Backward Elimination Methods for Associative Memory Network Pruning

Xia Hong[†], Chris Harris[*], Martin Brown[‡], Sheng Chen[*]

[†]Department of Cybernetics
University of Reading, Reading, UK
[*]Department of Electronics and Computer Science
University of Southampton, Southampton, UK
[‡] Department of Computing and Mathematics
Manchester Metropolitan University, Manchester, UK

Abstract. Three hybrid data based model construction/pruning formula are introduced by using backward elimination as automatic postprocessing approaches to improved model sparsity. Each of these approaches is based on a composite cost function between the model fit and one of three terms of A-/D-optimality / (parameter 1-norm in basis pursuit) that determines a pruning process. The A-/D-optimality based pruning formula contain some orthogonalisation between the pruned model and the deleted regressor. The basis pursuit cost function is derived as a simple formula without need for an orthogonalisation process. These different approaches to parsimonious data based modelling are applied to the same numerical examples in parallel to demonstrate their computational effectiveness.

Keywords: nonlinear modelling, backward elimination, forward regression, model sparsity, generalization.

1 Introduction

Associative memory networks (such as B-spline networks, radial basis function (RBF) networks and support vector machines (SVM)) are widely used in nonlinear modelling applications (Harris et al. 2002, Brown and Harris 1994, Murray-Smith and Johansen 1997, Vapnik 1998). A fundamental topic in system identification is the construction of a sparse model that best represents the underlying dynamics from noisy measurement data. Typically the problem is configured and solved in a forward regression or backward elimination manner by selecting a subset from a full model set that consists of a large number of features, e.g. the outputs of hidden nodes of an associative memory network. Forward orthogonal least squares (OLS) algorithm (Chen et al. 1989, Chen et al. 1999, Orr 1993) is a more popular approach than backward elimination because it is self-contained and more efficient than a complete backward elimination procedure that eliminates regressors from a full model, in particular when the full model is very large. Note that backward elimination as a postprocessing procedure is computationally affordable, and this can be used to form hybrid approaches to prune a model that is identified via other approaches (Hong and Billings 1997). In this manner, an identified model with some structural redundancy is subject to backward elimination for improved sparsity. Conventional forward (backward) approach includes (removes) a model regressor one at a time based on largest improvement (least deterioration) in model fit.

Optimum experimental designs have been used (Atkinson and Donev 1992) to construct smooth network response surfaces based on the setting of the experimental variables under well controlled experimental conditions. In optimum design, model adequacy is evaluated by design criteria that are statistical measures of goodness of experimental designs by virtue of design efficiency and experimental effort. Quantitatively, model adequacy is measured as function of the eigenvalues of the design matrix. Recently variants of the forward OLS algorithms have been introduced by modifying the selective criteria to include A- and D-optimality in forward regression (Hong and Harris 2001, Hong and Harris 2002), in which, structure optimality in design theory are combined with the well known system identification algorithm to form hybrid approaches applicable to associative memory networks neural networks modelling.

In order to achieve improved model sparsity, an overcomplex model determined via some identification method can be then pruned (Reed 1993). The model structural pruning is also important as a mechanism for improved generalization. Extending previous work (Hong and Harris 2001, Hong and Harris 2002) to backward elimination leads to alternative postprocessing approaches to prune an identified model. Even a model constructed via forward regression can gain extra sparsity, because a regressor selected in a forward manner may become insignificant at a later stage and can then be removed. Alternatively a sparse model can be

achieved by using a basis pursuit cost function to derive parameters (Chen et al. 1999, Orr 1993, Chen et al. 2001, Efron et al. 2003). Model sparsity may be achieved through parameter estimation via constrained optimization that penalizes the norm of parameter vector, as used in SVM (Vapnik 1998). The optimal parameter solutions would derive some parameter estimate as zero. This is equivalent to some regressors removal from the model. The basis pursuit cost function contains a 1-norm term, and its optimality is characterized by some zero parameters in the model but not a large number of parameters with near zero values (Chen et al. 2001).

In this paper three simple pruning formula are introduced for linear-in-the-parameters models such as associative memory networks (Harris et al. 2002). The similarity of these approaches is that each is based on the balance between the model fit and one of three alternative pruning approaches of A-/D-optimality /(parameter 1-norm in basis pursuit) in order to determine a pruning process. Such a balance (Hong and Harris 2001, Hong and Harris 2002) means that the backward elimination processes will be terminated automatically (this is also true for basis pursuit cost function based pruning as shown in this study). The A-/D-optimality based pruning formula contains some orthogonalisation between the pruned model and the deleted regressor. By derivation the pruning formula of basis pursuit cost function is shown to be an extremely simple formula without need for an orthogonalisation process. Numerical examples are applied to demonstrate the relative effectiveness of these approaches.

