Sparse Model Identification Using Orthogonal Forward Regression with Basis Pursuit and D-optimality

X. Hong[†], M. Brown[‡], S. Chen[§], C. J. Harris[§],

[†]Department of Cybernetics
University of Reading, Reading, RG6 6AY, UK
Email: x.hong@reading.ac.uk; Tel: +44 (0)118 9318222; Fax:+44(0)118 9318220

[‡] Department of Computing and Mathematics
Manchester Metropolitan University, Manchester, UK
Email: m.brown@mmu.ac.uk

[§] Department of Electronics and Computer Science
University of Southampton, Southampton SO17 1BJ, UK
Email: sqc@ecs.soton.ac.uk; cjh@ecs.soton.ac.uk.

Abstract — An efficient model identification algorithm for a large class of linear-in-the-parameters models is introduced that simultaneously optimizes the model approximation ability, sparsity and robustness. The derived model parameters, in each forward regression step, are initially estimated via the orthogonal least squares (OLS), followed by being tuned with a new gradient descent learning algorithm based on the basis pursuit that minimizes the l^1 norm of the parameter estimate vector. The model subset selection cost function includes a D-optimality design criterion that maximizes the determinant of the design matrix of the subset to ensure the model robustness and to enable that the model selection procedure automatically terminates at a sparse model. The proposed approach is based on the forward OLS algorithm using the modified Gram-Schmidt procedure. Both the parameter tuning procedure, based on basis pursuit, and the model selection criterion, based on the D-optimality that is effective in ensuring model robustness, are integrated with the forward regression. As a consequence, the inherent computational efficiency associated with the conventional forward OLS approach is maintained in the proposed algorithm. Illustrative examples are included to demonstrate the effectiveness of the new approach.

1 Introduction

Associative memory networks (such as B-spline networks, radial basis function (RBF) networks and support vector machines (SVM)) have been extensively studied [1, 2, 3, 4]. A main obstacle in nonlinear modelling using associative memory networks or fuzzy logic has been the problem of *the curse of dimensionality* [5]. This factor applies to all lattice based networks or knowledge representations such as fuzzy logic (FL), RBF, Karneva distributed memory maps, and all neurofuzzy networks (e.g. adaptive network based fuzzy inference system (ANFIS) [6], Takagi and Sugeno model [7], etc.). For these systems it is essential to use some model construction procedures to overcome the obstacle by deriving a model with an appropriate dimension. For general linear in the parameter systems, an orthogonal least squares (OLS) algorithm based on Gram-Schmidt orthogonal decomposition can be used to determine the significant model elements and associated parameter estimates, and the overall model structure [8]. Regularization techniques have been incorporated into the OLS algorithm to produce a regularized orthogonal least squares (ROLS) algorithm that reduces the variance of parameter estimates [9, 10]. To produce a model with good generalization capabilities, model selection criteria such as the Akaike information criterion (AIC) [11] are usually incorporated into the procedure to determinate the model construction process. Due to the fact that AIC or other information based criteria are usually simplified measures derived as an approximation formula that is particularly sensitive to model complexity. The use of AIC or other information based criteria, if used in forward regression, only affects the stopping point of the model selection, but does not penalize regressors that might cause poor model performance, e.g. too large parameter variance or ill-posedness of the regression matrix, if this is selected.

In optimum experimental design [12], it is common that the models are also in the form of linear-in-the-parameters. For these models, the design criteria are defined as function of the eigenvalues of the design matrix, hence quantitatively measure the model adequacy. In recent studies [13, 14], the authors have outlined efficient learning algorithms, in which composite cost functions were introduced to optimize the model approximation ability by using the forward OLS algorithm [8], and simultaneously the model adequacy by using an A-optimality design criterion (i.e. minimizes the variance of the parameter estimates), or a D-optimality criterion (i.e. optimizes the parameter efficiency and model robustness via the maximization of the determinant of the design matrix). It was shown that the resultant models can be improved based on A- or D-optimality. These algorithms lead automatically to an unbiased model parameter estimate with an overall robust and parsimonious model structure. Combining a locally regularized orthogonal least squares (LROLS) model selection [15] with D-optimality experimental design further enhances model robustness [16]. It has been shown [16, 17] that the parameter regularization is equivalent to a maximized *a posterior pdf* (MAP) of parameters from Bayesian viewpoint by adopting a Gaussian prior for parameters.

The regularization [9, 10] uses a penalty function on l^2 norms of the parameters. Alternatively the model sparsity can be achieved by a novel concept of the basis pursuit or least angle regression [18, 19] that aims to obtain a model by minimizing the l^1 norm of the parameters. The Bayesian interpretation for basis pursuit method is simply by adopting an exponential prior for parameters (see Section 2.1). The advantage of the basis pursuit is that it can achieve much sparser models by forcing more parameters to zero, than models derived from the minimization of the l^p norm, as most l^p norms will produces parameters small, but nonzero, values. Compared to method of the regularization [9, 10], the basis pursuit method, however, will not generally be computationally efficient, because by simply changing from l^2 norm to l^1 norm in the cost function, this effectively changes a quadratic optimization problem with a simple solution into a more sophisticated problem for which a convex, nonquadratic optimization is generally required[18, 19].

In this paper, a new model identification technique is introduced by using forward regression with basis pursuit and D-optimality design. Based on the previous work [13], we incorporate the concept of basis pursuit to tune the parameter estimates as derived from the orthogonal least squares method. A gradient descent parameter learning method is initially introduced with proven convergence, followed by its application to the parameters tuning in the modified Gram-Schmidt algorithm. It is shown that parameter tuning by basis pursuit, following the initialization of least squares inherent in the Gram-Schmidt procedure will enforce model sparsity, yet fit well in the procedure automated by the D-optimality model selective criterion. In the proposed algorithm, the gradient descent of the basis pursuit contributes as a tuning procedure, rather than the main optimization method, so the computational efficiency of the method due to the forward OLS regression maintains.

This paper is organized as follows. Section 2 introduces the current work on forward regression based on the modified Gram-Schmidt algorithm as a modelling approach. Section 3 initially introduces a new gradient descent method based on basis pursuit cost function, followed by the proposed algorithm itself that incorporates the basis pursuit optimization with the modified Gram-Schmidt algorithm. Numerical examples are provided in Section 4 to illustrate the effectiveness of the approach and Section 5 is devoted to conclusions.

