CDC-ECC 2005, Seville

# 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

# Nonlinear System Identification

- Modelling the nonlinear system

$$\begin{aligned} y_k &= f(y_{k-1}, \cdots, y_{k-n_y}, u_{k-1}, \cdots, u_{k-n_u}; \theta) + e_k \\ &= f(\mathbf{x}_k; \theta) + e_k \end{aligned}$$

  based on a set of $N$ training input-output data $\{\mathbf{x}_k, y_k\}_{k=1}^N$

- $u_k$ 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
- $\mathbf{x}_k = [y_{k-1} \cdots y_{k-n_y} \; u_{k-1} \cdots u_{k-n_u}]^T$, and $e_k$ is the system noise

# Existing Kernel Modellings

- Nonlinear optimisation to determine all the kernel centres, variances and weights

  $\Downarrow$ Local minimum and structure determination problems

- Clustering to determine kernel centres and variances

  $\Downarrow$ Structure determination problem

- Orthogonal Least Squares (OLS) forward selection, and sparse kernel methods, such as Support Vector Machine (SVM)

  $\Diamond$ 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
- S. Chen, X. Hong, C.J. Harris and P.M. Sharkey, “Sparse modelling using orthogonal forward regression with PRESS statistic and regularisation,” *IEEE Trans. Systems, Man and Cybernetics, Part B*, 34(2), 898–911, 2004
- 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 Model

- Generalised kernel modelling of the training data $\{\mathbf{x}_k, y_k\}_{k=1}^N$

$$y_k = \hat{y}_k + e_k = \sum_{i=1}^{M} w_i g_i(\mathbf{x}_k) + e_k = \mathbf{g}^T(k)\mathbf{w} + e_k$$

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

- Generic kernel regressor

$$g_i(\mathbf{x}) = K\left(\sqrt{(\mathbf{x} - \mu_i)^T \mathbf{\Sigma}_i^{-1} (\mathbf{x} - \mu_i)}\right)$$

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

# Orthogonal Decomposition

- The kernel model over the training set: $\mathbf{y} = \mathbf{G}\mathbf{w} + \mathbf{e}$, where the regression matrix $\mathbf{G} = [\mathbf{g}_1 \cdots \mathbf{g}_M]$
- Orthogonal decomposition: $\mathbf{G} = \mathbf{P}\mathbf{A}$, where the orthogonal matrix $\mathbf{P} = [\mathbf{p}_1 \cdots \mathbf{p}_M]$ has orthogonal columns
- The regression model becomes: $\mathbf{y} = \mathbf{P}\theta + \mathbf{e}$, with $\theta = \mathbf{A}\mathbf{w}$
- The space spanned by the original model bases is identical to the space spanned by the orthogonal model bases, and thus

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

- $\mathbf{g}^T(k)$is the $k$th row of $\mathbf{G}$ while $\mathbf{g}_k$ is the $k$th column of $\mathbf{G}$, and $\mathbf{p}^T(k)$ is the $k$th row of $\mathbf{P}$ while $\mathbf{p}_k$ is the $k$th column of $\mathbf{P}$

# 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

$$\begin{aligned} e_k^{(n)} &= y_k - \sum_{i=1}^{n}\theta_i p_i(k) = e_k^{(n-1)} - \theta_n p_n(k) \\ \eta_k^{(n)} &= 1 - \sum_{i=1}^{n}\frac{p_i^2(k)}{\mathbf{p}_i^T\mathbf{p}_i + \lambda} = \eta_k^{(n-1)} - \frac{p_n^2(k)}{\mathbf{p}_n^T\mathbf{p}_n + \lambda} \end{aligned}$$

where $\lambda \geq 0$ is a small regularisation parameter


University
of Southampton

# OFS-LOO Algorithm

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

$$\min_{\mu_n, \mathbf{\Sigma}_n} J_n(\mu_n, \mathbf{\Sigma}_n)$$

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

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

- 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 $M$

# Position and Shape Kernel

- Determine the $n$th kernel centre $\mu_n$ and covariance matrix $\boldsymbol{\Sigma}_n$ by minimising $J_n(\mu_n, \boldsymbol{\Sigma}_n)$ 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
- S. Chen, X.X. Wang and C.J. Harris, "Experiments with repeating weighted boosting search for optimisation in signal processing applications," *IEEE Trans. Systems, Man and Cybernetics, Part B*, 35(4), 682-693, 2005

