Sparse model identification using orthogonal forward regression with basis pursuit and D-optimality

X. Hong, M. Brown, S. Chen and C.J. Harris

Abstract: An efficient model identification algorithm for a large class of linear-in-the-parameters models is introduced that simultaneously optimises 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 minimises the l^1 norm of the parameter estimate vector. The model subset selection cost function includes a D-optimality design criterion that maximises the determinant of the design matrix of the subset to ensure model robustness and to enable the model selection procedure to automatically terminate 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. Examples 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 [1–4]. 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 procedure 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].

Regularisation techniques have been incorporated into the OLS algorithm to produce a regularised orthogonal least squares (ROLS) algorithm that reduces the variance of parameter estimates [9, 10]. To produce a model with good generalisation capabilities, model selection criteria such as the Akaike information criterion (AIC) [11] are usually incorporated into the procedure to determine 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 penalise regressors that might cause poor model performance, e.g. too large parameter variance or ill-posedness of the regression matrix, if this is selected.

While OLS is based on the standard QR factorisation, principal component analysis (PCA) is widely used to reduce the input dimensions based on the singular vector decompositions (SVD) [12] in signal processing applications. The derived model is based on an orthogonal basis that are a few significant hidden variables constructed by the full set of input variables. By using the full set of input variables, more sophisticated parameter regularisation (hierarchical prior) [13] and the Markov-chain Monte Carlo (MCMC) algorithm, improved approximation/generation performance can be achieved with a trade-off of high computational expense. However, the OLS remains a popular practical approach in dynamical system modelling due to less computational expense compared with SVD, and the ease of conversion from the orthogonal basis to only a few selected original input variables, as these are essential requirements for online system condition monitoring and control objectives.

In optimum experimental design [14], 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 [15, 16], we have outlined efficient learning algorithms in which composite cost functions were introduced to optimise the

IEE Proceedings online no. 20040693

doi: 10.1049/ip-cta:20040693

Paper first received 2nd July 2003 and in revised form 20th April 2004

X. Hong is with the Department of Cybernetics, University of Reading, Reading, RG6 6AY, UK

M. Brown is with the Department of Computing and Mathematics, Manchester Metropolitan University, Manchester, UK

S. Chen and C.J. Harris are with the Department of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, UK

model approximation ability by using the forward OLS algorithm [8], and simultaneously the model adequacy by using an A-optimality design criterion (i.e. minimises the variance of the parameter estimates), or a D-optimality criterion (i.e. optimises the parameter efficiency and model robustness via the maximisation 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 regularised orthogonal least squares (LROLS) model selection [17] with D-optimality experimental design further enhances model robustness [18]. It has been shown [18, 19] that the parameter regularisation is equivalent to a maximised *a posterior PDF* (MAP) of parameters from bayesian viewpoint by adopting a gaussian prior for parameters.

The regularisation [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 [20, 21] that aims to obtain a model by minimising the l^1 norm of the parameters. The bayesian interpretation for the basis pursuit method is simply by adopting an exponential prior for parameters (Section 2.1). The advantage of basis pursuit is that it can achieve much sparser models by forcing more parameters to zero than models derived from the minimisation of the l^p norm, as most l^p norms will produce parameters small, but nonzero, values. Compared with the method of regularisation [9, 10] the basis pursuit method 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 optimisation problem with a simple solution into a more sophisticated problem for which a convex, nonquadratic optimisation is generally required [20, 21].

In this paper a new model identification technique is introduced by using forward regression with basis pursuit and D-optimality design. Based on previous work [15] 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 initializations 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 optimisation method, so the computational efficiency of the method due to the forward OLS regression maintains.