2 Preliminaries

A linear regression model (RBF neural network, B-spline neurofuzzy network) can be formulated as [1, 2]

$$y(t) = \sum_{k=1}^{M} p_k(\mathbf{x}(t))\theta_k + \xi(t) \tag{1}$$

where $t = 1, 2, \cdots, N$, and N is the size of the estimation data set. $y(t)$ is system output variable, $\mathbf{x}(t) = [y(t-1), \cdots, y(t-n_y), u(t-1), \cdots, u(t-n_u)]^T$ is system input vector with assumed known dimension of $(n_y + n_u)$. $u(t)$ is system input variable. $p_k(\bullet)$ is a known nonlinear basis function, such as RBF, or B-spline fuzzy membership functions. $\xi(t)$ is an uncorrelated model residual sequence with zero mean and variance of σ^2. Eq.(1) can be written in the matrix form as

$$\mathbf{y} = \mathbf{P}\Theta + \Xi \tag{2}$$

where $\mathbf{y} = [y(1), \cdots, y(N)]^T$ is the output vector. $\Theta = [\theta_1, \cdots, \theta_M]^T$ is parameter vector, $\Xi = [\xi(1), \cdots, \xi(N)]^T$ is the residual vector, and $\mathbf{P}$ is the regression matrix

$$\mathbf{P} = \begin{bmatrix} p_1(1) & p_2(1) & \cdots & p_M(1) \\ p_1(2) & p_2(2) & \cdots & p_M(2) \\ \cdots & \cdots & \cdots & \cdots \\ p_1(N) & p_2(N) & \cdots & p_M(N) \end{bmatrix}$$

with $p_k(t) = p_k(\mathbf{x}(t))$. Denote the column vectors in $\mathbf{P}$ as $\mathbf{p}_k = [p_k(1), \cdots, p_k(N)]^T$, $k = 1, \cdots, M$. An orthogonal decomposition of $\mathbf{P}$ is

$$\mathbf{P} = \mathbf{W}\mathbf{A} \tag{3}$$

where $\mathbf{A} = \{\alpha_{ij}\}$ is an $M \times M$ unit upper triangular matrix and $\mathbf{W}$ is an $N \times M$ matrix with orthogonal columns that satisfy

$$\mathbf{W}^T\mathbf{W} = diag\{\kappa_1, \cdots, \kappa_M\} \tag{4}$$

with

$$\kappa_k = \mathbf{w}_k^T\mathbf{w}_k, \qquad k = 1, \cdots, M \tag{5}$$

so that (2) can be expressed as

$$\mathbf{y} = (\mathbf{P}\mathbf{A}^{-1})(\mathbf{A}\Theta) + \Xi = \mathbf{W}\Gamma + \Xi \tag{6}$$

where $\Gamma = [\gamma_1, \cdots, \gamma_M]^T$ is an auxiliary vector.

2.1 The modified Gram-Schmidt algorithm, parameter regularization and basis pursuit

Clearly for the orthogonalised system (6), the least squares estimates is given by

$$\gamma_k^{(0)} = \frac{\mathbf{w}_k^T \mathbf{y}}{\mathbf{w}_k^T \mathbf{w}_k}, \qquad k = 1, \cdots, M \tag{7}$$

The original model coefficient vector $\Theta = [\theta_1, \cdots, \theta_M]^T$ can then be calculated from $\mathbf{A}\Theta = \Gamma$ through back substitution.

The modified Gram-Schmidt procedure, described below, can be used to perform the orthogonalization of (3) and parameter estimation (7). Starting from $k = 1$, the columns $\mathbf{p}_j$, $k+1 \leq j \leq M$ are made orthogonal to the kth column at the kth stage. The operation is repeated for $1 \leq k \leq M-1$. Specifically, denoting $\mathbf{p}_j^{(0)} = \mathbf{p}_j$, $1 \leq j \leq M$, then for $k = 1, \cdots, M-1$

$$\begin{aligned} \mathbf{w}_k &= \mathbf{p}_k^{(k-1)} \\ \alpha_{kj} &= \frac{\mathbf{w}_k^T \mathbf{p}_j^{(k-1)}}{\mathbf{w}_k^T \mathbf{w}_k}, \qquad k+1 \leq j \leq M \\ \mathbf{p}_j^{(k)} &= \mathbf{p}_j^{(k-1)} - \alpha_{kj}\mathbf{w}_k, \qquad k+1 \leq j \leq M \end{aligned} \tag{8}$$

where α_{kj}'s are components of the upper triangular matrix $\mathbf{A}$. The last stage of the procedure is simply $\mathbf{w}_M = \mathbf{p}_M^{(M-1)}$. The elements of the auxiliary vector Γ are computed by transforming $\mathbf{y}^{(0)} = \mathbf{y}$ in a similar way. For $1 \leq k \leq M$

$$\begin{aligned} \gamma_k^{(0)} &= \frac{\mathbf{w}_k^T \mathbf{y}^{(k-1)}}{\mathbf{w}_k^T \mathbf{w}_k} \\ \mathbf{y}^{(k)} &= \mathbf{y}^{(k-1)} - \gamma_k^{(0)} \mathbf{w}_k \end{aligned} \tag{9}$$

It can be easily verified that $\gamma_k^{(0)}$ as derived from (9) is equivalent to (7). Geometrically the system output vector $\mathbf{y}$, at step k, is projected onto a set of orthogonal basis vectors, $\{\mathbf{w}_1, ... \mathbf{w}_k\}$. The model residual is decreased by projecting the system output vector $\mathbf{y}$ onto a new basis $\mathbf{w}_k$ at this step. Effectively, (9) can be regarded as a linear fitting of $\mathbf{y}^{(k-1)}$ by using a single variable $\mathbf{w}^{(k)}$, and to derive the new model residual $\mathbf{y}^{(k)}$, and so on. This observation will be explored further in Section 3.1 for the development of the proposed algorithm in Section 3.2.

For better model parameter estimation bias/variance tradeoff, the regularization can be applied. If the regularization is performed to the parameters in orthogonal space, γ_k, then (9) is simply replaced by the following

$$\begin{aligned} \gamma_k^{(r)} &= \frac{\mathbf{w}_k^T \mathbf{y}}{\mathbf{w}_k^T \mathbf{w}_k + \lambda_k}, \qquad k = 1, \cdots, M \\ \mathbf{y}^{(k)} &= \mathbf{y}^{(k-1)} - \gamma_k^{(r)} \mathbf{w}_k \end{aligned} \tag{10}$$

where $\lambda_k \geq 0$, are regularization parameters, which can be optimized by being treated as hyperparameters in Bayesian approach [16]. The above results are obtained by setting parameter optimizer as

$$V^{(r)} = \frac{1}{2}E[\xi^2(t)] + \sum_{k=1}^{M} \lambda_k \gamma_k^2$$

Because the regularization term is given as the l^2 norm, the closed form parameter estimates solution given by (10) is available as solution to a quadratic form optimization.

Alternatively the basis pursuit method is simply given by changing the l^2 norm into l^1 such that

$$V = \frac{1}{2}E[\xi^2(t)] + \boldsymbol{\lambda}^T ||\Gamma||_1 \tag{11}$$

where $\boldsymbol{\lambda} = [\lambda_1, ..., \lambda_{n_\theta}]^T$, $||\Gamma||_1 = [|\gamma_1|, ..., |\gamma_{n_\theta}|]^T$, and $n_\theta \leq M$ denotes the size of parameter vector of Γ with nonzero parameters. $\lambda_k \geq 0$, are basis pursuit parameters. Note that only nonzero parameters that are actually included in the model are penalized, because a regressor with zero parameter does not influence model performance.

The basis pursuit method tends to produce model with greater sparsity than that of l^2 parameter regularization. Because the solution of (11) is a nonquadratic optimization problem, there is no readily available closed form solution as simple as (10). In general, the basis pursuit will not be computationally efficient, since this is a more sophisticated problem for which a convex, nonquadratic optimization is required [18]. The objective of this paper is to tackle this problem by introducing some simple model identification algorithm using the idea of basis pursuit, as introduced in Section 3.