# RWBS for Minimising $J(\mathbf{u})$

*Outer Loop*: $N_G$ number of generations

*Initialisation*: Keep the best solution found in the previous generation as $\mathbf{u}_1$ and randomly choose rest of the population $\mathbf{u}_2, \cdots, \mathbf{u}_{P_S}$

*Inner Loop*: $N_I$ iterations

- Perform a convex combination

$$\mathbf{u}_{P_S+1} = \sum_{i=1}^{P_S} \delta_i \mathbf{u}_i \quad \text{where} \quad \delta_i \geq 0 \text{ and } \sum_{i=1}^{P_S} \delta_i = 1$$

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

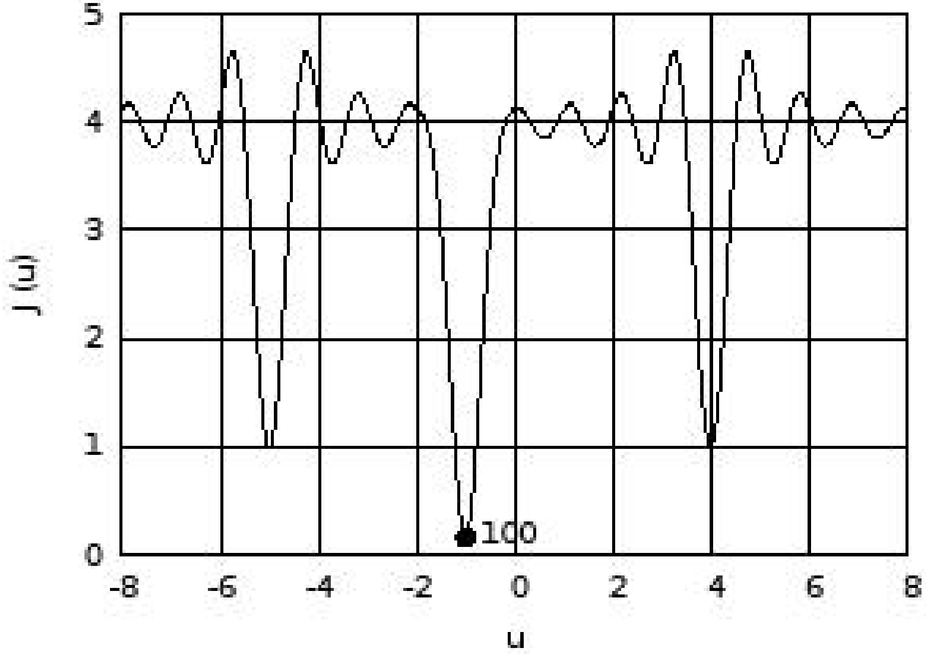
*End of Inner Loop*

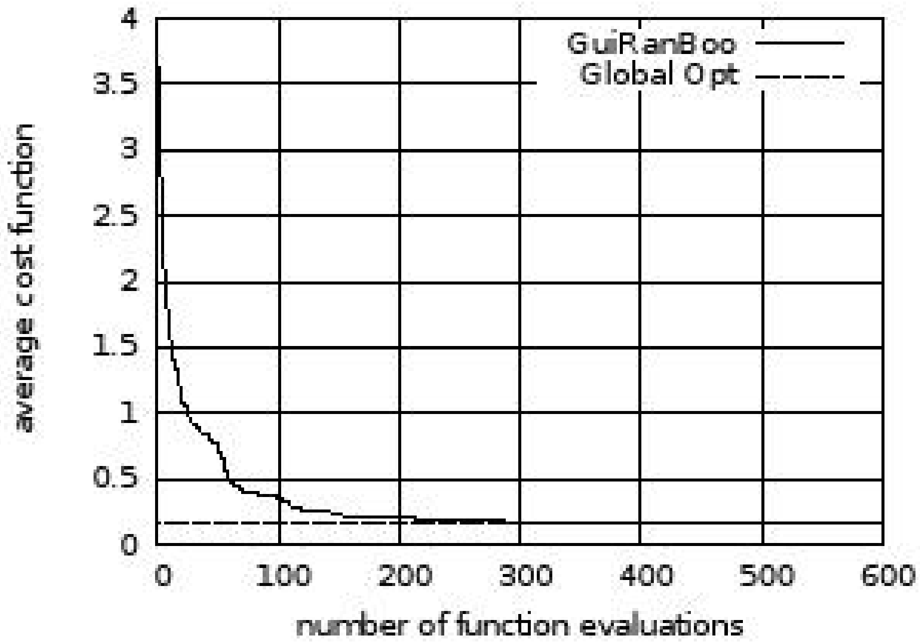
*End of Outer Loop*

# 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

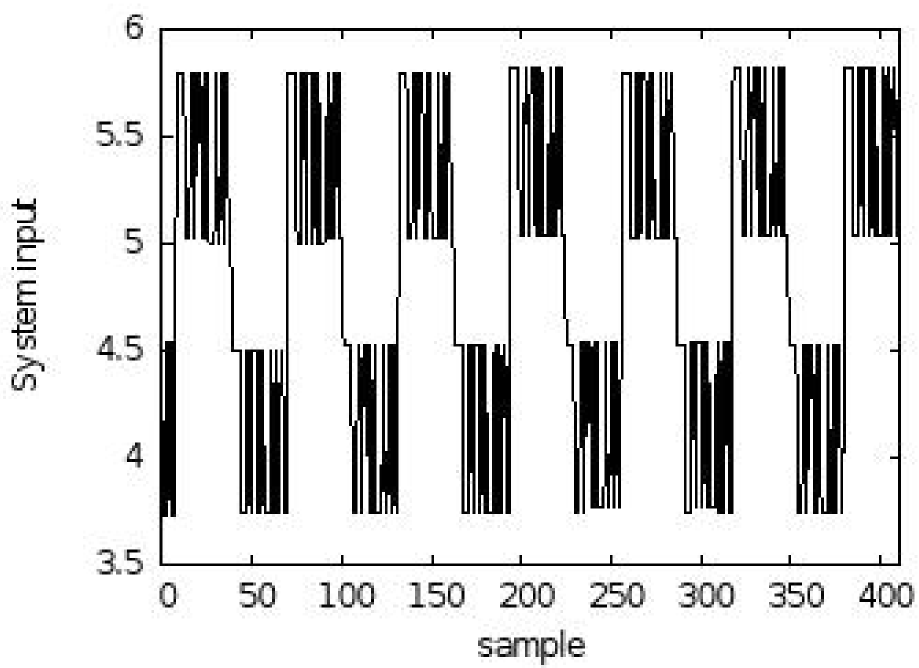


