Sparse Generalised Kernel Modelling for Nonlinear Systems

S. Chen†, X. Hong‡, X.X. Wang§ and C.J. Harris†

† School of ECS
University of Southampton

‡ Department of Cybernetics
University of Reading

§ Neural Computing Group
Aston University
Outline

- Introduction
- Generalised Kernel Modelling
- A Sparse Model Construction Algorithm
  - Orthogonal Forward Selection
  - Leave-One-Out Criterion
  - Repeated Weighted Boosting Search
- Modelling Results
- Conclusions
Introduction

Nonlinear System Identification

- Modelling the nonlinear system

\[
y_k = f(y_{k-1}, \cdots, y_{k-n_y}, u_{k-1}, \cdots, u_{k-n_u}; \theta) + e_k
\]

\[
= f(x_k; \theta) + e_k
\]

based on a set of \( N \) training input-output data \( \{x_k, y_k\}_{k=1}^N \)

- \( u_k \) and \( y_k \) are the system input and output variables with \( n_u \) and \( n_y \) indicating the lags in the input and output, respectively

- \( \theta \) is the unknown parameter vector associated with the system model structure yet to be determined

- \( x_k = [y_{k-1} \cdots y_{k-n_y} u_{k-1} \cdots u_{k-n_u}]^T \), and \( e_k \) is the system noise
Existing Kernel Modellings

- Nonlinear optimisation to determine all the kernel centres, variances and weights
  - Local minimum and structure determination problems
- Clustering to determine kernel centres and variances
  - Structure determination problem
- Orthogonal Least Squares (OLS) forward selection, and sparse kernel methods, such as Support Vector Machine (SVM)
  - Select centres from data points and use cross validation to determine a single common kernel variance for every kernel basis
The Previous State-of-the-Art

- Model selection should be based on generalisation capability, rather than training performance, and Leave-One-Out (LOO) criterion is a measure of generalisation


- This Locally Regularised OLS with LOO (LROLS-LOO) selects kernel centres from training data and adopts a single common kernel variance for every selected kernel
Novelty of the Proposed Algorithm

- Extend to tunable kernels
  - Kernel centre is not restricted to training data, and each kernel has an individual diagonal covariance matrix
- Combine OLS / nonlinear optimisation
  - Orthogonal Forward Selection (OFS) to select kernels one by one
  - Each kernel is determined by nonlinear optimisation based on the LOO criterion
- This OFS-LOO algorithm enables
  - Enhanced modelling capability and sparser representation
Generalised Kernel

Generalised Kernel Model

• Generalised kernel modelling of the training data \( \{x_k, y_k\}_{k=1}^N \)

\[
y_k = \hat{y}_k + e_k = \sum_{i=1}^{M} w_i g_i(x_k) + e_k = g^T(k)w + e_k
\]

where \( M \) is the number of kernels, \( w = [w_1 \cdots w_M]^T \) the kernel weight vector, and \( g(k) = [g_1(x_k) \cdots g_M(x_k)]^T \) the kernel regressors

• Generic kernel regressor

\[
g_i(x) = K \left( \sqrt{(x - \mu_i)^T \Sigma_i^{-1}(x - \mu_i)} \right)
\]

where \( \mu_i \) is the \( i \)th kernel centre, \( \Sigma_i = \text{diag}\{\sigma_{i,1}^2, \cdots, \sigma_{i,m}^2\} \) the \( i \)th diagonal kernel covariance matrix, \( K(\bullet) \) the chosen kernel function
Orthogonal Decomposition

- The kernel model over the training set: \( y = Gw + e \), where the regression matrix \( G = [g_1 \cdots g_M] \)

- Orthogonal decomposition: \( G = PA \), where the orthogonal matrix \( P = [p_1 \cdots p_M] \) has orthogonal columns

- The regression model becomes: \( y = P\theta + e \), with \( \theta = Aw \)

- The space spanned by the original model bases is identical to the space spanned by the orthogonal model bases, and thus

\[
\hat{y}_k = g^T(k)w = p^T(k)\theta
\]