Bayesian regularization and basis pursuit

The regularized parameter estimator by optimizing $V^{(r)}$ is equivalent to a maximized *a posterior pdf* (MAP) of parameters in a Bayesian approach [17, 16]. By Bayesian Theorem

$$p(\Gamma|D_N) \propto p(\Gamma)p(D_N,\Gamma) \tag{12}$$

It can be assumed that $\xi \sim N(0,\sigma^2)$, and observations are independent, so

$$p(D_N,\Gamma) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp[-\frac{1}{2\sigma^2}\sum_{t=1}^{N} \xi^2(t)] \tag{13}$$

whose maximization leads to maximum likelihood (ML) parameter estimator, which is equivalent to least squares estimator for linear-in-the-parameters models. The prior $p(\Gamma)$ serves as a solution to the inadequacy of ML estimator by using prior knowledge of $p(\Gamma)$ that controls superfluous parameters for improved generalization. If the prior $p(\Gamma)$ for the parameters is Gaussian

$$p(\Gamma) = \exp(-\frac{1}{\sigma^2}\sum_{k=1}^{M} \lambda_k \gamma_k^2)/Z_\Gamma^{(r)} \tag{14}$$

where $Z_\Gamma^{(r)}$ is a normalizing coefficient. The MAP estimator can be derived via minimizing $V^{(r)}$ [1, 17, 16]. Clearly for basis pursuit estimator, the prior $p(\Gamma)$ is simply set as

$$p(\Gamma) = \exp(-\frac{1}{\sigma^2}\boldsymbol{\lambda}^T ||\Gamma||_1)/Z_\Gamma \tag{15}$$

where Z_Γ is a normalizing coefficient. This means that, from Bayesian viewpoint, the basis pursuit method can be regarded as adopting a multivariable exponential distribution as a prior for parameters.

2.2 Model structure selection by D-optimality

A significant advantage due to orthogonalisation is that the contribution of model regressors to the model can be evaluated. The forward OLS estimator involves selecting a set of n_θ variables $\mathbf{p}_k = [p_k(1), \cdots, p_k(N)]^T$, $k = 1, \cdots, n_\theta$, from M regressors to form a set of orthogonal basis $\mathbf{w}_k$, $k = 1, \cdots, n_\theta$, in a forward regression manner. As the orthogonality property $\mathbf{w}_i^T \mathbf{w}_j = 0$ for $i \neq j$ holds, if (6) is multiplied by itself and then the time average is taken, the following equation is easily derived

$$\frac{1}{N}\mathbf{y}^T\mathbf{y} = \frac{1}{N}\sum_{k=1}^{M}\gamma_k^2 \mathbf{w}_k^T \mathbf{w}_k + \frac{1}{N}\Xi^T\Xi \tag{16}$$

The Error Reduction Ratio $[ERR]_k$, which is defined as the increment towards the overall output variance $E[y^2(t)]$ due to each regressor or input variable $p_k(t)$ divided by the overall output variance is computed through [8]

$$[ERR]_k = \frac{\gamma_k^2 \mathbf{w}_k^T \mathbf{w}_k}{\mathbf{y}^T\mathbf{y}}, \qquad k = 1, \cdots, M \tag{17}$$

The most relevant n_θ regressors can be forward selected according to the value of the error reduction ratio $[ERR]_k$. At the kth selection, a candidate regressor is selected as the kth basis of the subset if it produces the largest value of $[ERR]_k$ from the remaining $(M-k+1)$ candidates. By setting an appropriate tolerance ρ, which can be found by trial and error or via some statistical information criterion such as Akaike's information criterion(AIC) [11] that forms a compromise between the model performance and model complexity, the variable selection is terminated when

$$1 - \sum_{k=1}^{n_\theta}[ERR]_k < \rho \tag{18}$$

This procedure can automatically select a subset of n_θ regressors to construct a parsimonious model. Equivalently, this procedure can be expressed as

$$J^{(k)} = J^{(k-1)} - \frac{1}{N}\gamma_k^2 \kappa_k \tag{19}$$

where $J^{(0)} = \mathbf{y}^T\mathbf{y}$. At the kth forward regression stage, a candidate regressor is selected as the kth regressor if it produces the smallest $J^{(k)}$. (19) can be modified to form an alternative model selective criterion to enhance model robustness. D-optimality based cost function is one of robustness design criterion in experimental design criteria [12]. The D-optimality criterion is to maximize the determinant of the design matrix defined as $\mathbf{W}_k^T\mathbf{W}_k$, where $\mathbf{W}_k \in \Re^{N\times n_\theta}$ denotes the resultant regression matrix, consisting of n_θ regressors selected from M regressors in $\mathbf{W}$.

$$\max\{J_D = \det(\mathbf{W}_k^T\mathbf{W}_k) = \prod_{k=1}^{n_\theta}\kappa_k\} \tag{20}$$

It can be easily verified that the selection of the a subset of $\mathbf{W}_k$ from $\mathbf{W}$ is equivalent to the selection of the a subset of n_θ regressors from $\mathbf{P}$ [14]. In order to include D-optimality as a model selective criterion for improved model robustness, construct an augmented cost function as

$$\begin{aligned} J &= \frac{1}{N}\Xi^T\Xi + \alpha \log(\frac{1}{J_D}) \\ &= \frac{1}{N}(\mathbf{y}^T\mathbf{y} - \sum_{k=1}^{n_\theta} \gamma_k^2\kappa_k) + \alpha \sum_{k=1}^{n_\theta} \log[\frac{1}{\kappa_k}] \end{aligned} \tag{21}$$

where α is a positive small number. Note that this composite cost function simultaneously minimizes (19) and maximizes (20) [14]. Eq.(21) can be directly incorporated into the forward OLS algorithm to select the most relevant kth regressor at the kth forward regression stage, via

$$J^{(k)} = J^{(k-1)} - \frac{1}{N}\gamma_k^2\kappa_k + \alpha \log[\frac{1}{\kappa_k}] \tag{22}$$

At the kth forward regression stage, a candidate regressor is selected as the kth regressor if it produces the smallest $J^{(k)}$ and further reduction in $J^{(k-1)}$. Because $\log(\frac{1}{J_D})$ is an increasing function if $\kappa_k < 1$, which is true for some $k > K$, the selection procedure will terminate if $J^{(k)} \geq J^{(k-1)}$ at the derived model size n_θ if an proper α is set. This is significant because this means that the proposed approach can detect a parsimonious model size in an automatic manner. The D-optimality based model selective criterion will be applied in the proposed new model identification algorithm introduced in next section.

3 Model identification algorithm using Forward Regression with Basis Pursuit and D-optimality

3.1 Parameter estimation by basis pursuit function's gradient descent

Before the introduction of the proposed algorithm, we initially introduce a general concept (algorithm) of parameter estimation by basis pursuit function's gradient descent, followed by the basic idea as how to incorporate this algorithm in the modified Gram-Schmidt orthogonal procedure.

