi}$ and $\boldsymbol{\phi}_i$ is i-th column of $\boldsymbol{\Phi}$.

Orthogonalization

Orthogonal decomposition: $\mathbf{\Phi} = \mathbf{W}\mathbf{A}$, where

$$\mathbf{A} = \begin{bmatrix} 1 & a_{1,2} & \cdots & a_{1,n_M} \\ 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & a_{n_M-1,n_M} \\ 0 & \cdots & 0 & 1 \end{bmatrix}$$

and $\mathbf{W} = [\mathbf{w}_1 \cdots \mathbf{w}_{n_M}]$ with orthogonal columns: $\mathbf{w}_i^T \mathbf{w}_j = 0$, if $i \neq j$.
Let $\mathbf{g} = [g_1 \cdots g_{n_M}]^T$, satisfying $\mathbf{A}\boldsymbol{\theta} = \mathbf{g}$. Regression model becomes

$$\mathbf{y} = \mathbf{W}\mathbf{g} + \mathbf{e}$$

or

$$y(t) = \mathbf{w}^T(t)\mathbf{g} + e(t), \quad 1 \le t \le N$$

Note that $\mathbf{w}(t)$ is t-th row of $\mathbf{W}$ and $\mathbf{w}_i$ is i-th column of $\mathbf{W}$.

Leave-One-Out Generalization Error

Denoting k-term model error as $e_k(t)$, then LOO error for k-term model is

$$e_k^{(-t)}(t) = \frac{e_k(t)}{\beta_k(t)}$$

where super-index $^{(-t)}$ indicates that the model is obtained with t-th training sample removed, and LOO error weighting $\beta_k(t)$ is computed recursively

$$\beta_k(t) = \beta_{k-1}(t) - \frac{w_k^2(t)}{\mathbf{w}_k^T \mathbf{w}_k + \lambda}$$

where λ is a regularization parameter.

The LOO mean square error or test score is given by:

$$J_k = E\left[\left(e_k^{(-t)}(t)\right)^2\right] = \frac{1}{N}\sum_{t=1}^{N} \frac{e_k^2(t)}{\beta_k^2(t)}$$

Model Construction Algorithm

○ At selection step k, a model term is selected if it produces the smallest LOO test score J_k among the candidate model terms k to n_M.

In this algorithm,

$$J_k = \frac{1}{N}\sum_{t=1}^{N} \frac{e_k^2(t)}{\beta_k^2(t)}$$

This should be compared with original OLS with

$$J_k = \frac{1}{N}\sum_{t=1}^{N} e_k^2(t)$$

○ The model construction process is **fully automatic**, and ends with a n_θ-term model when

$$\Delta J = J_{n_\theta+1} - J_{n_\theta} \ge 0$$

User does not need to specify any separate termination criterion.

A Simple Scalar Function Modelling

$$f(x) = \frac{\sin(x)}{x}, \quad -10 \le x \le 10$$

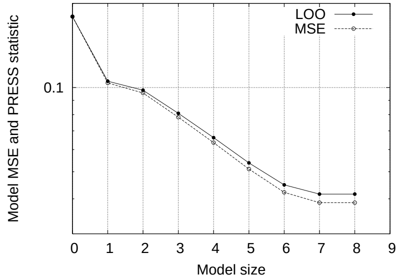
Give $y = f(x) + \epsilon$ and x. 400 x uniform distribution in $[-10,\ 10]$ and ϵ zero mean Gaussian with variance 0.04. First 200 samples as training set, the other 200 as testing set. Additional test set with 200 noise-free $f(x)$.

The RBF Gaussian kernel function with variance of 10.0. Each training data was considered as a candidate RBF center and $n_M = 200$. Regularization parameter fixed to $\lambda = 0.001$.