- \( g^T(k) \) is the \( k \)th row of \( G \) while \( g_k \) is the \( k \)th column of \( G \), and \( p^T(k) \) is the \( k \)th row of \( P \) while \( p_k \) is the \( k \)th column of \( P \)
Sparse Construction

Leave-One-Out criterion

- The LOO mean square error for the $n$-term kernel model

$$J_n = \frac{1}{N} \sum_{k=1}^{N} \left( e_{k}^{(n,-k)} \right)^2 = \frac{1}{N} \sum_{k=1}^{N} \left( \frac{e_{k}^{(n)}}{\eta_{k}^{(n)}} \right)^2$$

where $e_{k}^{(n,-k)}$ is the LOO modelling error, $e_{k}^{(n)}$ the usual modelling error, and $\eta_{k}^{(n)}$ the LOO weighting

- Computing the LOO criterion is very efficient, since

$$e_{k}^{(n)} = y_k - \sum_{i=1}^{n} \theta_i p_i(k) = e_{k}^{(n-1)} - \theta_n p_n(k)$$

$$\eta_{k}^{(n)} = 1 - \sum_{i=1}^{n} \frac{p_i^2(k)}{p_i^T p_i + \lambda} = \eta_{k}^{(n-1)} - \frac{p_n^2(k)}{p_n^T p_n + \lambda}$$

where $\lambda \geq 0$ is a small regularisation parameter
OFS-LOO Algorithm

- The algorithm constructs kernels one by one, i.e. at the $n$th stage, determines the $n$th kernel by minimising $J_n$

$$\min_{\mu_n, \Sigma_n} J_n(\mu_n, \Sigma_n)$$

- $J_n$ is at least locally convex, i.e. there exists an $M$ such that

$$J_{n-1} > J_n \text{ if } n \leq M \text{ and } J_M < J_{M+1}$$

- The construction procedure is terminated automatically, and the user does not need to specify any learning algorithmic parameter

- After construction, the LROLS-LOO can be called to optimise regularisation parameters and to further reduce the model size $M$
Sparse Construction

Position and Shape Kernel

- Determine the $n$th kernel centre $\mu_n$ and covariance matrix $\Sigma_n$ by minimizing $J_n(\mu_n, \Sigma_n)$ is a nonconvex nonlinear optimisation
  - Gradient-based techniques may be trapped at a local minimum
  - Global optimisation techniques are preferred, e.g. genetic algorithm
- We adopt a simple yet efficient global search algorithm called the Repeated Weighted Boosting Search (RWBS) to perform this task
Sparse Construction

**RWBS for Minimising** $J(u)$

**Outer Loop:** $N_G$ number of generations

- **Initialisation:** Keep the best solution found in the previous generation as $u_1$ and randomly choose rest of the population $u_2, \cdots, u_{P_S}$

**Inner Loop:** $N_I$ iterations

- Perform a convex combination

$$u_{P_S+1} = \sum_{i=1}^{P_S} \delta_i u_i \quad \text{where} \quad \delta_i \geq 0 \quad \text{and} \quad \sum_{i=1}^{P_S} \delta_i = 1$$

- The weightings $\delta_i$ are adapted by boosting to reflect goodness of $u_i$
- $u_{P_S+1}$ or its mirror image replaces the worst member in the population

*End of Inner Loop*

*End of Outer Loop*
Optimisation Example

Population size $P_S = 6$, number of inner iterations $N_I = 20$ and number of generations $N_G = 12$

100 random experiments, populations in all the 100 runs converge to the global minimum
Modelling the relationship between the fuel rack position (input $u_k$) and the engine speed (output $y_k$) for a Leyland TL11 turbocharged, direct injection diesel engine. Data set contains 410 pairs of input-output samples, modelled as $y_k = f(x_k) + e_k$ with $x_k = [y_{k-1}, u_{k-1}, u_{k-2}]^T$, first 210 data points for training and last 200 points for testing.
Engine Data Modelling

- The OFS-LOO using Gaussian kernels
  - The LOO mean square error as a function of model size for the engine data set
  - The OFS-LOO constructed 17 kernels
  - The LROLS-LOO reduced the model to 15 kernels