2 Problem formulation

A linear-in-the-parameters model (e.g. radial basis function (RBF) neural network, B-spline neurofuzzy network) can be formulated as (Harris et al. 2002, Brown and Harris 1994)

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

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 of observables with assumed known dimension of $(n_y + n_u)$. $u(t)$ is system input variable. θ_k is a parameter and $p_k(\cdot)$ 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. M is number of regressors in the model. (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 $\mathbf{P} = [\mathbf{p}_1, \mathbf{p}_2, \cdots, \mathbf{p}_M]$, where $\mathbf{p}_i = [p_i(\mathbf{x}(1)), ...p_i(\mathbf{x}(N))]^T$. (1) can be written as

$$y(t) = \sum_{k=1, k\neq i}^{M} p_k(\mathbf{x}(t))\theta_k + p_i(\mathbf{x}(t))\theta_i + \xi(t), \quad i = 1, ..., M \tag{3}$$

whose matrix form is given by

$$\mathbf{y} = \mathbf{P}_{(-i)}\Theta_{(-i)} + \mathbf{p}_i\theta_i + \Xi \tag{4}$$

where $\mathbf{P}_{(-i)}$ is the new regression matrix by removing the ith column from $\mathbf{P}$, $\Theta_{(-i)}$ is derived from Θ by deleting the ith element. $\mathbf{p}_i$ can be made orthogonal with $\mathbf{P}_{(-i)}$ via an orthogonal projection operation of

$$\mathbf{w}_i = \left\{ \mathbf{I} - \mathbf{P}_{(-i)} \left[\mathbf{P}_{(-i)}^T \mathbf{P}_{(-i)} \right]^{-1} \mathbf{P}_{(-i)}^T \right\} \mathbf{p}_i \tag{5}$$

where $\mathbf{I}$ is a unit matrix with appropriate dimension, so that (4) can be rewritten as

$$\mathbf{y} = \mathbf{P}_{(-i)}\mathbf{g}_{(-i)} + \mathbf{w}_i\theta_i + \Xi \tag{6}$$

where $\mathbf{g}_{(-i)} = \Theta_{(-i)} + \left[\mathbf{P}_{(-i)}^T \mathbf{P}_{(-i)} \right]^{-1} \mathbf{P}_{(-i)}^T \mathbf{p}_i\theta_i$, and the least squares estimates of θ_i and $\mathbf{g}_{(-i)}$ is simply given by (Hong and Billings 1997)

$$\theta_i = \frac{\mathbf{w}_i^T \mathbf{y}}{\mathbf{w}_i^T \mathbf{w}_i} \tag{7}$$

and

$$\mathbf{g}_{(-i)} = \left[\mathbf{P}_{(-i)}^T \mathbf{P}_{(-i)}\right]^{-1} \mathbf{P}_{(-i)}^T \mathbf{y} \tag{8}$$

due to orthogonality between $\mathbf{P}_{(-i)}$ and $\mathbf{w}_i$. Note that $\mathbf{g}_{(-i)}$ (not $\mathbf{\Theta}_{(-i)}$) is the least squares estimator for the model with ith regressor p_i removed.

By setting $i = 1, 2, ..., M$ in turn, the deterioration of model fit of removing each regressor, p_i, from model (2) can be calculated as the increment of error variance (IEV) (Hong and Billings 1997), given by

$$IEV_i = \theta_i^2 \mathbf{w}_i^T \mathbf{w}_i \tag{9}$$

(6) can also be written as

$$\mathbf{y} = \mathbf{W}\mathbf{g} + \Xi \tag{10}$$

by denoting an auxiliary parameter vector $\mathbf{g} = [\mathbf{g}_{(-i)}^T, \theta_i]^T$, and an auxiliary regression matrix $\mathbf{W} = [\mathbf{P}_{(-i)} \mathbf{w}_i]$, with a subspace orthogonality property given by

$$\mathbf{W}^T\mathbf{W} = \begin{bmatrix} \mathbf{P}_{(-i)}^T \mathbf{P}_{(-i)} & \mathbf{0} \\ \mathbf{0}^T & \mathbf{w}_i^T \mathbf{w}_i \end{bmatrix} \tag{11}$$

where $\mathbf{0}$ is a zero vector with appropriate dimension.