Theorem 1: Suppose that the dynamics underlying data set D_N can be described by

$$y(t) = f(\mathbf{x}(t), \Theta) + \xi(t) \tag{23}$$

where functional $f(\bullet)$ is given as appropriate. If the following parameter learning law is applied

$$\Theta(t+1) = \Theta(t) + \overline{\eta\xi(t)\frac{\partial f}{\partial \Theta}} - \eta\ \boldsymbol{\lambda}^T\ \mathrm{sgn}(\Theta(t)) \tag{24}$$

where the operator $\overline{(\bullet)}$ denotes the time averaging, and $\mathrm{sgn}(\Theta) = [\mathrm{sgn}(\theta_1), ..., \mathrm{sgn}(\theta_M)]^T$, in which,

$$\mathrm{sgn}(u) = \begin{cases} 1 & \text{if} \quad u > 0 \\ 0 & \text{if} \quad u = 0 \\ -1 & \text{if} \quad u < 0 \end{cases} \tag{25}$$

η is an arbitrarily small positive number, then

$$\begin{aligned} &\text{(i)} \quad \lim_{t\to+\infty} V(t) \to c \\ &\text{(ii)} \quad \lim_{t\to+\infty} ||\Theta(t) - \Theta(t-k)|| = 0 \text{ for any finite } k \end{aligned} \tag{26}$$

where the basis pursuit cost function $V(t) = \frac{1}{2}\overline{\xi^2(t)} + \boldsymbol{\lambda}^T||\Theta||_1$, and $||\Theta||_1 = [|\theta_1|, ..., |\theta_{n_\theta}|]^T$ is constructed based on a subvector of Θ with nonzero parameters (see also (11)). $c = \min V(t)$ is the lower bound of $V(t)$.

Proof: Consider $V(t) = \frac{1}{2}\overline{\xi^2(t)} + \boldsymbol{\lambda}^T||\Theta||_1 > 0$ as a Lyapunov function. For an arbitrarily small neighborhood around a current parameter estimate $\Theta(t) = [\theta_1(t), ...\theta_M(t)]^T$, by the first order of Taylor series expansion of $V(t)$

$$\begin{aligned} \triangle V(t) &\approx [\frac{\partial V(t)}{\partial \Theta}]^T \triangle \Theta(t) \\ &= \{-\overline{\xi(t)\frac{\partial f}{\partial \Theta}} + \boldsymbol{\lambda}^T \text{ sgn}(\Theta(t))\} \triangle \Theta(t) \end{aligned} \tag{27}$$

where $\triangle\Theta(t) = \Theta(t+1) - \Theta(t)$, $\triangle V(t+1) = V(t+1) - V(t)$. When the learning law of (24) is applied, we have

$$\triangle V(t) = -\eta\{\overline{\xi(t)\frac{\partial f}{\partial \Theta}} - \boldsymbol{\lambda}^T \text{ sgn}(\Theta(t))\}^T\{\overline{\xi(t)\frac{\partial f}{\partial \Theta}} - \boldsymbol{\lambda}^T \text{ sgn}(\Theta(t))\} \leq 0 \tag{28}$$

that is, $V(t)$ is non-increasing with a lower bound. Hence

$$\lim_{t\to+\infty} \triangle V(t) = 0 \tag{29}$$

Hence property (i) is established.

$$\begin{aligned} \lim_{t\to+\infty} \triangle V(t) &= \eta \triangle \Theta^T(t) \triangle \Theta(t) \\ &= \eta||\Theta(t) - \Theta(t-1)||^2 \end{aligned} \tag{30}$$

yielding

$$\lim_{t\to+\infty} ||\Theta(t) - \Theta(t-1)|| = 0 \tag{31}$$

for a finite k

$$\begin{aligned} ||\Theta(t) - \Theta(t-k)||^2 &= ||\sum_{i=1}^{k} \Theta(t-i+1) - \Theta(t-i)||^2 \\ &= \sum_{i=1}^{k} ||\Theta(t-i+1) - \Theta(t-i)||^2 \to 0 \end{aligned} \tag{32}$$

so property (ii) follows;

□

In the proposed algorithm of Subsection 3.2, the above gradient descent of basis pursuit error function is combined with the modified Gram-Schmidt algorithm of Section 2.1 to derive a new model identification procedure. The basic idea is introduced here. Consider (9), which can be regarded as a linear fitting of $\mathbf{y}^{(k-1)}$ by using a single variable $\mathbf{w}^{(k)}$ with the least squares method. The derived model residual vector Ξ is then set as $\mathbf{y}^{(k)}$. This observation suggests that for each step k in the modified Gram-Schmidt algorithm, the parameter estimates, calculated by (9) can be further tuned by learning algorithm of (24) that optimizes the basis pursuit's function given by (11). Following (9), denote $\mathbf{y}^{(k-1)} = [y^{(k-1)}(1), y^{(k-1)}(2), ..., y^{(k-1)}(N)]^T$ and $\mathbf{w}_k = [w_k(1), ..., w_k(N)]^T$. The tuning process is an extremely simple case based on Theorem 1, as illustrated by the following

Theorem.

Theorem 2: If the learning law given by (24) is applied to a special case of one dimensional linear system

$$y^{(k-1)}(t) = \gamma_k w_k(t) + \xi(t) \tag{33}$$

with the parameter estimates γ_k initialized as the least square parameter estimate $\gamma_k^{(0)} \neq 0$, given by (9), and if $\lambda_k < \frac{1}{2N}|\mathbf{w}_k^T\mathbf{y}|$, then the final converged parameter estimate γ_k

$$\begin{aligned} &\text{(i)} \quad |\gamma_k| < |\gamma_k^{(0)}| \\ &\text{(ii)} \quad \mathrm{sgn}(\gamma_k) = \mathrm{sgn}(\gamma_k^{(0)}) \end{aligned} \tag{34}$$

Proof: (i) The learning law given by (24), when applied to the system (33), can be rewritten as

$$\gamma_k(t+1) = \gamma_k(t) + \eta\overline{\xi(t)w_k(t)} - \eta\lambda_k\ \mathrm{sgn}(\gamma_k(t)) \tag{35}$$

The least squares solution means that $\overline{\frac{1}{2}\xi^2(t,\gamma_k)} \geq \overline{\frac{1}{2}\xi^2(t,\gamma_k^{(0)})}$, and $V(t) = \overline{\frac{1}{2}\xi^2(t)} + \lambda_k|\gamma_k|$ is non-increasing, with an initial value as $\overline{\frac{1}{2}\xi^2(t)} + \lambda_k|\gamma_k^{(0)}|$, so for $t \to \infty$,

$$V(t) = \overline{\frac{1}{2}\xi^2(t,\gamma_k)} + \lambda_k|\gamma_k| \leq \overline{\frac{1}{2}\xi^2(t,\gamma_k^{(0)})} + \lambda_k|\gamma_k^{(0)}| \tag{36}$$

yields $|\gamma_k| < |\gamma_k^{(0)}|$. Hence (i) follows.

