An Elastic Net Orthogonal Forward Regression Algorithm

Xia Hong * **Sheng Chen** **

* *School of Systems Engineering, University of Reading, UK.*
** *School of Electronics and Computer Science, University of Southampton SO17 1BJ, UK. (also with the Faculty of Engineering, King Abdulaziz University, Jeddah 21589, Saudi Arabia)*

Abstract: In this paper we propose an efficient two-level model identification method for a large class of linear-in-the-parameters models from the observational data. A new elastic net orthogonal forward regression (ENOFR) algorithm is employed at the lower level to carry out simultaneous model selection and elastic net parameter estimation. The two regularization parameters in the elastic net are optimized using a particle swarm optimization (PSO) algorithm at the upper level by minimizing the leave one out (LOO) mean square error (LOOMSE). Illustrative examples are included to demonstrate the effectiveness of the new approaches.

Keywords: Data sets, Models, Neural networks, Regularization, System identification.

1. INTRODUCTION

A basic principle in practical nonlinear data modelling is the parsimonious principle that ensures the smallest possible model for the explanation of the observational data. A large class of nonlinear models and neural networks can be classified as linear models which include statistically linear or linear-in-the-parameters models. Regularization methods are developed to carry out parameter estimation and model structure selection simultaneously (Chen et al. [2003], Zou and Hastie [2005]). From Bayesian viewpoint, it has been shown that the parameter regularization using a penalty function on l^2 norms of the parameters is equivalent to a maximized *a posterior* probability (MAP) estimate of parameters by adopting a Gaussian prior for parameters. An iterative evidence procedure can be used for solving the optimal l^2 regularization parameters (MacKay [1991], Chen et al. [2003], Chen [2002]).

Alternatively the model sparsity can be achieved by minimizing the l^1 norm of the parameters. The l^1 norm minimization is fundamental to the basis pursuit or least absolute shrinkage and selection operator (LASSO) (Chen et al. [1998], Tibshirani [1996]). The least angle regression (LAR) procedure is developed for solving the problem efficiently, see Efron et al. [2004]. The Bayesian interpretation for LASSO is simply by adopting an Laplacian prior for parameters. The advantage of LASSO 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 small, but nonzero, values. Unfortunately introducing nondifferentiable l^1 norm in the cost function brings difficulties of model parameter estimation and finding an appropriate l^1 regularizer. Another disadvantage of using l^1 optimization is that a group of correlated terms cannot be selected together, which is not desirable for the sake of interpretability of the model in some applications. On the other hand, the use of l^2 will improve model generalization, but cannot be used for model selection by itself.

Recently a promising concept of the elastic net (EN) has been proposed by minimizing the l^1 and l^2 norms of the parameters together, see Zou and Hastie [2005]. The EN keeps the model sparsity of LASSO, while strongly correlated terms tend to be in or out of the model together. It is shown that the elastic net problem can be transformed into an equivalent LASSO problem on an augmented data, based on which the LAR procedure is applicable, referred to as LARS-EN, see Zou and Hastie [2005]. Note that because there are two regularization parameters in the elastic net, the cross validation has to be performed over a two-dimensional space. The ten fold cross validation was used in the choosing two regularization parameters by searching over a grid of l^2 norm regularization parameter values. Then for each setting of the l^2 norm regularization parameter, the algorithm LARS-EN produces the entire solution path of the elastic net, which is used to select l^1 norm regularization parameter by tenfold CV. Clearly this may not yield the optimal parameters if the grid search is set at a coarse level, but increasing the grid search at a very fine level would inevitably increase the computational cost. It would be desirable that the two regularisation parameters can be optimized simultaneously based on cross validation as well as in an efficient manner.

In this paper we propose an efficient model identification method aiming at maximizing a model's generalisation capability. A new elastic net cost function is defined and applied based on orthogonal decomposition, which facilitates the automatic model structure selection process with no need of using a predetermined error tolerance to terminate the forward selection process. The analytical evaluation of LOOMSE was presented based on the resultant ENOFR models without actually splitting the data set. Consequently a fully automated procedure is achieved

⋆ This work was supported in part by the UK EPSRC

without resort to any other validation data set for iterative model evaluation. The algorithm has a two level structure. At the upper level, the two regularization parameters in the elastic net are optimized using PSO by minimizing the LOOMSE. At the lower level are the simultaneous model selection and elastic net parameter estimation. Illustrative examples are included to demonstrate the effectiveness of the new approaches.