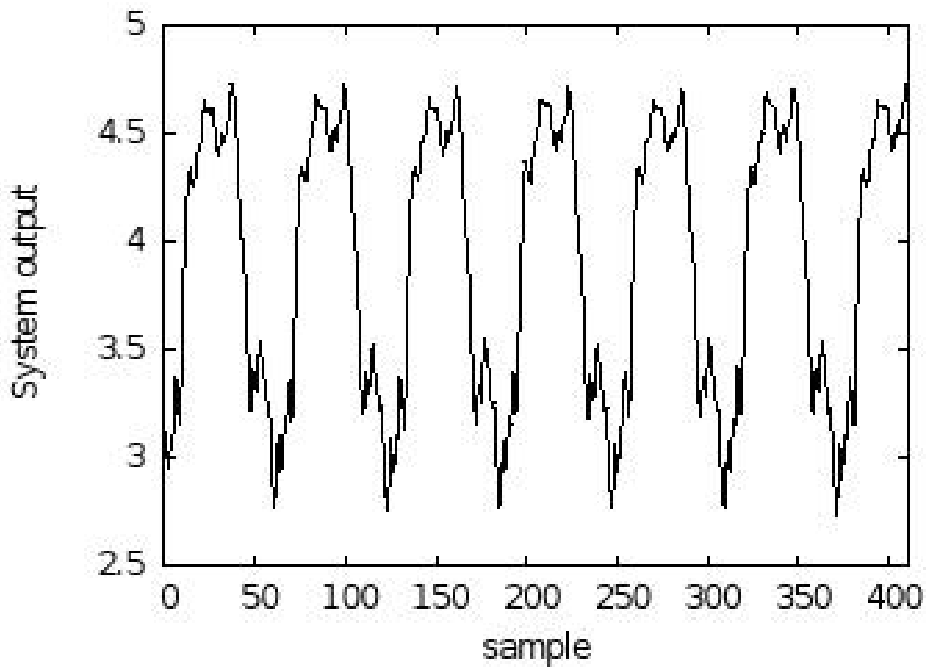



# Engine Data

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

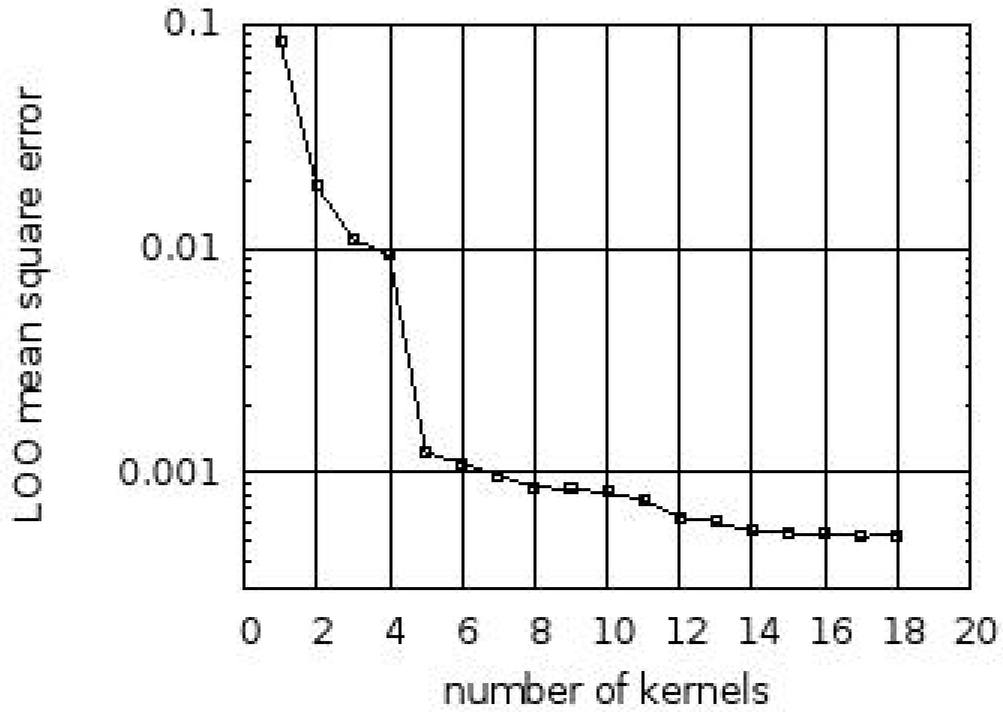
Data set contains 410 pairs of input-output samples, modelled as $y_k = f(\mathbf{x}_k) + e_k$ with $\mathbf{x}_k = [y_{k-1}\ u_{k-1}\ u_{k-2}]^T$, first 210 data points for training and last 200 points for testing