(ii) For an arbitrary small learning rate η, it can be assumed that the parameter changes in arbitrary small range per time step. Initially it is assumed that γ_k change sign at a time step denoted as t', i.e., the parameter trajectory needs to pass zero at a point t', $\gamma_k(t') = \varepsilon$, where $\varepsilon \approx 0$, and by the property that $V(t)$ is non-increasing, yields

$$\begin{aligned} V(t') &= \overline{\frac{1}{2}\xi^2(t,\varepsilon)} + \lambda_k|\varepsilon| = \overline{\frac{1}{2}[y^{(k-1)}(t) - \varepsilon w_k(t)]^2} + \lambda_k|\varepsilon| \approx \frac{1}{2N}[\mathbf{y}^{(k-1)}]^T\mathbf{y}^{(k-1)} \\ &\leq \overline{\frac{1}{2}\xi^2(t,\gamma_k^{(0)})} + \lambda_k|\gamma_k^{(0)}| = \frac{1}{2N}[\mathbf{y}^{(k-1)}]^T\mathbf{y}^{(k-1)} - \frac{1}{2N}[\gamma_k^{(0)}]^2\mathbf{w}_k^T\mathbf{w}_k + \lambda_k|\gamma_k^{(0)}| \end{aligned} \tag{37}$$

So

$$\frac{1}{2N}[\mathbf{y}^{(k-1)}]^T\mathbf{y}^{(k-1)} \leq \frac{1}{2N}[\mathbf{y}^{(k-1)}]^T\mathbf{y}^{(k-1)} - \frac{1}{2N}[\gamma_k^{(0)}]^2\mathbf{w}_k^T\mathbf{w}_k + \lambda_k|\gamma_k^{(0)}| \tag{38}$$

$$\lambda_k|\gamma_k^{(0)}| \geq \frac{1}{2N}[\gamma_k^{(0)}]^2\mathbf{w}_k^T\mathbf{w}_k \tag{39}$$

and by applying the least square solution $\gamma_k^{(0)} = \frac{\mathbf{w}_k^T\mathbf{y}^{(k-1)}}{\mathbf{w}_k^T\mathbf{w}_k}$, yields

$$\lambda_k \geq \frac{1}{2N}|\mathbf{w}_k^T\mathbf{y}^{(k-1)}| = \frac{1}{2N}|\mathbf{w}_k^T\mathbf{y}| \tag{40}$$

This is contradictory to the assumption for λ_k. Therefore γ_k should not change sign throughout conditional on $\lambda_k < \frac{1}{2N}|\mathbf{w}_k^T\mathbf{y}|$, hence property (ii) follows.

□

The significance of Theorem 2 is that by setting the basis pursuit parameters λ_k below a certain value, for each step k, the overall effect of the tuning process is that the parameters γ_k

is pulled towards 0. In forward regression, as model size k increases, the parameter estimates γ_k, as initialized by least squares algorithm with very small magnitudes, followed by basis pursuit gradient tuning, will shrink below some threshold value, and can therefore be obtained as zero, to achieve model sparsity. For a sufficiently small λ_k, the optimality condition can be derived as

$$\overline{\xi(t)w_k(t)} - \lambda_k \ \mathrm{sgn}(\gamma_k(t)) = 0 \tag{41}$$

or

$$\begin{aligned} \gamma_k &= \frac{\mathbf{w}_k^T \mathbf{y}^{(k-1)} - N\lambda_k \mathrm{sgn}(\gamma_k)}{\mathbf{w}_k^T \mathbf{w}_k} \\ &= \gamma_k^{(0)} - \frac{N\lambda_k \mathrm{sgn}(\gamma_k^{(0)})}{\mathbf{w}_k^T \mathbf{w}_k} \end{aligned} \tag{42}$$

3.2 The new algorithm using combined modified Gram-Schmidt algorithm, basis pursuit and D-optimality

In this section a new algorithm is introduced that combines the modified Gram-Schmidt algorithm with the basis pursuit gradient tuning for new parameter estimation. The model selective criteria by D-optimality of Section 2.2 [14] is applied in the proposed algorithm. The algorithm is introduced as follows, in which, the basis pursuit parameters are assumed to be predetermined.

The modified Gram-Schmidt algorithm combining basis pursuit and D-optimality:

The Gram-Schmidt orthogonalisation scheme can be used to derive a simple and efficient algorithm for selecting subset models. Introducing the definition of $\mathbf{P}^{(k-1)}$ as

$$\mathbf{P}^{(k-1)} = [\mathbf{w}_1, \cdots, \mathbf{w}_{k-1}, \mathbf{p}_k^{(k-1)}, \cdots, \mathbf{p}_M^{(k-1)}] \tag{43}$$

If some of the columns $\mathbf{p}_k^{(k-1)}, \cdots, \mathbf{p}_M^{(k-1)}$ in $\mathbf{P}^{(k-1)}$ have been interchanged, this will still be referred to as $\mathbf{P}^{(k-1)}$ for notational convenience. The kth stage of the forward regression selection procedure is given below

1. For $k \le j \le M$, compute

$$\gamma_k^{(j)} = \frac{(\mathbf{p}_j^{(k-1)})^T \mathbf{y}^{(k-1)}}{(\mathbf{p}_j^{(k-1)})^T \mathbf{p}_j^{(k-1)}} \tag{44}$$

$$J_j^{(k)} = J^{(k-1)} - \frac{1}{N}[\gamma_k^{(j)}]^2 \kappa_k^{(j)} + \alpha \log[\frac{1}{\kappa_k^{(j)}}] \tag{45}$$

2. Find

$$J^{(k)} = J_{j_k}^{(k)} = \min\{J_j^{(k)}, \quad k \le j \le M\} \tag{46}$$

Then the j_kth column of $\mathbf{P}^{(k-1)}$ is interchanged with the kth column of $\mathbf{P}^{(k-1)}$, and the j_kth column of $\mathbf{A}$ up to the $(k-1)$th row is interchanged with the kth column of $\mathbf{A}$. This effectively selects the j_kth candidates as the kth regressor in the subset model. Then set $\gamma_k^{(0)} = \gamma_k^{(j_k)}$.

3. Perform the orthogonalization as follows

$$\begin{aligned}
\mathbf{w}_k &= \mathbf{p}_k^{(k-1)} \\
\alpha_{kj} &= \frac{\mathbf{w}_k^T \mathbf{p}_j^{(k-1)}}{\mathbf{w}_k^T \mathbf{w}_k}, \qquad k+1 \leq j \leq M \\
\mathbf{p}_j^{(k)} &= \mathbf{p}_j^{(k-1)} - \alpha_{kj}\mathbf{w}_k, \qquad k+1 \leq j \leq M
\end{aligned} \tag{47}$$

   to transform $\mathbf{P}^{(k-1)}$ into $\mathbf{P}^{(k)}$ and derive the kth row of $\mathbf{A}$. Update κ_k.

4. With $\gamma_k^{(0)} \neq 0$ as initialized parameter estimates, the optimal solution of learning law (35) is given by (42), and is rewritten here

$$\gamma_k = \gamma_k^{(0)} - \frac{N\lambda_k \mathrm{sgn}(\gamma_k^{(0)})}{\mathbf{w}_k^T \mathbf{w}_k} \tag{48}$$

   where $\lambda_k < \frac{1}{2N}|\mathbf{w}_k^T \mathbf{y}^{(k-1)}|$.