It is natural to consider model subset selection from an initial model base with M regressors in the framework of the optimal experiment design. If $\mathbf{P}^T\mathbf{P}$ is nonsingular, then for the least squares parameter estimator $\hat{\mathbf{\Theta}}$ of $\mathbf{\Theta}$ in (2)

$$\begin{aligned} &(i)\ E\hat{\mathbf{\Theta}} = \mathbf{\Theta} \\ &(ii)\ \mathrm{cov}\hat{\mathbf{\Theta}} = \sigma^2(\mathbf{P}^T\mathbf{P})^{-1} \end{aligned} \tag{12}$$

where the matrix $(\mathbf{P}^T\mathbf{P})$ is called the design matrix. Consider the application of experimental design criteria in the context of model subset selection. In this section we initially introduce the concepts of experimental design criteria including A-optimality and D-optimality based on using a fixed sized subset. The subset model is constructed from the full model with regression matrix $\mathbf{P}$ by using n_θ regressors selected from M regressors in $\mathbf{P}$, $n_\theta \ll M$. The resultant regression matrix is denoted $\mathbf{P}_k \in \Re^{N \times n_\theta}$, the resultant design matrix by $\mathbf{P}_k^T\mathbf{P}_k$, and μ_k, $k = 1, ..., n_\theta$ are the eigenvalues of $\mathbf{P}_k^T\mathbf{P}_k$.

Definition 1: A-optimality criterion minimises the sum of the variance of a parameter estimate vector $\hat{\mathbf{\Theta}}$

$$\min\{J^A = \mathrm{tr}\left[\mathrm{cov}\hat{\mathbf{\Theta}}\right] = \sigma^2 \sum_{k=1}^{n_\theta} \frac{1}{\mu_k}\}. \tag{13}$$

Definition 2: The D-optimality criterion maximises the determinant of the design matrix of $\mathbf{P}_k^T\mathbf{P}_k$

$$\max\{J^D = \det(\mathbf{P}_k^T\mathbf{P}_k) = \prod_{k=1}^{n_\theta} \mu_k\}. \tag{14}$$

It is well known that a model based on least squares estimates tend to be unsatisfactory for a near ill-conditioned regression matrix (or design matrix). The D-optimality criterion (Atkinson and Donev 1992) inherently improves model robustness by favoring model with good design matrix condition.

3 Backward elimination approaches

Conventional backward elimination removes a model regressor one at a time based on the least deterioration in model fit, i.e., for an identified model given by (1), it can be pruned by removing the regressor which produces the smallest value of $IEV_i, \forall i$. The procedure is repeated for $(M - n_\theta) > 0$ steps, until some stopping criterion is satisfied to derive a model with $n_\theta < M$ regressors. For the k^{th} $(k = 1, 2, ...)$ step of procedure, the pruning is operated on a previous model with $(M - k + 1)$ regressors. In the following, three variants of the backward elimination approaches are introduced, based on experimental design criteria (A- and D-optimality) and basis pursuit parameter optimization cost function respectively.

3.1 Backward elimination approaches using experimental design criteria

3.1.1 Backward elimination based on A-optimality

In experimental design (Atkinson and Donev 1992), $\mathbf{P}^T\mathbf{P}$ is called a design matrix. The A-optimality criteria minimizes the sum of variance of least squares parameter estimates, i.e. the trace of the covariance matrix via (13). Consider a model given by (10) with auxiliary parameter vector $\mathbf{g}$ and regression matrix $\mathbf{W}$, the trace of the covariance matrix $\text{cov}(\mathbf{g}) = \sigma^2\text{trace}([\mathbf{W}^T\mathbf{W}]^{-1})$. From (11)

$$\text{trace}[(\mathbf{W}^T\mathbf{W})^{-1}] = \text{trace}[(\mathbf{P}_{(-i)}^T\mathbf{P}_{(-i)})^{-1}] + \frac{1}{\mathbf{w}_i^T\mathbf{w}_i}. \tag{15}$$

or

$$\text{trace}[\text{cov}(\mathbf{g})] = \text{trace}[\text{cov}(\mathbf{g}_{(-i)})] + \frac{\sigma^2}{\mathbf{w}_i^T\mathbf{w}_i}. \tag{16}$$

where $\text{trace}[\text{cov}(\mathbf{g}_{(-i)})] = \sigma^2\text{trace}[(\mathbf{P}_{(-i)}^T\mathbf{P}_{(-i)})]^{-1}]$ is the sum of variance of least squares parameter estimates based on the model after the i^{th} regressor is removed. This means that the reduction in A-optimality by deleting p_i is given by $\frac{\sigma^2}{\mathbf{w}_i^T\mathbf{w}_i}$. The reduction in a composite A-optimality cost function (Hong and Harris 2001) during each backward elimination step is given by