2 Preliminaries

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

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

where $t = 1, 2, \ldots, N$, and N is the size of the estimation data set, $y(t)$ is system output variable, $\boldsymbol{x}(t) = [y(t-1), \ldots, y(t-n_y), u(t-1), \ldots, u(t-n_u)]^T$ is the 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 and $\xi(t)$ is an uncorrelated model residual sequence with zero mean and variance of σ^2. Equation (1) can be written in matrix form as

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

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

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

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

$$\boldsymbol{P} = \boldsymbol{WA} \tag{3}$$

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

$$\boldsymbol{W}^T\boldsymbol{W} = diag\{\kappa_1, \ldots, \kappa_M\} \tag{4}$$

with

$$\kappa_k = \boldsymbol{w}_k^T\boldsymbol{w}_k, \quad k = 1, \ldots, M \tag{5}$$

so that (2) can be expressed as

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

where $\boldsymbol{\Gamma} = [\gamma_1, \ldots, \gamma_M]^T$ is an auxiliary vector.

2.1 Modified Gram–Schmidt algorithm, parameter regularisation and basis pursuit

For the orthogonalised system (6) the least squares estimates is given by

$$\gamma_k^{(0)} = \frac{\boldsymbol{w}_k^T\boldsymbol{y}}{\boldsymbol{w}_k^T\boldsymbol{w}_k}, \quad k = 1, \ldots, M \tag{7}$$

The original model coefficient vector $\boldsymbol{\Theta} = [\theta_1, \ldots, \theta_M]^T$ can then be calculated from $\boldsymbol{A\Theta} = \boldsymbol{\Gamma}$ through back substitution. The modified Gram–Schmidt procedure, described subsequently, can be used to perform the orthogonalisation of (3) and parameter estimation (7). Starting from $k = 1$, the columns $\boldsymbol{p}_j$, $k+1 \le j \le M$ are made orthogonal to the kth column at the kth stage. The operation is repeated for $1 \le k \le M-1$. Specifically, denoting $\boldsymbol{p}_j^{(0)} = \boldsymbol{p}_j, 1 \le j \le M$, then for $k = 1, \ldots, M-1$

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

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

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

It can be easily verified that $\gamma_k^{(0)}$ as derived from (9) is equivalent to (7). Geometrically the system output vector $\boldsymbol{y}$

at step k is projected onto a set of orthogonal basis vectors $\{\boldsymbol{w}_1, \ldots \boldsymbol{w}_k\}$. The model residual is decreased by projecting the system output vector $\boldsymbol{y}$ onto a new basis $\boldsymbol{w}_k$ at this step. Effectively (9) can be regarded as a linear fitting of $\boldsymbol{y}^{(k-1)}$ by using a single variable $\boldsymbol{w}^{(k)}$, and to derive the new model residual $\boldsymbol{y}^{(k)}$, and so on. This observation is 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, regularisation can be applied. If regularisation is performed to the parameter in orthogonal space, γ_k, then (9) is simply replaced by the following

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

where $\lambda_k \leq 0$ are regularisation parameters which can be optimised by being treated as hyperparameters in the Bayesian approach [18]. These results are obtained by setting the parameter optimiser as

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

Because the regularisation 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 optimisation.

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

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

where $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_{n_\theta}]^T$, $\|\Gamma\|_1 = [|\gamma_1|, \ldots, |\gamma_{n_\theta}|]^T$, 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 penalised, 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 regularisation. Because the solution of (11) is a nonquadratic optimisation 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 optimisation is required [20]. 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.

2.2 Bayesian regularisation and basis pursuit

The regularised parameter estimator by optimising $V^{(r)}$ is equivalent to a maximised *a posterior PDF* (MAP) of parameters in a bayesian approach [19, 18]. By the 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

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

whose maximisation leads to maximum likelihood (ML) parameter estimator, which is equivalent to the 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

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

where $Z_\Gamma^{(r)}$ is a normalising coefficient. The MAP estimator can be derived by minimising $V^{(r)}$ [1, 18, 19]. Clearly for basis pursuit estimator, the prior $p(\Gamma)$ is simply set as