- The SVM and LROLS-LOO were also used for comparison
## Engine Data Results

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Kernel Type</th>
<th>Model Size</th>
<th>Training MSE</th>
<th>Test MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>fixed Gaussian</td>
<td>92</td>
<td>0.000447</td>
<td>0.000498</td>
</tr>
<tr>
<td>LROLS-LOO</td>
<td>fixed Gaussian</td>
<td>22</td>
<td>0.000453</td>
<td>0.000490</td>
</tr>
<tr>
<td>OFS-LOO</td>
<td>tunable Gaussian</td>
<td>15</td>
<td>0.000466</td>
<td>0.000480</td>
</tr>
</tbody>
</table>

![Graph 1](image1.png)

![Graph 2](image2.png)
Modelling Results

Gas Furnace Data

Modelling the relationship between the coded input gas feed rate (input $u_k$) and the CO$_2$ concentration (output $y_k$) for a gas furnace data set. Data set contains 296 pairs of input-output samples, modelled as $y_k = f(x_k) + e_k$ with $x_k = [y_{k-1} \ y_{k-2} \ y_{k-3} \ u_{k-1} \ u_{k-2} \ u_{k-3}]^T$, all the data points for training.
Gas Furnace Modelling

- The OFS-LOO using Gaussian kernels
  - The LOO mean square error as a function of model size for the gas furnace data set
  - The OFS-LOO constructed 16 kernels
  - The LROLS-LOO reduced the model to 15 kernels
- The SVM and LROLS-LOO were also used for comparison
## Gas Furnace Results

<table>
<thead>
<tr>
<th>algorithm</th>
<th>kernel type</th>
<th>model size</th>
<th>training MSE</th>
<th>LOO MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>fixed Gaussian</td>
<td>62</td>
<td>0.052416</td>
<td>0.054376</td>
</tr>
<tr>
<td>LROLS-LOO</td>
<td>fixed thin-plate-spline</td>
<td>28</td>
<td>0.053306</td>
<td>0.053685</td>
</tr>
<tr>
<td>OFS-LOO</td>
<td>tunable Gaussian</td>
<td>15</td>
<td>0.054306</td>
<td>0.054306</td>
</tr>
</tbody>
</table>

![Graph showing system output and model prediction](image1)

![Graph showing model prediction error](image2)
Modelling Results

Boston Housing Data

- **Boston Housing**: [http://www.ics.uci.edu/~mlearn/MLRepository.html](http://www.ics.uci.edu/~mlearn/MLRepository.html)
  - Data set comprises 506 data points with 14 variables
  - Predicting the median house value from the remaining 13 attributes

- **Modelling**: randomly selected 456 data points from the data set for training and used the remaining 50 data points to form test set
  - Average results were given over 100 repetitions

- The SVM, LROLS-LOO and OFS-LOO algorithms using Gaussian kernels

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Kernel Type</th>
<th>Model Size</th>
<th>Training MSE</th>
<th>Test MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>fixed</td>
<td>243.2 ± 5.3</td>
<td>6.7986 ± 0.4444</td>
<td>23.1750 ± 9.0459</td>
</tr>
<tr>
<td>LROLS-LOO</td>
<td>fixed</td>
<td>58.6 ± 11.3</td>
<td>12.9690 ± 2.6628</td>
<td>17.4157 ± 4.6670</td>
</tr>
<tr>
<td>OFS-LOO</td>
<td>tunable</td>
<td>34.6 ± 8.4</td>
<td>10.0997 ± 3.4047</td>
<td>14.0745 ± 3.6178</td>
</tr>
</tbody>
</table>
Conclusions

- A construction algorithm has been proposed for nonlinear system identification using the generalised kernel model
  - The algorithm has ability to tune the centre and covariance matrix of individual kernel to minimise the leave-one-out error
  - A global search algorithm is used to construct the generalised kernel model in an orthogonal forward selection procedure
  - The model construction procedure is fully automatic and user does not need to specify any learning algorithmic parameter

- It offers enhanced modelling capability with sparser representation

S. Chen wish to thank the support of the United Kingdom Royal Academy of Engineering.