$$V_i^A = \frac{\alpha}{\mathbf{w}_i^T\mathbf{w}_i} - IEV_i \tag{17}$$

where α is a small positive parameter that controls the trade-off between A-optimality and model error variance. If $i_A = \arg[\max\{V_i^A, \forall i\} > 0]$, the model can be pruned by using backward elimination to remove the i_A^{th} regressor and reduce composite A-optimality cost function (Hong and Harris 2001). The procedure repeats until it is automatically terminated when $\max\{V_i^A, \forall i\} < 0$.

3.1.2 Backward elimination based on D-optimality

The D-optimality criteria maximizes the determinant of the design matrix $[\mathbf{P}^T\mathbf{P}]$. The relation between $\mathbf{W}$ and $\mathbf{P}$ can be expressed as

$$\mathbf{P} = [\mathbf{P}_{(-i)} \;\; \mathbf{p}_i] = [\mathbf{P}_{(-i)} \;\; \mathbf{w}_i]\mathbf{A} = \mathbf{W}\mathbf{A} \tag{18}$$

where

$$\mathbf{A} = \begin{bmatrix} \mathbf{I} & \left[\mathbf{P}_{(-i)}^T\mathbf{P}_{(-i)}\right]^{-1}\mathbf{P}_{(-i)}^T\mathbf{p}_i \\ \mathbf{0}^T & 1 \end{bmatrix} \tag{19}$$

Substitute (11) into (18) while applying $\det \mathbf{A} = 1$ to yield

$$\det[\mathbf{P}^T\mathbf{P}] = \det[\mathbf{P}_{(-i)}^T\mathbf{P}_{(-i)}][\mathbf{w}_i^T\mathbf{w}_i] \tag{20}$$

or

$$\log\frac{1}{\det[\mathbf{P}^T\mathbf{P}]} = \log\frac{1}{\det[\mathbf{P}_{(-i)}^T\mathbf{P}_{(-i)}]} + \log\frac{1}{\mathbf{w}_i^T\mathbf{w}_i} \tag{21}$$

Since the maximization of the determinant of the design matrix $[\mathbf{P}^T\mathbf{P}]$ is equivalent to the minimization of the left hand side (21), the reduction to left hand side of (21) by deleting p_i is given by $\log\frac{1}{\mathbf{w}_i^T\mathbf{w}_i}$. The reduction in a composite D-optimality cost function (Hong and Harris 2002) in backward elimination is given by

$$V_i^D = \beta\log\frac{1}{\mathbf{w}_i^T\mathbf{w}_i} - IEV_i \tag{22}$$

where β is a small positive parameter that controls the trade-off between D-optimality and model error variance. If $i_D = \arg[\max\{V_i^D, \forall i\} > 0]$, the backward elimination can be applied to reduce the composite D-optimality cost function by deleting the i_D^{th} regressor (Hong and Harris 2002). The process repeats until it is terminated when $\max\{V_i^D, \forall i\} < 0$.

In the above proposed algorithms, the orthogonalization scheme is different from that of (Hong and Harris 2001, Hong and Harris 2002). In forward orthogonal least squares (Chen et al. 1989, Chen et al. 1999), the regressors in the model are orthogonal to each other, but for backward elimination approaches introduced above, each model regressor is made orthogonal to a subspace spanned by the model after this regressor is removed ($\mathbf{P}_{(-i)}$) (for subspace orthogonal decomposition, see (Hong and Harris 2003)). The orthogonality for regressors in the model (Hong and Billings 1997) may gain some computation efficiency, but of course will make coding of the algorithm more complex.

3.2 Backward elimination by using basis pursuit

By replacing the 2-norm term in ridge regression that regulates parameter estimates vector with a 1-norm term, the derived parameter estimates is called basis pursuit parameter estimates (Chen et al. 2001). The 1-norm parameter regularization via basis pursuit is connected to linear programming (LP)(Chen et al. 2001), i.e. the problem can be configured as a standard form of constrained optimization problem. By using some parameter searching strategy for the basis pursuit based parameter cost function can lead to a sparse model (Chen et al. 2001), because the optimization of basis pursuit cost function will result in some parameters as zeros (Chen et al. 2001). A feasible strategy to sparsify the model is via backward elimination to remove regressors one at time by maximizing the reduction in the basis pursuit cost function, while maintaining feasible parameters based on least squares for the pruned model, until the model structure is sufficiently small. Consider the parameter vector $\mathbf{g} = [\mathbf{g}_{(-i)}^T, \theta_i]^T$ based on regression matrix $\mathbf{W}$. The basis pursuit cost function in this paper is defined by