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

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

2.3 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 $\boldsymbol{p}_k = [p_k(1), \ldots, p_k(N)]^T$, $k = 1, \ldots, n_\theta$, from M regressors to form a set of orthogonal basis $\boldsymbol{w}_k$, $k = 1, \ldots, n_\theta$, in a forward regression manner. As the orthogonality property $\boldsymbol{w}_i^T \boldsymbol{w}_j = 0$ for $i \neq j$ holds, if (6) is multiplied by itself and then the time average is taken, the following equation is easily derived:

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

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]

$$[ERR]_k = \frac{\gamma_k^2 \boldsymbol{w}_k^T \boldsymbol{w}_k}{\boldsymbol{y}^T \boldsymbol{y}}, \quad k = 1, \ldots, M \tag{17}$$

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 $J^{(0)} = \boldsymbol{y}^T \boldsymbol{y}$. At the kth forward regression stage a candidate regressor is selected as the kth regressor if it produces the smallest $J^{(k)}$. Equation (19) can be modified to form an alternative model selective criterion to enhance model robustness. The D-optimality-based cost function is one of robustness design criterion in experimental design

criteria [14]. The D-optimality criterion is to maximise the determinant of the design matrix defined as $\boldsymbol{W}_k^T\boldsymbol{W}_k$, where $\boldsymbol{W}_k \in \Re^{N\times n_\theta}$ denotes the resultant regression matrix, consisting of n_θ regressors selected from M regressors in $\boldsymbol{W}$

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

It can be easily verified that the selection of the a subset of $\boldsymbol{W}_k$ from $\boldsymbol{W}$ is equivalent to the selection of the a subset of n_θ regressors from $\boldsymbol{P}$ [16]. To include D-optimality as a model selective criterion for improved model robustness, construct an augmented cost function as

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

where α is a positive small number. Note that this composite cost function simultaneously minimises (19) and maximises (20) [16]. Equation (21) can be directly incorporated into the forward OLS algorithm to select the most relevant kth regressor at the kth forward regression stage, via

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

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 $\log(1/J_D)$ 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 is applied in the proposed new model identification algorithm introduced in following 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 basis 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

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

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 $\mathrm{sgn}\,\Theta = [\mathrm{sgn}\,\theta_1, \ldots, \mathrm{sgn}\,\theta_M]^T$, in which

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

η is an arbitrarily small positive number, then

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

where the basis pursuit cost function $V(t) = \frac{1}{2}\overline{\xi^2(t)} + \boldsymbol{\lambda}^T\|\Theta\|_1$, and $\|\Theta\|_1 = [|\theta_1|, \ldots, |\theta_{n_\theta}|]^T$ 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 neighbourhood around a current parameter estimate $\Theta(t) = [\theta_1(t), \ldots, \theta_M(t)]^T$, by the first-order Taylor series expansion of $V(t)$

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

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

$$\begin{aligned} \Delta V(t) = &-\eta\left\{\overline{\xi(t)\frac{\partial f}{\partial\Theta}} - \boldsymbol{\lambda}^T\mathrm{sgn}\,\Theta(t)\right\}^T \\ &\times\left\{\overline{\xi(t)\frac{\partial f}{\partial\Theta}} - \boldsymbol{\lambda}^T\mathrm{sgn}\,\Theta(t)\right\} \leq 0 \end{aligned} \tag{28}$$

that is, $V(t)$ is nonincreasing with a lower bound. Hence

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

Hence property (i) is established.

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

yielding

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

for a finite k

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

so property (ii) follows.

In the proposed algorithm of Section 3.2, this 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 $\boldsymbol{y}^{(k-1)}$ by using a single variable $\boldsymbol{w}^{(k)}$ with the least-squares method. The derived model residual vector $\boldsymbol{\Xi}$ is then set as $\boldsymbol{y}^{(k)}$. 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 the learning algorithm of (24) that optimises the basis pursuit's function given by (11). Following (9), denote $\boldsymbol{y}^{(k-1)} = [y^{(k-1)}(1), y^{(k-1)}(2), \ldots, y^{(k-1)}(N)]^T$ and