# Engine Data Modelling

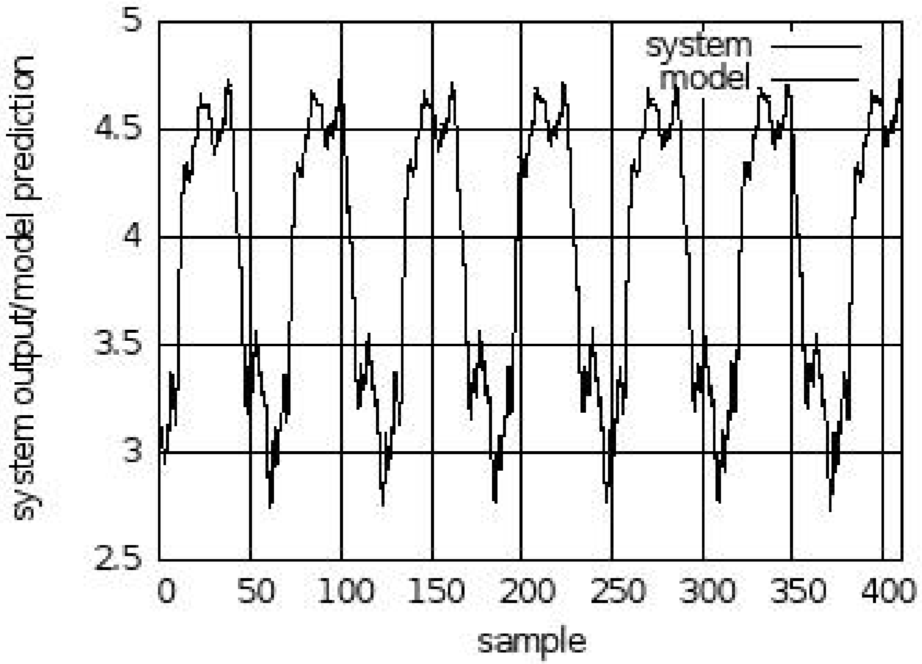
- The OFS-LOO using Gaussian kernels
  - The LOO mean square error as a function of model size for the engine data set
  - The OFS-LOO constructed 17 kernels
  - The LROLS-LOO reduced the model to 15 kernels

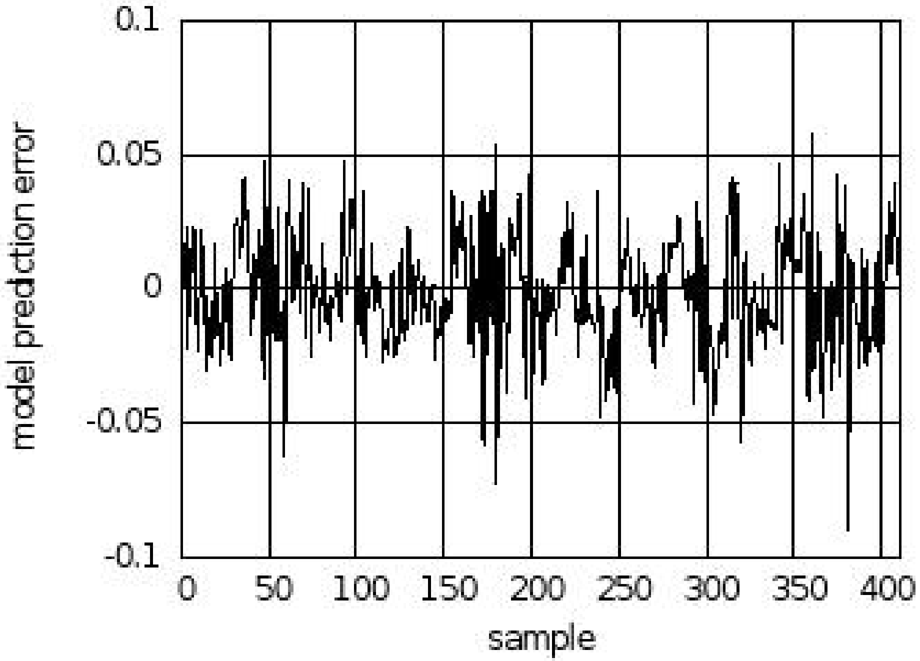



- The SVM and LROLS-LOO were also used for comparison

# Engine Data Results

| algorithm | kernel type | model size | training MSE | test MSE |
|---|---|---|---|---|
| SVM | fixed Gaussian | 92 | 0.000447 | 0.000498 |
| LROLS-LOO | fixed Gaussian | 22 | 0.000453 | 0.000490 |
| **OFS-LOO** | **tunable Gaussian** | **15** | **0.000466** | **0.000480** |