$$J^B = \frac{1}{N}\Xi^T\Xi + \lambda|\mathbf{g}| \tag{23}$$

where $\lambda|\mathbf{g}| = \lambda\sum_{i=1}^{M}|g_i|$, g_i is the ith component of $\mathbf{g}$. λ is the basis pursuit parameter constant (Chen et al. 2001). Due to the orthogonality between $\mathbf{P}_{(-i)}$ and $\mathbf{w}_i$, substituting (6) into (23) yields

$$\begin{aligned} J^B &= \frac{1}{N}(\mathbf{y} - \mathbf{P}_{(-i)}\mathbf{g}_{(-i)} - \mathbf{w}_i\theta_i)^T(\mathbf{y} - \mathbf{P}_{(-i)}\mathbf{g}_{(-i)} - \mathbf{w}_i\theta_i) + \lambda|\mathbf{g}| \\ &= J_{(-i)}^B + \lambda|\theta_i| - \frac{1}{N}\theta_i^2\mathbf{w}_i^T\mathbf{w}_i \\ &= J_{(-i)}^B + \lambda|\theta_i| - \frac{1}{N}IEV_i \end{aligned} \tag{24}$$

where $J_{(-i)}^B = \frac{1}{N}\Xi_{(-i)}^T\Xi_{(-i)} + \lambda|\mathbf{g}_{(-i)}|$. $\Xi_{(-i)}$ is the model residual vector after deleting the ith regressor from $\mathbf{P}$. Therefore by deleting the ith regressor, the change in the cost function J^B is given by

$$V_i^B = |\theta_i|(\lambda - \frac{1}{N}|\theta_i|\mathbf{w}_i^T\mathbf{w}_i) \tag{25}$$

Substituting (7) into (25), followed by applying (5) and (8) to yield

$$\begin{aligned} V_i^B &= |\theta_i|(\lambda - \frac{1}{N}|\mathbf{w}_i^T\mathbf{y}|) \\ &= |\theta_i|(\lambda - \frac{1}{N}|\mathbf{p}_i^T(\mathbf{I} - \mathbf{P}_{(-i)}\left[\mathbf{P}_{(-i)}^T\mathbf{P}_{(-i)}\right]^{-1}\mathbf{P}_{(-i)}^T)\mathbf{y}|) \\ &= |\theta_i|(\lambda - \frac{1}{N}|\mathbf{p}_i^T(\mathbf{y} - \mathbf{P}_{(-i)}\mathbf{g}_{(-i)}|) \\ &= |\theta_i|(\lambda - \frac{1}{N}|\mathbf{p}_i^T\Xi_{(-i)}|) \end{aligned} \tag{26}$$

If $\lambda > \frac{1}{N}|\mathbf{p}_i^T\Xi_{(-i)}|$, this means that a deduction in the overall cost function can be achieved by deleting the ith regressor, and if $\lambda < \frac{1}{N}|\mathbf{p}_i^T\Xi_{(-i)}|$, $\forall i$, this means the overall cost function can not be reduced by deleting any of the regressors. If $i_B = \arg[\max\{V_i^B, \forall i\} > 0]$, the model can be pruned using backward elimination to reduce basis pursuit cost function by deleting the i_B^{th} regressor. The process will be automatically terminated when $\lambda < \min\{\frac{1}{N}|\mathbf{p}_i^T\Xi_{(-i)}|, \forall i\}$, which is controlled by the basis pursuit parameter as a threshold value.

Note that (26) is that it does not obviously contain the orthogonalised regressor $\mathbf{w}_i$, and the algorithm does not need an orthogonalisation. Generally for a model identified via least squares that is subject to backward elimination, θ_i is directly available as the current model parameters. This makes the coding of backward elimination procedure via basis pursuit extremely simple.

Note that as a postprocessing procedure designed to prune to a model with sufficient approximation capability, the range of the α, β and λ should be quite easily set, such that a model with sufficient approximation capability can be pruned without loosing too much on approximation capability. If these parameters are too small, then pruning processing won't start. If these are too large, the model MSE will increase too much during pruning. Experience has shown that a model with sufficient approximation capability can be pruned and terminated before a large deterioration in model fit for a range of α, β and λ, respectively, i.e. the pruning results are not particular sensitive to the these values in a large range (Hong and Harris 2001, Hong and Harris 2002).