$\boldsymbol{w}_k = [w_k(1), \ldots, w_k(N)]^T$. 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 initialised as the least-square parameter estimate $\gamma_k^{(0)} \neq 0$, given by (9), and if $\lambda_k < \frac{1}{2N}\left|\boldsymbol{w}_k^T \boldsymbol{y}\right|$, then the final converged parameter estimate γ_k

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

Proof:

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

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

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

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

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

(ii) For an arbitrary small learning rate η it can be assumed that the parameter changes in an arbitrarily 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 nonincreasing, yields

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

So

$$\begin{aligned} \frac{1}{2N}[\boldsymbol{y}^{(k-1)}]^T\boldsymbol{y}^{(k-1)} &\leq \frac{1}{2N}[\boldsymbol{y}^{(k-1)}]^T\boldsymbol{y}^{(k-1)} \\ &\quad - \frac{1}{2N}\left[\gamma_k^{(0)}\right]^2 \boldsymbol{w}_k^T\boldsymbol{w}_k + \lambda_k\left|\gamma_k^{(0)}\right| \end{aligned} \tag{38}$$

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

and by applying the least-square solution $\left[\gamma_k^{(0)} = \frac{\boldsymbol{w}_k^T\boldsymbol{y}^{(k-1)}}{\boldsymbol{w}_k^T\boldsymbol{w}_k}\right]$ yields

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

This is contradictory to the assumption for λ_k. Therefore γ_k should not change sign throughout conditional on $\lambda_k < \frac{1}{2N}\left|\boldsymbol{w}_k^T\boldsymbol{y}\right|$, 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 the model size k increases, the parameter estimates γ_k as initialised 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

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

or

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

3.2 New algorithm using combined modified Gram–Schmidt algorithm, basis pursuit and D-optimality

The model selective criteria by D-optimality of Section 2.2 [16] is applied in the proposed algorithm. The algorithm is introduced as follows, in which, the basis pursuit parameters are assumed to be predetermined.

3.2.1 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 $\boldsymbol{P}^{(k-1)}$ as

$$\boldsymbol{P}^{(k-1)} = \left[\boldsymbol{w}_1, \ldots, \boldsymbol{w}_{k-1}, \boldsymbol{p}_k^{(k-1)}, \ldots, \boldsymbol{p}_M^{(k-1)}\right] \tag{43}$$

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

(i) For $k \leq j \leq M$, compute

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

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

(ii) Find

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

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

(iii) Perform the orthogonalisation as follows:

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

to transform $\boldsymbol{P}^{(k-1)}$ into $\boldsymbol{P}^{(k)}$ and derive the kth row of $\boldsymbol{A}$. Update κ_k.
(iv) With $\gamma_k^{(0)} \neq 0$ as initialised parameter estimates, the optimal solution of learning law (35) is given by (42), and is rewritten here

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

where

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

(v) Update $\boldsymbol{y}^{(k-1)}$ into $\boldsymbol{y}^{(k)}$ by

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

and update

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

(vi) 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 (v), it is clear that if $\gamma_k = 0$, the procedure terminates. In forward regression selection each regressor is selected from step (ii) characterised by the largest reduction in $J^{(k)}$, hence $\gamma_k^{(0)} \neq 0$, before the current model residual $\boldsymbol{y}^{(k-1)}$ becomes white. Clearly, as the model size k increases, if the parameter estimates are initialised with very small magnitudes from least-squares estimates the basis pursuit gradient tuning procedure in step (iv) 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.

3.2.2 Method of choosing $\boldsymbol{\lambda}$: The identification algorithm introduced uses a predetermined basis pursuit parameter $\boldsymbol{\lambda}$, 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 a model with generalisation is preferred, the choice of $\boldsymbol{\lambda}$ may be derived based on the commonly used method of cross-validation. In the following we introduce 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 is only a heuristic approach; other optimisation methods of $\boldsymbol{\lambda}$ are still under investigation. For simplicity a single global basis pursuit λ is used, i.e. $\lambda_1 = \lambda_2 = \ldots = \lambda$. By using the constraints of $\lambda_k < (1/2N)|\boldsymbol{w}_k^T \boldsymbol{y}|$, a feasible initial choice of λ is determined as $\lambda = (1/2N)|\boldsymbol{w}_{n_\theta^{(0)}}^T \boldsymbol{y}|$, 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 [16]. To derive a model with excellent generalisation the complete modelling procedure of iterating the proposed algorithm by incrementally increasing λ from zero in a controlled manner is given as follows.