5. Update $\mathbf{y}^{(k-1)}$ into $\mathbf{y}^{(k)}$ by

$$\mathbf{y}^{(k)} = \mathbf{y}^{(k-1)} - \gamma_k \mathbf{w}_k \tag{49}$$

   and update

$$J^{(k)} = J^{(k-1)} - \frac{1}{N}\gamma_k^2 \kappa_k + \alpha \log[\frac{1}{\kappa_k}] \tag{50}$$

6. The selection is terminated at the n_θth stage where a subset model containing n_θ significant regressors by the D-optimality model selective criteria $J^{(k)}$ achieves a minimum.

Note that the assumption $\gamma_k^{(0)} \neq 0$ in Theorem 2 is actually true for the selected regressors before the model achieves sufficient approximation. By (50) of step (5), it is clear that if $\gamma_k = 0$, the procedure terminates. In forward regression selection, each regressor is selected from Step (2), characterized by the largest reduction in $J^{(k)}$, hence $\gamma_k^{(0)} \neq 0$, before the current model residual $\mathbf{y}^{(k-1)}$ becomes white. Clearly, as model size k increases, if the parameter estimates are initialized with very small magnitudes from least squares estimates, the basis pursuit gradient tuning procedure in Step (4), will pull it even more towards zero by Theorem 2. If an arbitrary small threshold was set for zero, the parameter γ_k is obtained as zero. $J^{(k)}$ will then increase to terminate the selection procedure, at a sparser model than that of without basis pursuit gradient tuning procedure.

A method of choosing $\boldsymbol{\lambda}$

The identification algorithm introduced above uses a predetermined basis pursuit parameters $\boldsymbol{\lambda}$, which reflects a tradeoff between modelling errors and the l^1 norm of parameter vector. An inappropriate choice of $\boldsymbol{\lambda}$ (too large) will cause the term representing the modelling error in V of (11) to become insignificant in deriving parameter estimates and result in poor model approximation. By the general principle in data modelling of that a model with generalization is preferred, the choice of $\boldsymbol{\lambda}$ may be derived based on the commonly used method of cross-validation. In the following, we introduced a simple method of choosing $\boldsymbol{\lambda}$ by the basic principle of cross-validation. i.e. using two data sets, one for training and another for testing. This method however is only a heuristic approach, while other optimization methods of $\boldsymbol{\lambda}$ are still under investigation. For simplicity a single global basis pursuit λ is used, that is, $\lambda_1 = \lambda_2 = ... = \lambda$. By using the constraints of $\lambda_k < \frac{1}{2N}|\mathbf{w}_k^T\mathbf{y}|$, a feasible initial choice of λ is determined as $\lambda = \frac{1}{2N}|\mathbf{w}_{n_\theta^{(0)}}^T\mathbf{y}|$, where $n_\theta^{(0)}$ is the size of the model derived with the D-optimality selective criterion, by setting α arbitrarily small, without using basis pursuit [14]. In order to derive a model with excellent generalization, the complete modelling procedure of iterating the proposed algorithm, by incrementally increasing λ from zero in a controlled manner, is given as follows.

The iterative procedure of the proposed algorithm including choosing basis pursuit parameters

1. Initialization. Set an arbitrarily small α, applying the modelling procedure of [14] to derive a model with size $n_\theta^{(0)}$. (This is equivalent to the proposed algorithm with $\lambda = 0$.) and set $\lambda = \frac{1}{2N}|\mathbf{w}_{n_\theta^{(0)}}^T\mathbf{y}|$. Set a counter for iteration $j = 1$;
2. Applying the proposed algorithm with the new λ, to derive a model with the size of $n_\theta^{(j)} < n_\theta^{(j-1)}$. Set a new $\lambda = \frac{1}{2N}|\mathbf{w}_{n_\theta^{(j)}}^T\mathbf{y}|$ for next iteration of this step, while the mean squares errors (MSE) of the test data set is monitored; $j = j + 1$;
3. Step 2 is terminated when the MSE of the test data set achieves a minimum.

Note that heuristically, for each step j, $\lambda \propto |\mathbf{w}_{n_\theta^{(j)}}|$. By the property of forward regression that selects the term with the largest reduction of modelling error. It can be assumed that $|\mathbf{w}_i| > |\mathbf{w}_j|$, for $i > j$. This means that $\lambda_k = \lambda = \frac{1}{2N}|\mathbf{w}_{n_\theta^{(j)}}^T\mathbf{y}| < \frac{1}{2N}|\mathbf{w}_k^T\mathbf{y}^{(k-1)}|$, for $k < n_\theta^{(j)}$. As the iteration step j increases, the effect of basis pursuit cost function (shrinking the small parameters to zero) would derive at the smaller size $n_\theta^{(j)}$ compared to previous iteration step. Because a smaller model size means a larger value of $|\mathbf{w}_{n_\theta^{(j)}}|$, λ increases gradually with the iteration, which is terminated at a proper stage via it performance over the test data set. Alternatively, λ can be set as a very small value for general improvement in model sparseness.

4 Modelling examples

Example 1: Consider the benchmark *Henon* time series given by

$$z(t) = 1.4 - z^2(t-1) + 0.3z(t-2) \tag{51}$$

1000 data points were generated with an initial condition $y(0) = 0, y(1) = 0$. The data set was then added a very small noise $e(t)$ $N(0, 0.001^2)$ to form a noisy data set $y(t) = z(t) + e(t)$. The input vector is set as $\mathbf{x}(t) = [y(t-1), y(t-2)]^T$. 498 data samples from $t = 1 \sim 500$, were used as estimation set, and 500 data samples $t = 499 \sim 1000$ were used as test data. The Gaussian radial basis function was used to construct a full model set by using all the data in the estimation data set as centers $\mathbf{c}_i$, $i = 1, ...498$, and $p_i(\mathbf{x}(t)) = \exp\{-\frac{\|\mathbf{x}(t)-\mathbf{c}_i\|^2}{\sigma_i^2}\}$, with $\sigma_i = 1$, $\forall\ i$. The modelling starts with $\lambda = 0$, and $\alpha = 10^{-8}$ (an arbitrarily small coefficient for D-optimality). The iterative procedure of the proposed algorithm was applied. The model was automatically terminated at a 30 centers networks. The final basis pursuit parameter was derived at $\lambda = 1.77 \times 10^{-8}$. The modelling MSE for the test data set is derived at 4.7841×10^{-5}. Equivalently 99.97% output variance of the test data has been explained by the model. The modelling results for test data set is shown in Figure 1.