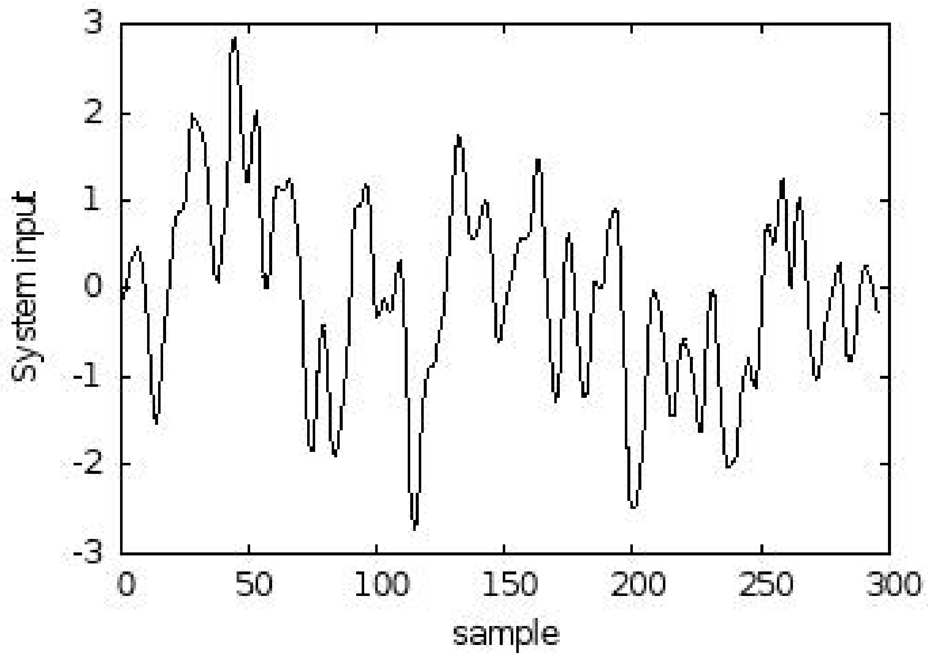


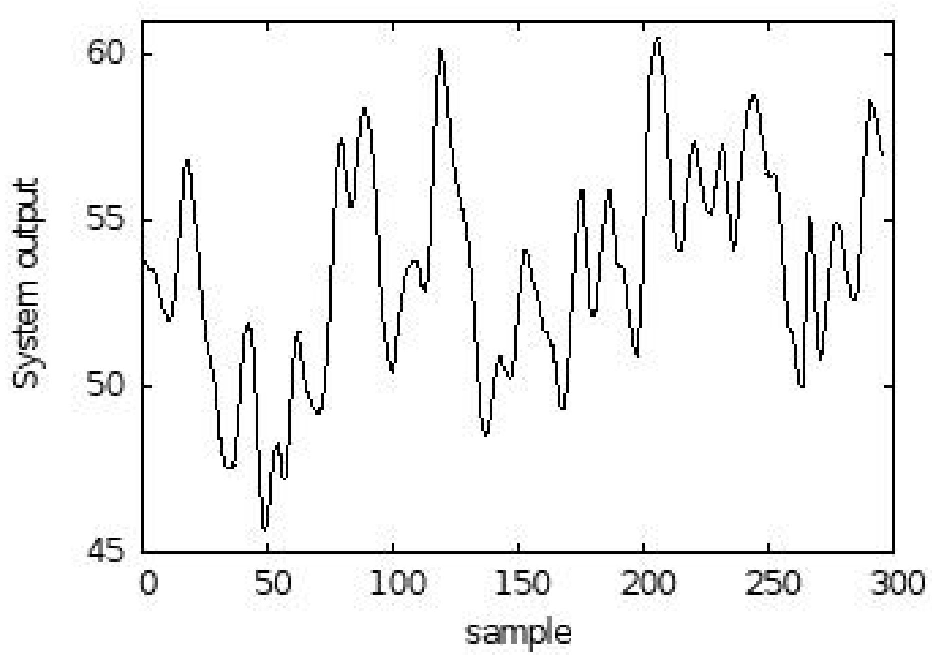



# Gas Furnace Data

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

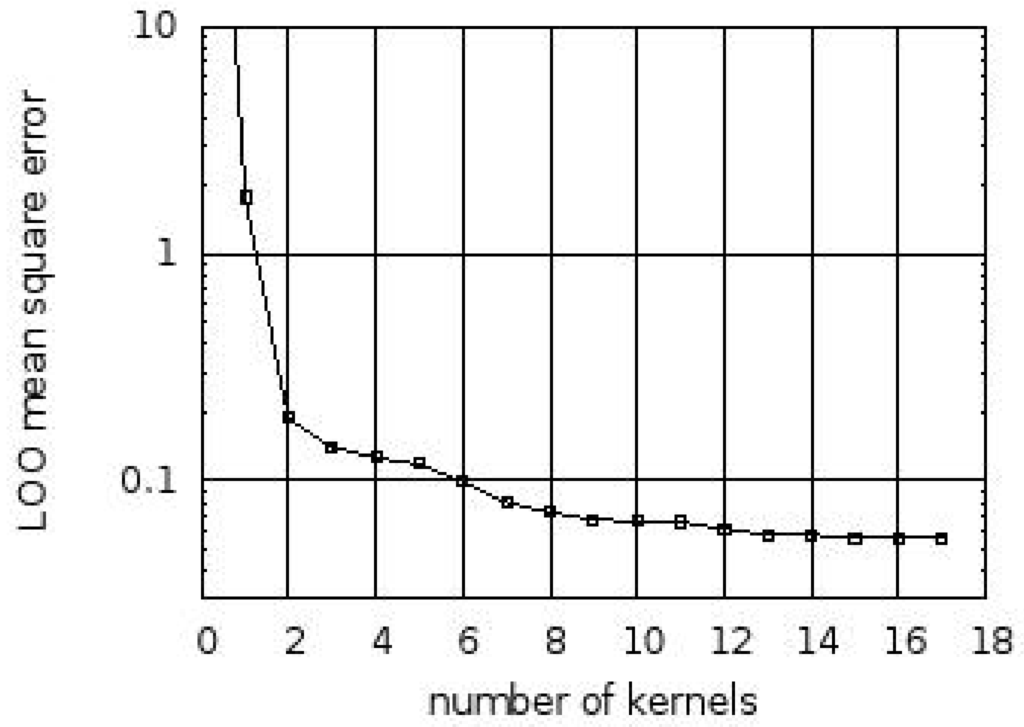
Data set contains 296 pairs of input-output samples, modelled as $y_k = f(\mathbf{x}_k) + e_k$ with $\mathbf{x}_k = [y_{k-1}\ y_{k-2}\ y_{k-3}\ u_{k-1}\ u_{k-2}\ u_{k-3}]^T$, all the data points for training



# Gas Furnace Modelling

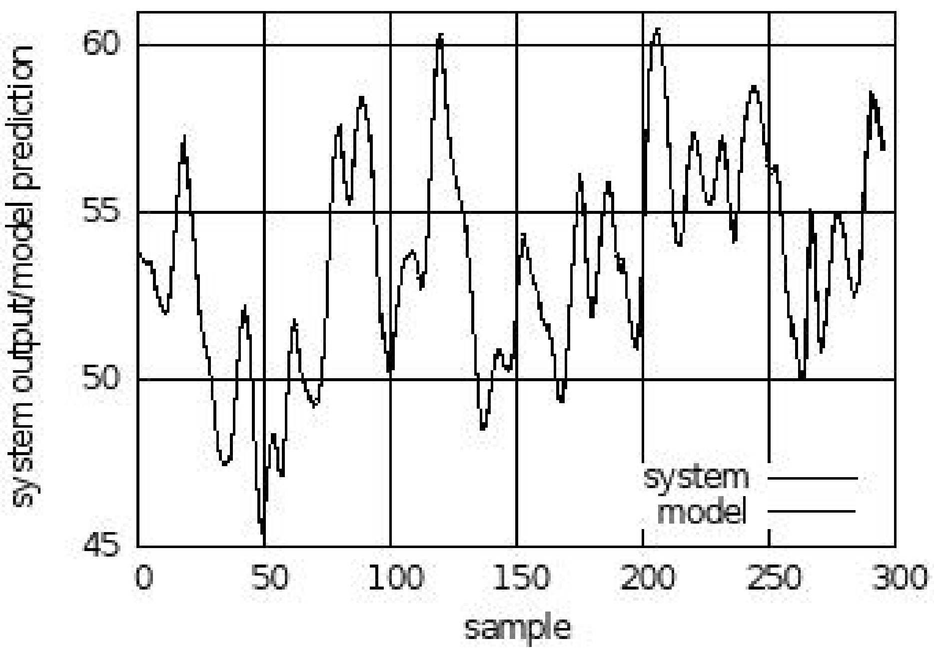
- The OFS-LOO using Gaussian kernels
  - The LOO mean square error as a function of model size for the gas furnace data set
  - The OFS-LOO constructed 16 kernels
  - The LROLS-LOO reduced the model to 15 kernels