3.2.3 Iterative procedure of proposed algorithm including choosing basis pursuit parameters:

(i) Initialisation. Set an arbitrarily small α, applying the modelling procedure of [16] to derive a model with size $n_\theta^{(0)}$. (This is equivalent to the proposed algorithm with $\lambda = 0$) and set $\lambda = (1/2N)|\boldsymbol{w}_{n_\theta^{(0)}}^T \boldsymbol{y}|$. Set a counter for iteration $j = 1$.

(ii) Apply the proposed algorithm with the new λ to derive a model with size $n_\theta^{(j)} < n_\theta^{(j-1)}$. Set a new $\lambda = (1/2N)|\boldsymbol{w}_{n_\theta^{(j)}}^T(j)\boldsymbol{y}|$ for the next iteration of this step, while the mean squares error (MSE) of the test data set is monitored; $j = j + 1$.
(iii) Step (ii) is terminated when the MSE of the test data set achieves a minimum.

Note that heuristically, for each step j, $\lambda \propto |\boldsymbol{w}_{n_\theta^{(j)}}|$. Forward regression selects the term with the largest reduction of modelling error. It can be assumed that $|\boldsymbol{w}_i| > |\boldsymbol{w}_j|$, for $i > j$. This means that $\lambda_k = \lambda = (1/2N)|\boldsymbol{w}_{n_\theta^{(j)}}^T \boldsymbol{y}| < (1/2N)|\boldsymbol{w}_k^T \boldsymbol{y}^{(k-1)}|$, 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 with the previous iteration step. Because a smaller model size means a larger value of $|\boldsymbol{w}_{n_\theta^{(j)}}|$, 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

4.1 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 $z(0) = 0, z(1) = 0$. 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 $\boldsymbol{x}(t) = [y(t-1), y(t-2)]^T$; 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 centres $\boldsymbol{c}_i$, $\boldsymbol{i} = 1, \ldots 498$, and $p_i(\boldsymbol{x}(t)) = \exp\{-\|\boldsymbol{x}(t) - \boldsymbol{c}_i\|^2/\sigma_i^2\}$, 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 30 centres. The final basis-pursuit parameter was derived at $\lambda = 1.77 \times 10^{-8}$. The modelling MSE for the test data set is derived as 4.7841×10^{-5}. Equivalently a 99.97% output variance of the test data has been explained by the model. The modelling results for the test data set are shown in Fig. 1.

4.2 Example 2

Consider the chaotic two-dimensional time series, *Ikeda map* [22], 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 $\boldsymbol{x}(t) = [x(t-1), y(t-1)]^T$. A total of 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 centres $\boldsymbol{c}_i,\ \boldsymbol{i} = 1, \ldots, 498$, and $p_i(\boldsymbol{x}(t)) = \exp\{-\|\boldsymbol{x}(t) - c_i\|^2/\sigma_i^2\}$, with $\sigma_i = 0.5,\ \forall i$.