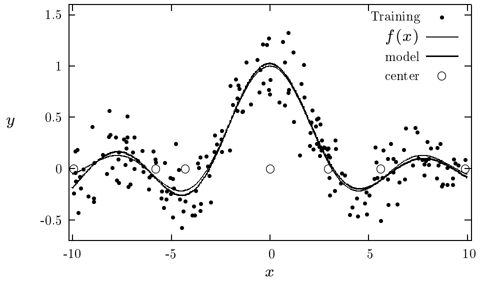
• Modelling accuracy (mean±std) averaged over ten different sets of data realizations

model terms	7.8 ± 0.6
MSE (noisy training set)	0.037703 ± 0.003708
LOO test score	0.040725 ± 0.003893
MSE (noisy test set)	0.041692 ± 0.002458
MSE (noise-free test set)	0.001749 ± 0.000630

- Training MSE and LOO test score in $\log$ scale for a typical set of noisy training data. Note the algorithm terminated with a 7-term model when $J_8 = 0.041589 \geq J_7 = 0.041589$.

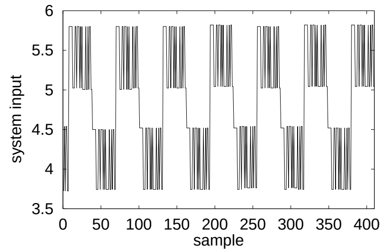


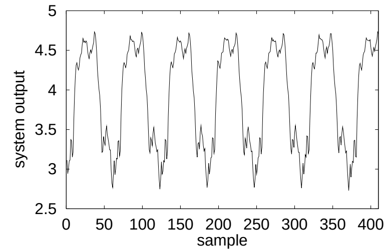
- The noisy training points y and the underlying function $f(x)$ together with the mapping generated using this 7-term model identified.



Engine Data Modelling

System input $u(t)$ and output $y(t)$





First 210 data points for modelling, last 200 points for testing

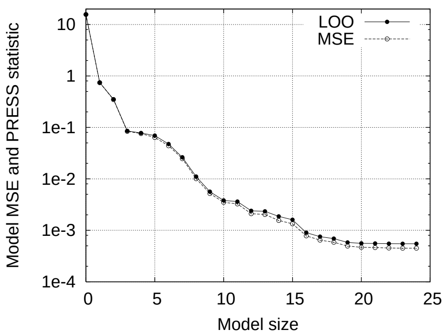
RBF model:

$$\hat{y}(t) = \hat{f}_{RBF}(y(t-1), u(t-1), u(t-2))$$

Gaussian kernel function variance 1.69. Regularization parameter fixed to 10^{-7}

Modelling Results

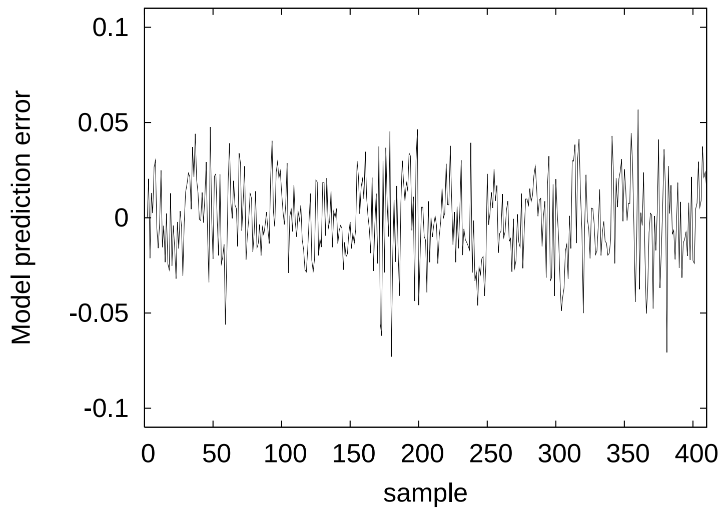
- Training MSE and LOO test score in $\log$ scale for engine data set. Note the algorithm terminated with a 23-term model when $J_{24} = 0.000548 \geq J_{23} = 0.000548$.



- Modelling accuracy for engine data set.

model terms	23
MSE over training set	0.000449
LOO test score	0.000548
MSE over test set	0.000487

- Modelling error $y(t) - \hat{y}(t)$ by the constructed 23-term model:



Conclusions

- A fully automatic model construction algorithm for linear-in-the-parameters nonlinear models has been developed based directly on maximizing model generalization capability

- The leave-one-out test score in the framework of regularized orthogonal least squares has been derived and, in particular, an efficient recursive computation formula for LOO errors has been presented

- The proposed algorithm is based on orthogonal forward regression with LOO test score to optimize model structure without resorting to another validation data set for model assessment