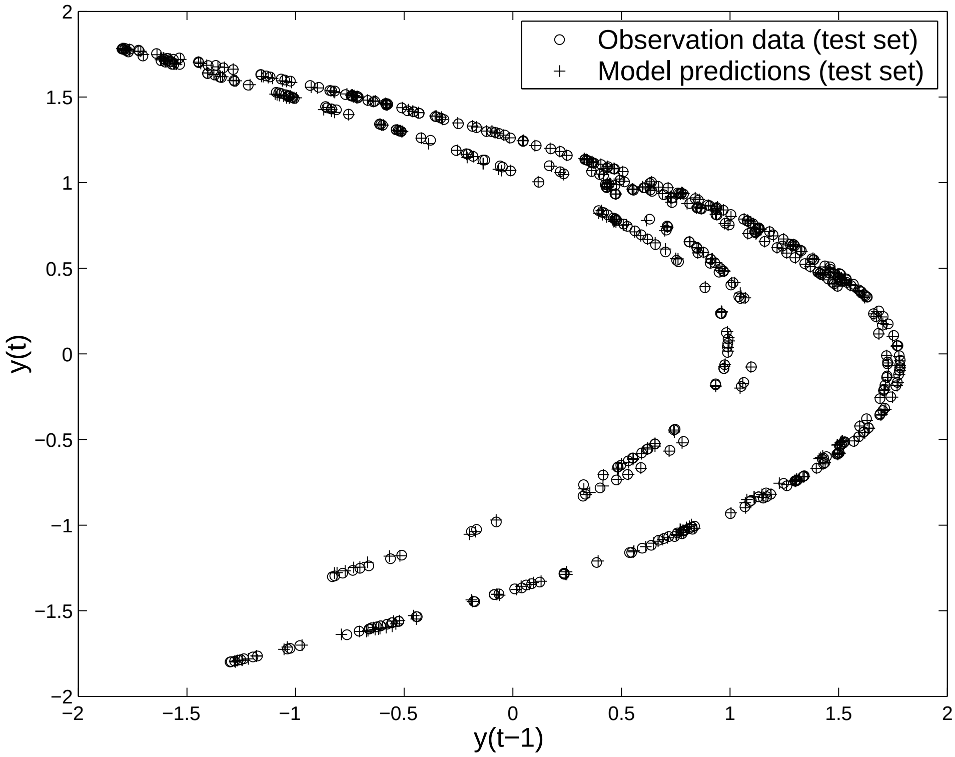


Figure 1: Modelling results for Example 1.

Example 2: Consider the chaotic two dimensional time series, *Ikeda map* [20], given by

$$\begin{bmatrix} x(t) \\ y(t) \end{bmatrix} = \begin{bmatrix} 1 + 0.9[x(t-1)\cos(r) - y(t-1)\sin(r)] \\ 0.9[x(t-1)\sin(r) + y(t-1)\cos(r)] \end{bmatrix}$$

$$\text{with} \qquad r = 0.4 - \frac{6.0}{1 + x^2(t-1) + y^2(t-1)} \tag{52}$$

1000 data points were generated with an initial condition $x(1) = 0.1, y(1) = 0.1$. Two models were constructed to model $x(t)$ and $y(t)$ respectively. For both models, the input vector is set as $\mathbf{x}(t) = [x(t-1), y(t-1)]^T$. 498 data samples from $t = 1 \sim 500$, were used as estimation set, and 500 data samples $t = 499 \sim 1000$ were used as test data. The Gaussian radial basis function was used to construct full model sets by using all the data in the estimation data set as centers $\mathbf{c}_i$, $i = 1, ...498$, and $p_i(\mathbf{x}(t)) = \exp\{-\frac{\|\mathbf{x}(t)-\mathbf{c}_i\|^2}{\sigma_i^2}\}$, with $\sigma_i = 0.5$, $\forall\ i$.

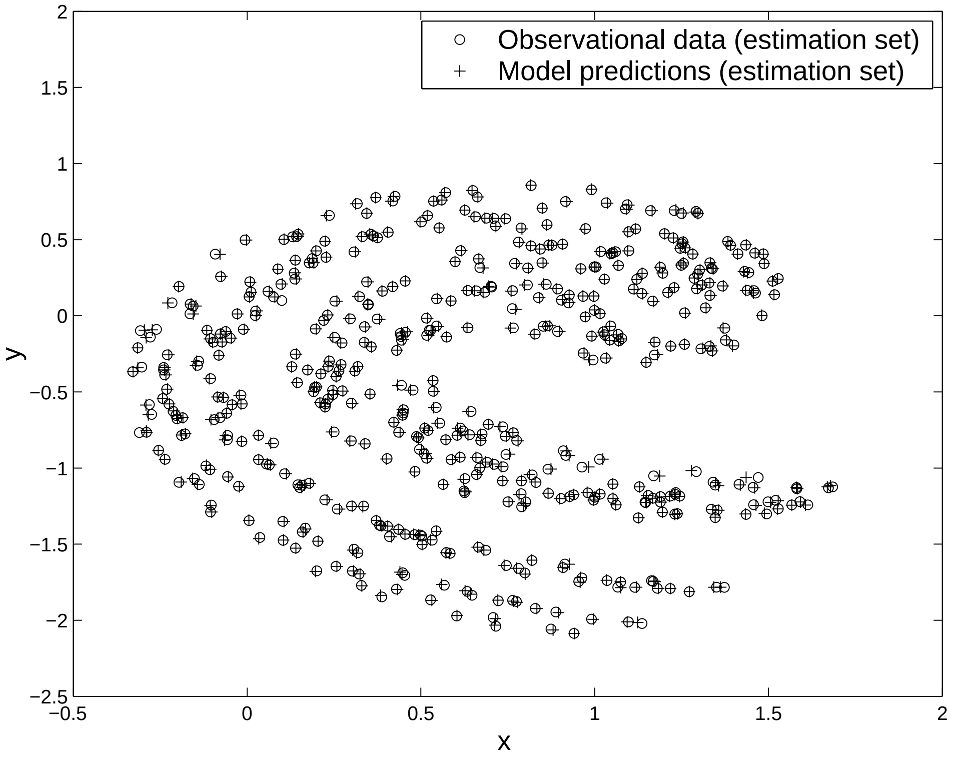
For the first model that models $x(t)$, the modelling starts with $\lambda = 0$, and $\alpha = 10^{-8}$ (an arbitrarily small coefficient for D-optimality). The iterative procedure of the proposed algorithm was applied. The model was automatically terminated at a 63 centers networks. The final basis pursuit parameter was derived at $\lambda = 7.7 \times 10^{-8}$. The modelling MSE for the test data set is derived at 3.13×10^{-5}. Equivalently 99.81% output variance of the test data has been explained by the model. For the second model that models $y(t)$, the modelling starts with $\lambda = 0$, and $\alpha = 10^{-8}$ (an arbitrarily small coefficient for D-optimality). The iterative procedure of the proposed algorithm was applied. The model was automatically terminated at a 66 centers networks. The final basis pursuit parameter was derived at $\lambda = 1.7 \times 10^{-8}$. The modelling MSE for the test data set is derived at 1.36×10^{-5}. Equivalently 99.94% output variance of the test data has been explained by the model. To illustrate the overall performance of the model in capturing the underlying system dynamics, the modelling results for both estimation and test data set is shown in Figure 2.