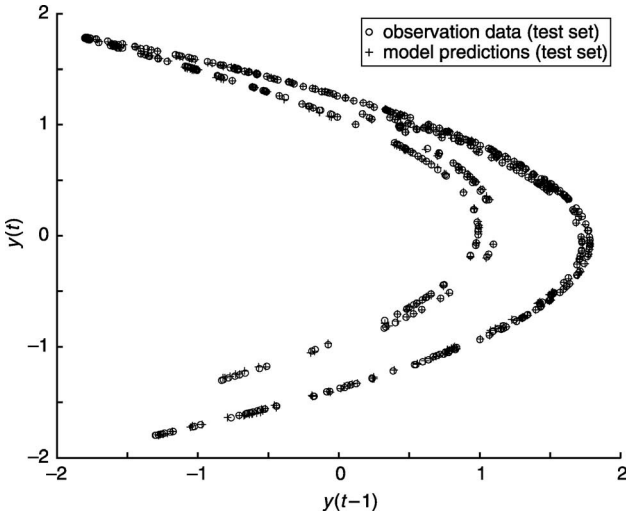

Fig. 1 *Modelling results for example 1*

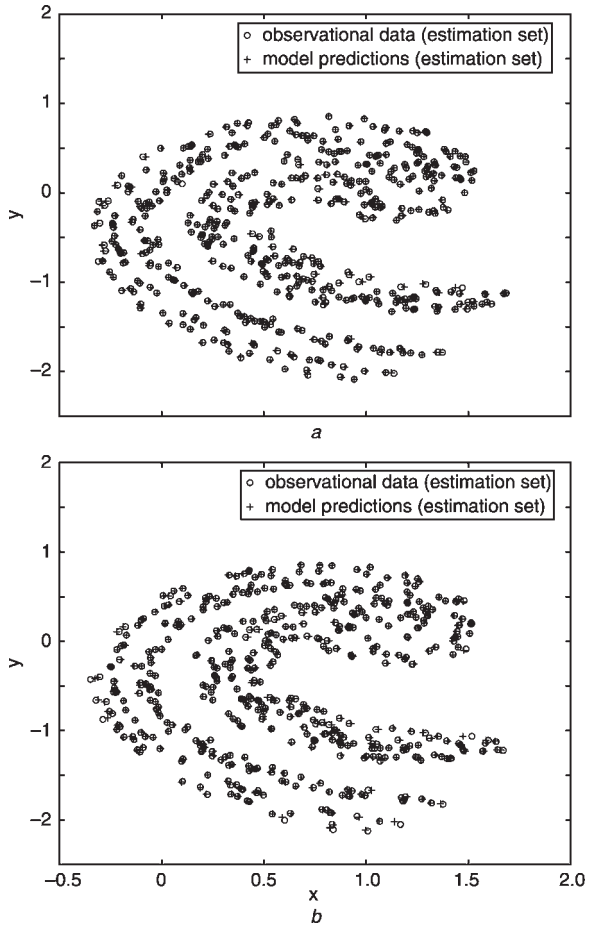

Fig. 2 *Modelling results for example 2*
a Training set
b Test set

For the first model that model $x(t)$, 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 63 centres. 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 a 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