4 Numerical examples

Here we consider two characteristic modelling problems a) data based prediction and b) nonlinear dynamic modelling to illustrate the efficacy of the three approaches.

Example 1. Consider the Benchmark diabetes data set (Efron et al. 2003). There are ten input variables of measurements of age, sex, body mass index, average blood pressure, and six blood serum measurements. The output variable is response of interest, a quantitative measure of disease progression one year after the input measurements were obtained. For comparison, the same normalization used in (Efron et al. 2003) was applied in this study (the variance of the output is 5943, and each of input vectors has zero mean and unit length). The three proposed backward elimination approaches were applied separately to a full linear regression model with ten input variables to predict the response, with $\alpha = 10000$, $\beta = 80000$ and $\lambda = 0.2$, to derive some comparable results (note that for this example α, β appear large due to different scales for inputs and output). In Figure 1, the model MSE were plotted versus the pruning steps. All of them converge to the same MSE after 6 regressors were removed, after following some different paths, as shown in Table 1. Each procedure can be automatically terminated via detecting the sign change in maximum values of V_i^A, V_i^D, V_i^B (from positive to negative). As indicated in Figure 1, the stopping point is typified prior to a sudden increase in MSE. After deleting 5 common regressors via separate backward elimination, the model is pruned as the same structure with a set of regressors $\{3,9,4,7,2\}$. The least squares estimates for this sparsified model gives $R^2 = 1 - \frac{\text{var}(\xi)}{\text{var}(y)} = 0.5086$.

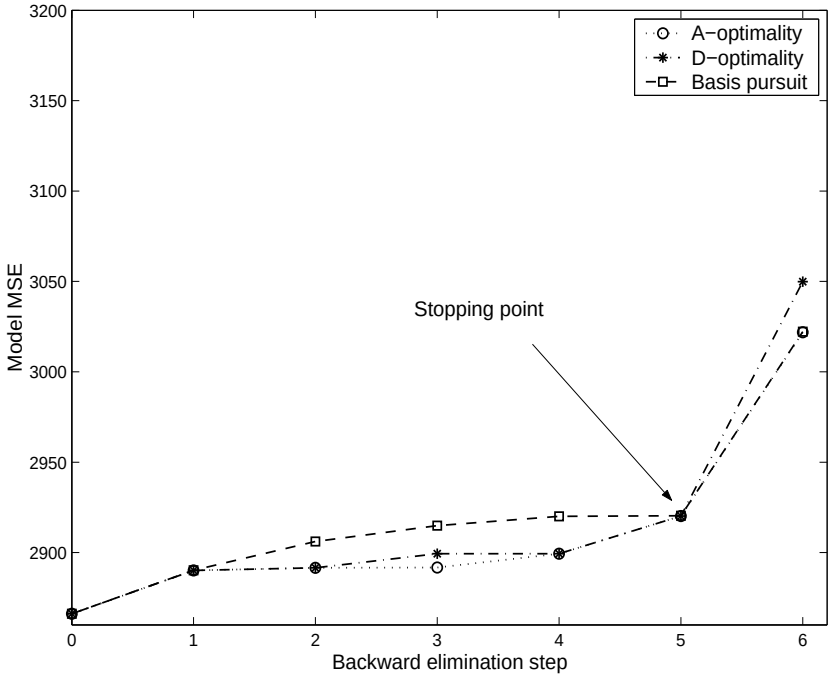

Figure 1. The evolution of MSE during pruning via A-/D- optimality and basis pursuit cost functions (Example 1).

Table 1. The removal order of regressors via A-/D- optimality and basis pursuit cost function (Example 1).

Method	**Backward elimination step**	**1**	**2**	**3**	**4**	**5**
A-optimality	Removed regressors via V_i^A	5	8	1	10	6
D-optimality	Removed regressors via V_i^D	5	8	10	1	6
Basis pursuit	Removed regressors via V_i^B	5	6	8	10	1

Example 2. Backward elimination to postprocess a model derived by D-optimality based forward orthogonal least squares algorithm (Hong and Harris 2002). Consider a nonlinear discrete dynamical system based on

the rational model

$$y(t) = \frac{y(t-1)y(t-2)[y(t-1)+y(t-2)u(t-1)-2.8]}{1+y^2(t-1)+y(t-1)y(t-2)+y^2(t-2)} + u(t-1) \tag{27}$$

where the system input $u(k)$ is a uniformly distributed random sequence in the range of $[-0.80.8]$. 500 system input–output data pairs $\{y(t), u(t)\}$ for $t = 1, ..., 500$ were generated. The RBF networks based on Gaussian function $p_k(\mathbf{x}(t)) = \phi(\mathbf{x}(t), \mathbf{c}_k) = \exp\{-\|\mathbf{x}(t)-\mathbf{c}_k\|^2/\tau^2\}$ are used to model the system based on the system input–output data set, with $n_y = 2$, $n_u = 1$. The width as set $\tau = 1$ for simplicity. Initially all 500 training data is used as the candidate centre set $\mathbf{c}_k$. The D-optimality algorithm (Hong and Harris 2002) was applied with $\alpha = 1 \times 10^{-5}$ to derive a model with 46 centres.

