Locally Regularised Orthogonal Least Squares Algorithm for the Construction of Sparse Kernel Regression Models

Sheng Chen

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

Presented at 6th International Conference on Signal Processing
Beijing, China, August 26-30, 2002

Motivation

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

○ Parsimonious principle: subset model selection

⋆ OLS: significance of individual selected terms

○ Bayesian learning: hyperparameters/regularisation to enforce sparsity

- Bayesian framework, maximum *a posteriori* (MAP)
  ⋆ Evidence procedure ⋆ Markov chain Monte Carlo sampling
  ⋆ variational learning method
- Kernel-based data modelling
  ⋆ support vector machines (structural risk minimisation)
  ⋆ relevance vector machines (individual hyperparameters)

○ **OLS with individual regularisation.**

Regression Model

$$y(k) = \hat{y}(k) + e(k) = \sum_{i=1}^{n_M} \theta_i \phi_i(k) + e(k), \; 1 \le k \le N$$

$y(k)$: target or desired output, $e(k) = y(k) - \hat{y}(k)$, $\hat{y}(k)$: model output, θ_i: model weights, $\phi_i(k)$: regressors, n_M: number of candidate regressors, N: number of training samples.

Defining

$$\mathbf{y} = [y(1) \cdots y(N)]^T, \quad \mathbf{e} = [e(1) \cdots e(N)]^T, \quad \boldsymbol{\theta} = [\theta_1 \cdots \theta_{n_M}]^T$$

$$\boldsymbol{\Phi} = [\boldsymbol{\phi}_1 \cdots \boldsymbol{\phi}_{n_M}] \quad \text{with} \quad \boldsymbol{\phi}_i = [\phi_i(1) \cdots \phi_i(N)]^T$$

leads to matrix form

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

Orthogonalisation

Orthogonal decomposition: $\boldsymbol{\Phi} = \mathbf{W}\mathbf{A}$, 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 $i \ne j$.

Regression model becomes

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

with orthogonal weight vector $\mathbf{g} = [g_1 \cdots g_{n_M}]^T$ satisfying

$$\mathbf{A}\boldsymbol{\theta} = \mathbf{g}$$

Locally Regularised OLS Regression

Given regularisation parameter vector $\boldsymbol{\lambda} = [\lambda_1 \cdots \lambda_{n_M}]^T$ and denoting $\boldsymbol{\Lambda} = \text{diag}\{\lambda_1, \cdots, \lambda_{n_M}\}$, locally regularised error criterion:

$$J_R(\mathbf{g}, \boldsymbol{\lambda}) = \mathbf{e}^T\mathbf{e} + \mathbf{g}^T\boldsymbol{\Lambda}\mathbf{g} = \mathbf{y}^T\mathbf{y} - \sum_{i=1}^{n_M} \left(\mathbf{w}_i^T\mathbf{w}_i + \lambda_i\right) g_i^2$$

• Forward-regression procedure selects significant regressors according to regularised error reduction ratio due to each regressor $\mathbf{w}_i$

$$[\text{rerr}]_i = \left(\mathbf{w}_i^T\mathbf{w}_i + \lambda_i\right) g_i^2/\mathbf{y}^T\mathbf{y}$$

Selection terminated with n_s-term sub-model at the n_s-th stage when

$$1 - \sum_{l=1}^{n_s} [\text{rerr}]_l < \xi$$

Regularisation Parameter Update

Bayesian evidence procedure for updating regularisation parameters:

$$\lambda_i^{\text{new}} = \frac{\gamma_i^{\text{old}}}{N - \gamma^{\text{old}}} \frac{\mathbf{e}^T\mathbf{e}}{g_i^2}, \quad 1 \le i \le n_M$$

where

$$\gamma_i = \frac{\mathbf{w}_i^T\mathbf{w}_i}{\lambda_i + \mathbf{w}_i^T\mathbf{w}_i} \quad \text{and} \quad \gamma = \sum_{i=1}^{n_M} \gamma_i$$

Special cases of this LROLS — original OLS: $\lambda_i = 0$, $\forall i$

Initialization. Set all λ_i to same small positive value (e.g. 0.001).