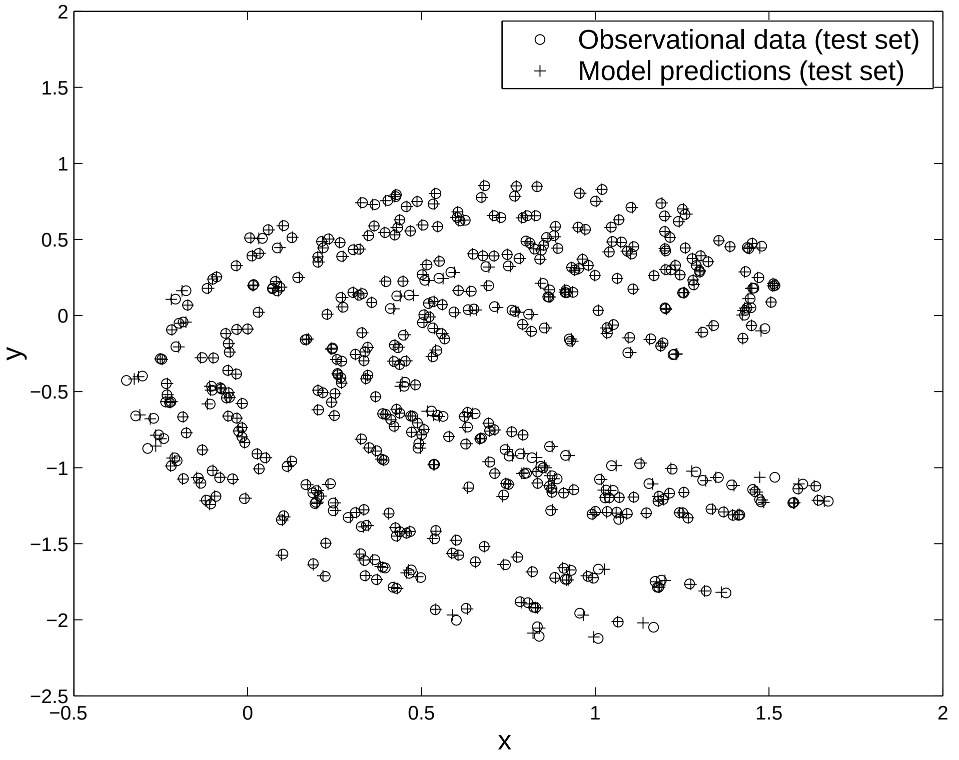
5 Conclusions

This paper has introduced a novel model identification algorithm for linear-in-the-parameters models. The proposed approach is based on the forward orthogonal least square algorithm using the modified Gram-Schmidt procedure. The approach aims to simultaneously optimize the model approximation ability, sparsity and robustness by combining the modified Gram-Schmidt algorithm with basis pursuit and D-optimality design. The main contribution is to tune the model parameters, in each forward regression step, with the basis pursuit that minimizes the l^1 norm of the parameter estimates vector. The D-optimality design criterion is used for model selection to ensure the model robustness and automatically terminates at a sparse model. The choice of basis pursuit parameters is discussed and a simple iterative procedure of the proposed algorithm is introduced to obtain a model with good generalization. Both the parameter tuning procedure, based on basis pursuit, and the model selection criterion, based on the D-optimality that is effective in ensuring model robustness, are integrated with the forward regression to maintain computational efficiency.

Acknowledgements

XH gratefully acknowledges that part of this work was supported by EPSRC in the UK.


Observational data (estimation set)
Model predictions (estimation set)
2
1.5
1
0.5
0
−0.5
−1
−1.5
−2
−2.5
y
−0.5
0
0.5
1
1.5
2
x


Observational data (test set)
Model predictions (test set)
2
1.5
1
0.5
0
−0.5
−1
−1.5
−2
−2.5
y
−0.5
0
0.5
1
1.5
2
x

References

[1] C. J. Harris, X. Hong, and Q. Gan, *Adaptive Modelling, Estimation and Fusion from Data: A Neurofuzzy Approach*, Springer-Verlag, 2002.

[2] M. Brown and C. J. Harris, *Neurofuzzy Adaptive Modelling and Control*, Prentice Hall, Hemel Hempstead, 1994.

[3] K. M. Bossley, *Neurofuzzy Modelling Approaches in System Identification*, Ph.D. thesis, Dept of ECS, University of Southampton, 1997.

[4] R. Murray-Smith and T. A. Johansen, *Multiple Model Approaches to Modelling and Control*, Taylor and Francis, 1997.

[5] R. Bellman, *Adaptive Control Processes*, Princeton University Press, 1966.

[6] J. S. R. Jang, C.T. Sun, and E. Mizutani, *Neuro-Fuzzy and Soft Computing: A Computational Approach to Learning and Machine Intelligence*, Upper Saddle River, NJ : Prentice Hall, 1997.

[7] T. Takagi and M. Sugeno, "Fuzzy identification of systems and its applications to modelling and control," *IEEE Trans. on Systems, Man, and Cybernetics*, vol. 15, pp. 116–132, 1985.

[8] S. Chen, S. A. Billings, and W. Luo, "Orthogonal least squares methods and their applications to non-linear system identification," *International Journal of Control*, vol. 50, pp. 1873–1896, 1989.

[9] S. Chen, Y. Wu, and B. L. Luk, "Combined genetic algorithm optimization and regularized orthogonal least squares learning for radial basis function networks," *IEEE Trans. on Neural Networks*, vol. 10, pp. 1239–1243, 1999.

[10] M. J. L. Orr, "Regularisation in the selection of radial basis function centers," *Neural Computation*, vol. 7, no. 3, pp. 954–975, 1995.

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

[12] A. C. Atkinson and A. N. Donev, *Optimum Experimental Designs*, Clarendon Press, Oxford, 1992.

[13] X. Hong and C. J Harris, "Nonlinear model structure detection using optimum experimental design and orthogonal least squares," *IEEE Transactions on Neural Networks*, vol. 12, no. 2, pp. 435–439, 2001.

[14] X. Hong and C. J Harris, "Nonlinear model structure design and construction using orthogonal least squares and d-optimality design," *IEEE Transactions on Neural Networks*, vol. 13, no. 5, pp. 1245–1250, 2001.

[15] S. Chen, "Local regularization assisted orthogonal least squares regression," *Submitted to International Journal of Control*, 2003.

[16] S. Chen, X. Hong, and C. J. Harris, “Sparse kernel regression modelling using combined locally regularised orthogonal least squares and d-optimality experimental design,” *IEEE Trans. on Automatic Control*, p. Accepted., 2003.

[17] D. J. C. MacKay, *Bayesian Methods for Adaptive Models*, Ph.D. thesis, California Institute of Technology, USA, 1991.

[18] S. S. Chen, D. L. Donoho, and M. A. Saunders, “Atomic decomposition by basis pursuit,” *SIAM Review*, vol. 43, no. 1, pp. pp129–159, 2001.

[19] B. Efron, I. Johnstone, T. Hastie, and R. Tibshirani, “Least angle regression,” *Annals of Statistics*, p. To Appear, 2003.

[20] K. Ikeda, “Multiple-valued stationary state and its instability of the transmitted light by a rign cavity system,” *Optics Communications*, vol. 30, no. 2, pp. 257–261, 1979.