Given an initial model with 46 centres, which are indexed in the order of being selected in the initial forward regression, a normalisation procedure was applied initially to all the regressors so that they are all with zero mean, and unit variance, then three backward elimination approaches were applied separately with $\alpha = 1\times10^{-3}$, $\beta = 1\times10^{-2}$ and $\lambda = 5\times10^{-5}$, to derive models with comparable final sizes. In Figure 2, the model MSE were plotted versus the pruning steps. Each procedure can be automatically terminated via detecting the sign change in maximum values of V_i^A, V_i^D, V_i^B (from positive to negative). After following different paths, for A-/D-optimality based algorithms 5 regressors were removed, and for basis pursuit based algorithms 6 regressors were removed, as shown in Figure 2, where the stopping point is one step later. The index of these removed regressors and when they were removed are list in Table 1. Clearly there are some common regressors that are removed for different approaches. The pruned models have very similar MSE and sparsity. (For basis pursuit approach, the MSE is slightly worse). For 3 derived models gives $R^2 = 1 - \frac{\text{var}(\xi)}{\text{var}(y)} = 0.966$. The effects of these backward elimination pruning approaches to the model as in demonstrated in Figure 3 is that all these approaches are successful in gaining extra sparsity yet maintain sufficient approximation capabilities.

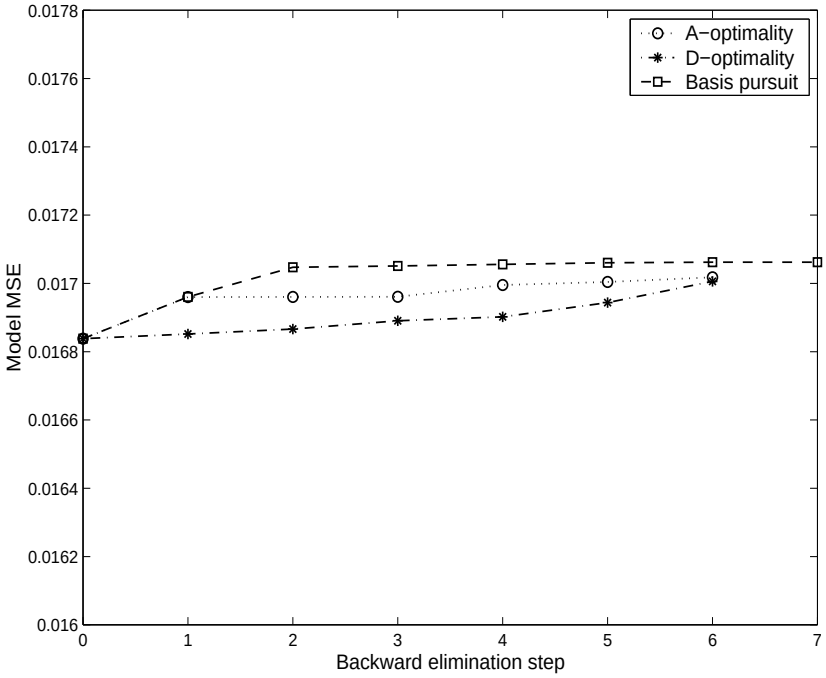

Figure 2. The evolution of MSE during pruning via A-/D- optimality and basis pursuit cost functions (Example 2).

Table 2. The removal order of regressors via A-/D- optimality and basis pursuit cost function (Example 2).