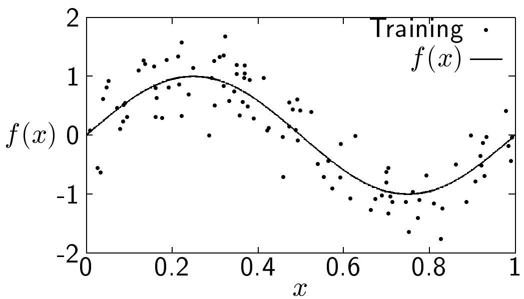
Step 1. Given current $\boldsymbol{\lambda}$, orthogonal forward procedure selects n_s-term subset model.

Step 2. Update $\boldsymbol{\lambda}$. If $\boldsymbol{\lambda}$ remains sufficiently unchanged in two successive iterations or a pre-set maximum iteration number is reached, stop; otherwise go to *Step 1.*

A Simple Example

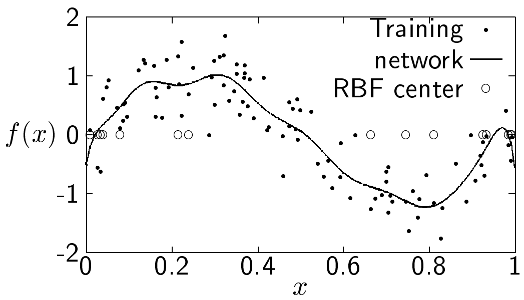
Modelling $f(x)$ given $y = f(x) + \epsilon$ and x. 100 x uniform distribution in $(0,\ 1)$ and ϵ zero mean Gaussian with variance 0.16.

The RBF Gaussian kernel function with variance of 0.04. Each training data was considered as a candidate RBF center and $n_M = 100$.



stage l	accuracy $1 - \sum[\text{err}]_l$	weight θ_l	
1	0.6461718264	2.60935e+06	
2	0.2840641827	-2.28370e+06	
3	0.2416057207	-1.29831e+08	
4	0.2260673781	-2.21722e+09	
5	0.2189319619	3.63027e+08	
6	0.2179112365	1.66438e+09	
7	0.2169210404	-3.19282e+09	
8	0.2156145110	1.70011e+09	OLS
9	0.2135190658	4.06932e+09	Selection
10	0.2113153903	-1.94658e+09	
11	0.2108713704	-2.72236e+08	
12	0.2095033180	-4.28658e+07	
13	0.2093349973	5.60372e+06	
14	0.2091282455	-1.59224e+06	
15	0.2068241235	3.83400e+05	
stop due to no term selected at 16 stage			
MSE over noisy training set: 0.147430			

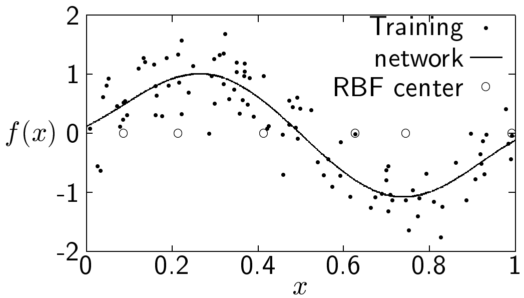
OLS Modelling Result



15-term model mapping (curve) produced by the OLS algorithm for the simple scalar function modelling problem. Dots indicate noisy training data y and circles the RBF centers.