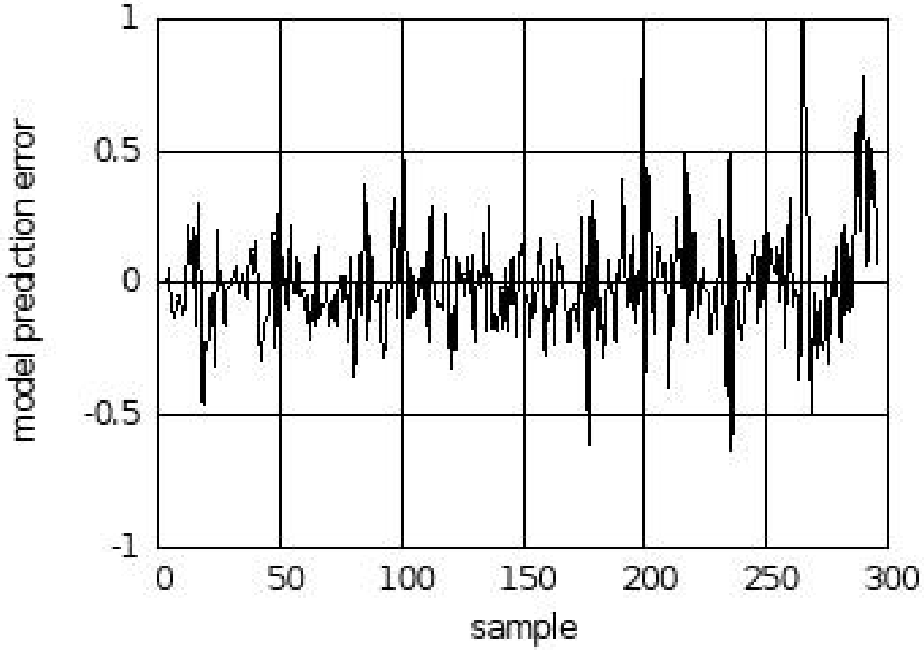



- The SVM and LROLS-LOO were also used for comparison

# Gas Furnace Results

| algorithm | kernel type | model size | training MSE | LOO MSE |
|---|---|---|---|---|
| SVM | fixed Gaussian | 62 | 0.052416 | 0.054376 |
| LROLS-LOO | fixed thin-plate-spline | 28 | 0.053306 | 0.053685 |
| **OFS-LOO** | **tunable Gaussian** | **15** | **0.054306** | **0.054306** |



# Boston Housing Data

- Boston Housing: 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

| algorithm | kernel type | model size | training MSE | test MSE |
|---|---|---|---|---|
| SVM | fixed | $243.2 \pm 5.3$ | $6.7986 \pm 0.4444$ | $23.1750 \pm 9.0459$ |
| LROLS-LOO | fixed | $58.6 \pm 11.3$ | $12.9690 \pm 2.6628$ | $17.4157 \pm 4.6670$ |
| **OFS-LOO** | **tunable** | $\mathbf{34.6 \pm 8.4}$ | $\mathbf{10.0997 \pm 3.4047}$ | $\mathbf{14.0745 \pm 3.6178}$ |

# 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.**