Method	**Backward elimination step**	**1**	**2**	**3**	**4**	**5**	**6**
A-optimality	Removed regressors via V_i^A	40	23	14	31	38	
D-optimality	Removed regressors via V_i^D	23	14	39	38	27	
Basis pursuit	Removed regressors via V_i^B	40	28	14	31	27	38

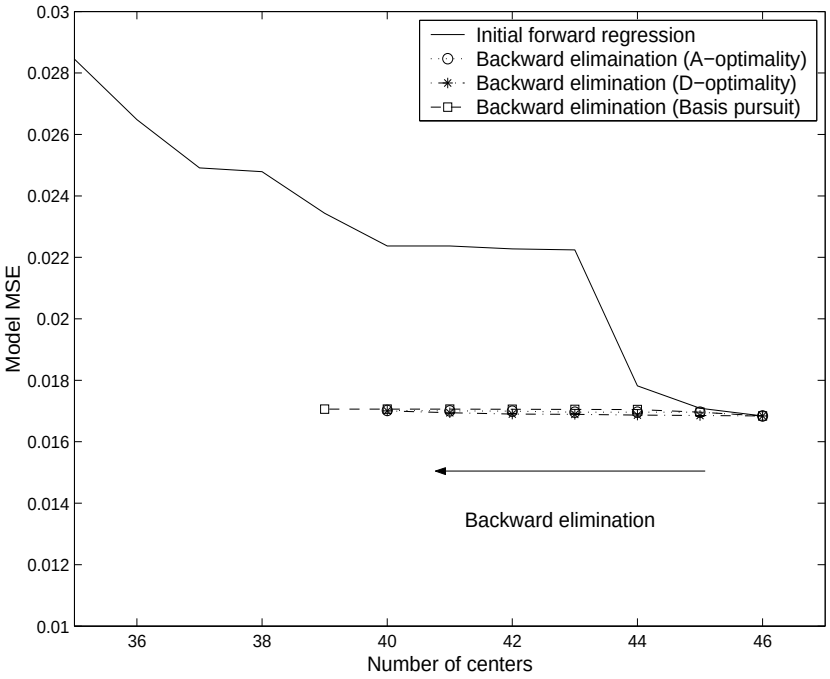

Figure 3. The effect of backward elimination approaches in approximation/sparsity (Example 2).

5 Conclusions

In this paper, several variants of the backward elimination approach have been introduced as an automatic postprocessing procedure that can be used in improving model sparsity in data based approaches. Composite or hybrid cost functions have been introduced to determine model pruning in a backward elimination manner, based on A-/D-optimality, and basis pursuit. These approaches take some similar "learning" patterns by derivation via a subspace orthogonalisation between each regressor and pruned model. It is shown that the newly introduced basis pursuit cost function based approach uses a simple formula without a need for any orthogonalisation. The model structural pruning process is important as a mechanism for improved generalization. For an identified model with linear-in-the-parameter structure and sufficient approximation capability, the proposed algorithms can be applied as an additional post processing procedure to improve model sparsity.

Acknowledgements

The authors gratefully acknowledge that part of this work was supported by EPSRC in the UK.

References

Atkinson, A.C. and Donev, A.N. (1992), *Optimum Experimental Designs*, Clarendon Press, Oxford.

Brown, M. and Harris, C.J. (1994), *Neurofuzzy Adaptive Modelling and Control*, Prentice Hall, Hemel Hempstead.

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

Chen, S., Wu, Y. and Luk, B.L. (1999), "Combined genetic algorithmoptimization and regularized orthogonal least squares learning for radial basis function networks," *IEEE Trans. on Neural Networks*, vol. 10, pp. 1239-1243.

Chen, S.S., Donoho, D.L. and Saunders, M.A. (2001), "Atomic decomposition by basis pursuit," *SIAM Review*, vol. 43, no. 1, pp. 129-159.

Efron, B., Johnstone, I., Hastie, T. and Tibshirani, R. (2003), "Least angle regression," *Annals of Statistics*, To Appear.

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

Hong, X. and Billings, S.A. (1997), "Givens rotation based fast backward elimination algorithm for RBF neural networks pruning," *IEE Proc. - Control Theory and Applications*, vol. 144, no. 5, pp. 381-384.

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

Hong, X. and Harris, C.J. (2002), "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.

Hong, X. and Harris, C.J. (2003), "A neurofuzzy network knowledge extraction and extended gram-schmidt algorithm for model subspace decomposition," *IEEE Transactions on Fuzzy Systems*, vol. 11, no. 4, pp.528-541.

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

Orr, M.J.L. (1993), "Regularisation in the selection of radial basis function centers," *Neural Computation*, vol. 7, no. 3, pp. 954-975.

Reed, R. (1993), "Pruning algorithm – a survey," *IEEE Transactions on Neural Networks*, vol. NN-4, pp. 740-747.

Vapnik, V. (1998), *Statistical Leaning Theory: Adaptive Learning Systems for Signal Processing, Communication and Control*, Wiley, Chichester.