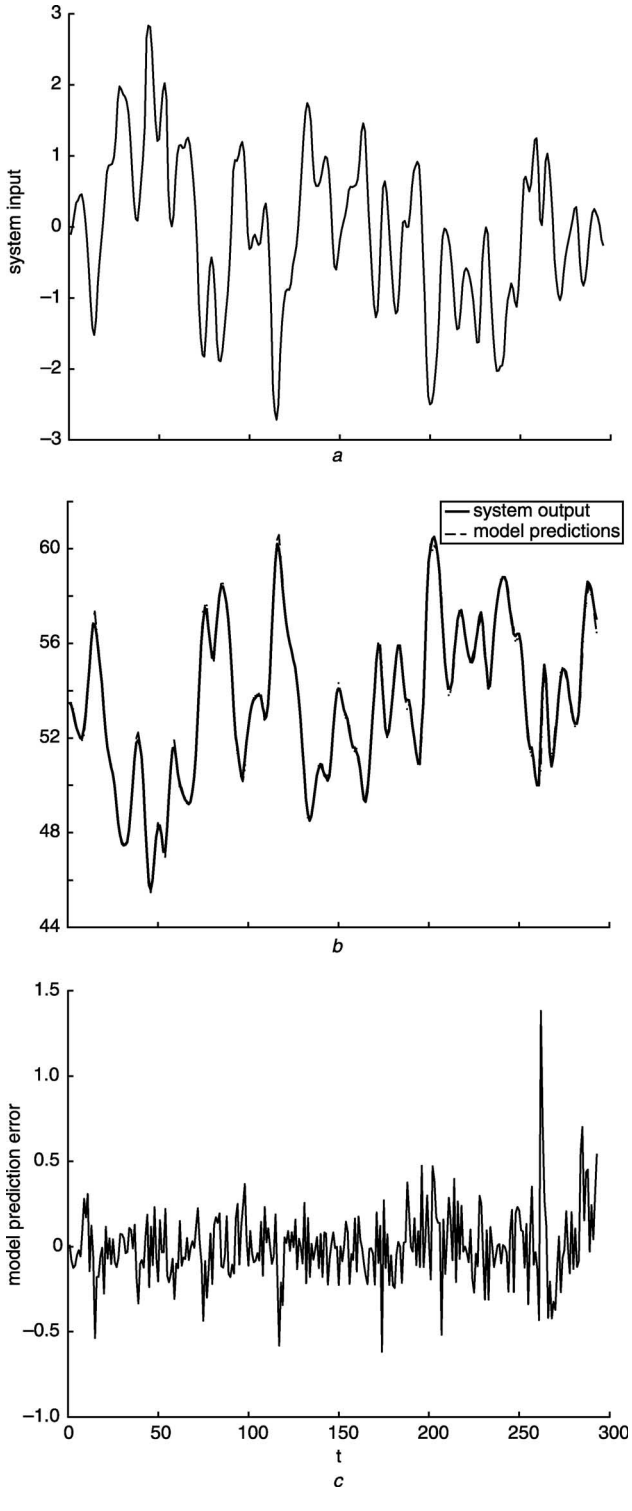

Fig. 3 *Modelling results for example 3*
a System input
b Model prediction and model output
c Model prediction error

small coefficient for D-optimality). The iterative procedure of the proposed algorithm was applied. The model was automatically terminated at 66 centres. 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 a 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 5 are shown in Fig. 2.

4.3 Example 3

The benchmarking gas furnace data (series J in [23]) set consists of 296 input / output pairs representing coded input gas-feed rate as input, $u(t)$, and CO_2 concentration from the gas furnace as output $y(t)$. All the data were used as training data set. A RBF network with the input vector $\boldsymbol{x}(t) = [u(t-1), u(t-2), u(t-3), y(t-1), y(t-2), y(t-3)]^T$, and the thin-plate-spline basis function $p_i(\boldsymbol{x}(t)) = \|\boldsymbol{x}(t) - \boldsymbol{c}_i\|^2 \log \|\boldsymbol{x}(t) - \boldsymbol{c}_i\|$ was used as the basis function with all data sets as candidate centres $\boldsymbol{c}_i$.

The iterative procedure of the proposed algorithm was applied with $\beta = 10^{-4}$. The model was automatically terminated at 36-centres. The final basis-pursuit parameter was derived at $\lambda = 0.0052$. The modelling MSE for the test data set is derived at 0.045. A list of results on the same data can be found in [24]. It can be seen that the results obtained in this study are comparable. The modelling results for both estimation and test data sets are shown in Fig. 3.

5 Conclusions

This paper has introduced a 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 optimise 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 minimises 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 generalisation. 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 forward regression to maintain computational efficiency.

6 Acknowledgments

XH gratefully acknowledges that part of this work was supported by the UK EPSRC. The authors would like to thank the referees for their constructive comments.