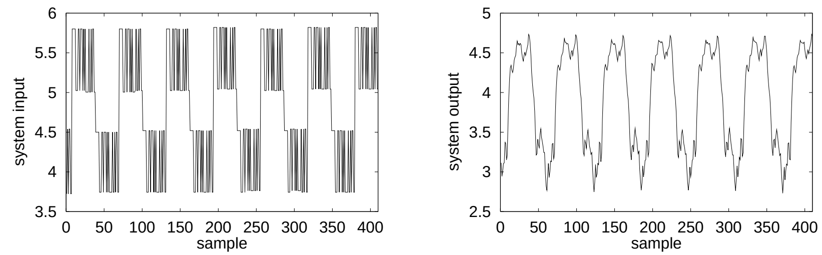
stage l	accuracy $1 - \sum[\text{rerr}]_l$	weight θ_l	regularizer λ_l	
1	0.6485054202	1.87494e+00	2.53227e-01	
2	0.2887313702	-1.70014e+00	1.81540e-01	
3	0.2500895914	-1.00970e+00	2.01490e-01	
4	0.2349327688	5.67310e-01	8.64601e-01	
5	0.2336724743	4.17979e-01	1.36357e+00	
6	0.2332827490	-1.51352e-01	6.93984e-01	LROLS
7	0.2332827490	-9.49873e-10	5.67623e+07	Selection
8	0.2332827490	-2.79967e-10	1.11770e+08	
9	0.2332827490	7.14157e-11	1.03860e+07	
10	0.2332827490	-2.05313e-12	1.92708e+08	
11	0.2332827490	-1.32386e-13	7.85977e+08	
12	0.2332827490	2.29641e-14	4.09979e+08	
13	0.2332827490	-2.53260e-38	1.15132e+32	
stop due to no term selected at 14 stage				
MSE over noisy training set: 0.159167				

LROLS Modelling Result



6-term model mapping (curve) produced by the LROLS algorithm for the simple scalar function modelling problem. Dots indicate noisy training data y and circles the RBF centers.

Engine Data Modelling



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

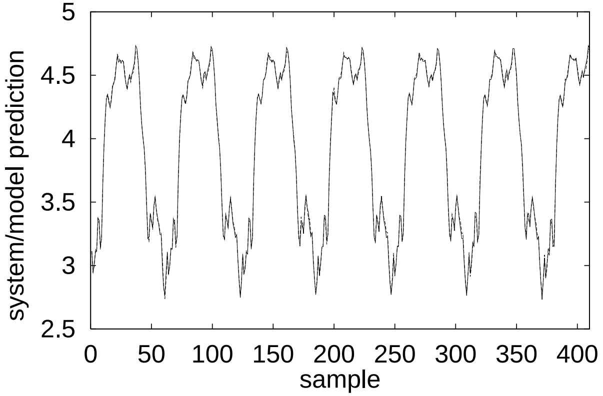
RBF model: $\hat{y}(k) = f_{RBF}(y(k-1), u(k-1), u(k-2))$

model	MSE for training	MSE for testing
60-term (OLS)	0.000336	0.000872
34-term (LROLS)	0.000435	0.000487

Model one-step prediction with 34-term model by LROLS

$$\hat{y}(k) = f_{RBF}(y(k-1), u(k-1), u(k-2))$$

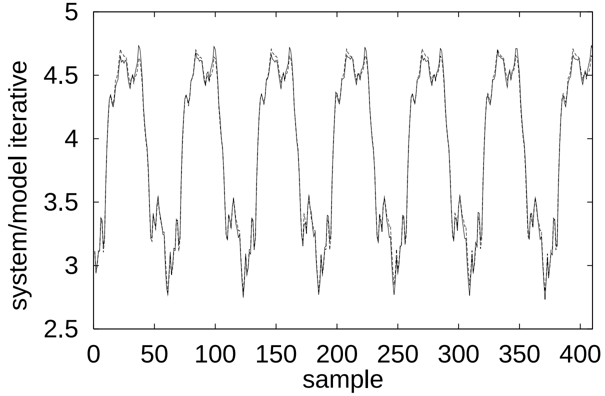
$y(k)$: solid $\hat{y}(k)$: dashed



Model iterative output with 34-term model by LROLS

$$\hat{y}_d(k) = f_{RBF}(\hat{y}_d(k-1), u(k-1), u(k-2))$$

$y(k)$: solid $\hat{y}_d(k)$: dashed



Conclusions

- Parsimonious principle based subset model selection
  - ⋆ OLS algorithm selects significant model terms
- Local regularisation re-enforces sparsity of selected model
  - ⋆ When to terminate subset model selection becomes obvious
- Combined algorithm is very efficient and capable of producing small-size models that generalize well