7 References

1 Harris, C.J., Hong, X., and Gan, Q.: 'Adaptive modelling, estimation and fusion from data: a neurofuzzy approach' (Springer-Verlag, Berlin, 2002)
2 Brown, M., and Harris, C.J.: 'Neurofuzzy adaptive modelling and control' (Prentice Hall, Hemel Hempstead, 1994)
3 Bossley, K.M.: 'Neurofuzzy modelling approaches in system identification'. PhD thesis, Dept of ECS, University of Southampton, 1997
4 Murray-Smith, R., and Johansen, T.A.: 'Multiple model approaches to modelling and control' (Taylor and Francis, London, 1997)
5 Bellman, R.: 'Adaptive control processes' (Princeton University Press, Princeton, 1966)
6 Jang, J.S.R., Sun, C.T., and Mizutani, E.: 'Neurofuzzy and soft computing: a computational approach to learning and machine intelligence' (Prentice Hall, Upper Saddle River, NJ, 1997)
7 Takagi, T., and Sugeno, M.: 'Fuzzy identification of systems and its applications to modelling and control', *IEEE Trans. Syst. Man Cybern.*, 1985, **15**, pp. 116–132
8 Chen, S., Billings, S.A., and Luo, W.: 'Orthogonal least squares methods and their applications to nonlinear system identification', *Int. J. Control*, 1989, **50**, pp. 1873–1896
9 Chen, S., Wu, Y., and Luk, B.L.: 'Combined genetic algorithm optimization and regularized orthogonal least-squares learning for radial basis function networks', *IEEE Trans. Neural Netw.*, 1999, **10**, pp. 1239–1243
10 Orr, M.J.L.: 'Regularisation in the selection of radial basis function centers', *Neural Comput.*, 1995, **7**, (3), pp. 954–975
11 Akaike, H.: 'A new look at the statistical model identification', *IEEE Trans. Autom. Control*, 1974, **19**, pp. 716–723
12 Hansen, P.C.: 'Rank-deficient and discrete ill-posed problems' (SIAM, Philadelphia, 1998)
13 Neal, R.M., and Zhang, J.: 'Classification for high dimensional problems using bayesian neural networks and dirichlet diffusion trees'. Presented at the NIPS Workshop on Feature Selection, Whistler, BC, 11–13 December 2003
14 Atkinson, A.C., and Donev, A.N.: 'Optimum experimental designs' (Clarendon Press, Oxford, 1992)
15 Hong, X., and Harris, C.J.: 'Nonlinear model structure detection using optimum experimental design and orthogonal least squares', *IEEE Trans. Neural Netw.*, 2001, **12**, (2), pp. 435–439
16 Hong, X., and Harris, C.J.: 'Nonlinear model structure design and construction using orthogonal least squares and D-optimality design', *IEEE Trans Neural Netw.*, 2001, **13**, (5), pp. 1245–1250
17 Chen, S.: 'Locally regularised orthogonal least squares algorithm for the construction of sparse kernel regression models'. Proc. 6th. Int. Conf. on Signal Processing, Beijing, China, 26–30 August 2002, pp. 1229–1232
18 Chen, S., Hong, X., and Harris, C.J.: 'Sparse kernel regression modelling using combined locally regularised orthogonal least squares and D-optimality experimental design', *IEEE Trans. Autom. Control*, 2003, **48**, (6), pp. 1029–1036
19 MacKay, D.J.C.: 'Bayesian methods for adaptive models'. PhD thesis, California Institute of Technology, USA, 1991
20 Chen, S.S., Donoho, D.L., and Saunders, M.A.: 'Atomic decomposition by basis pursuit', *SIAM Rev.*, 2001, **43**, (1), pp. 129–159
21 Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R.: 'Leastangle regression', *Ann. Sta.*, 2004, **32**, pp. 407–451
22 Ikeda, K.: 'Multiple-valued stationary state and its instability of the transmitted light by a ring cavity system', *Opt. Commun.*, 1979, **30**, (2), pp. 257–261
23 Box, G.E.P., and Jenkins, G.M.: 'Time series analysis, forecasting and control' (Holden-Day, London, 1976)
24 Chen, S., Hong, X., Harris, C.J., and Sharkey, P.M.: 'Sparse modelling using orthogonal forward regression with press statistic and regularisation', *IEEE Trans. Syst., Man Cybern., B, Cybern.*, 2004, **34**, (2), pp. 898–911