2. PRELIMINARIES

Consider the general nonlinear system represented by the nonlinear model, see Chen and Billings [1989]:

$$y(k) = f(\mathbf{x}(k)) + e(k), \tag{1}$$

where $\mathbf{x}(k) \in \Re^m$ denotes the system input vector and $y(k)$ is the system output variable, respectively. $e(k)$ is the system white noise and $f(\bullet)$ is the unknown system mapping. The system model (1) is to be identified from an observation data set $D_N = \{\mathbf{x}(k), y(k)\}_{k=1}^N$ using some suitable functional which can approximate $f(\bullet)$ with arbitrary accuracy. One class of such functionals is the kernel regression model of the form:

$$y(k) = \hat{y}(k) + e(k) = \sum_{i=1}^{n_M} \theta_i \phi_i(\mathbf{x}(k)) + e(k), \tag{2}$$

where $\hat{y}(k)$ denotes the model output, θ_i are the model weights, $\phi_i(\mathbf{x}(k))$ are the regressors, and n_M is the total number of candidate regressors or model terms.

By letting $\boldsymbol{\phi}_i = [\phi_i(\mathbf{x}(1)) \cdots \phi_i(\mathbf{x}(N))]^T$, for $1 \le i \le n_M$, and defining

$$\mathbf{y} = \begin{bmatrix} y(1) \\ \vdots \\ y(N) \end{bmatrix}, \quad \boldsymbol{\Phi} = [\boldsymbol{\phi}_1 \cdots \boldsymbol{\phi}_{n_M}],$$

$$\boldsymbol{\theta} = \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_{n_M} \end{bmatrix}, \quad \mathbf{e} = \begin{bmatrix} e(1) \\ \vdots \\ e(N) \end{bmatrix}, \tag{3}$$

the regression model (2) can be written in the matrix form

$$\mathbf{y} = \boldsymbol{\Phi}\boldsymbol{\theta} + \mathbf{e}. \tag{4}$$

Let an orthogonal decomposition of the matrix $\boldsymbol{\Phi}$ be

$$\boldsymbol{\Phi} = \mathbf{W}\mathbf{A}, \tag{5}$$

where

$$\mathbf{A} = \begin{bmatrix} 1 & a_{1,2} & \cdots & a_{1,n_M} \\ 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & a_{n_M-1,n_M} \\ 0 & \cdots & 0 & 1 \end{bmatrix} \tag{6}$$

and

$$\mathbf{W} = [\mathbf{w}_1 \cdots \mathbf{w}_{n_M}] \tag{7}$$

with columns satisfying $\mathbf{w}_i^T \mathbf{w}_j = 0$, if $i \neq j$. The regression model (4) can alternatively be expressed as

$$\mathbf{y} = \mathbf{W}\mathbf{g} + \mathbf{e}, \tag{8}$$

where the orthogonal weight vector $\mathbf{g} = [g_1 \cdots g_{n_M}]^T$ satisfy the triangular system $\mathbf{A}\boldsymbol{\theta} = \mathbf{g}$, which can be used to determine model parameters $\boldsymbol{\theta}$, given $\mathbf{A}$ and $\mathbf{g}$.

3. AUTOMATIC KERNEL REGRESSION MODEL CONSTRUCTION ALGORITHM USING ENOFR ASSISTED BY PSO

3.1 Elastic net orthogonal forward regression

For any fixed positive λ_1 and λ_2, the naive elastic net (NEN) criterion is defined as Zou and Hastie [2005]

$$L(\lambda_1, \lambda_2, \boldsymbol{\theta}) = \|\mathbf{y} - \boldsymbol{\Phi}\boldsymbol{\theta}\|^2 + \lambda_2 \|\boldsymbol{\theta}\|^2 + \lambda_1 \|\boldsymbol{\theta}\|_1 \tag{9}$$

where $\| \bullet \|$ denotes Euclidean norm, and $\|\boldsymbol{\theta}\|_1 = \sum_{i=1}^{n_M} |\theta_i|$. The naive elastic net estimator is the minimizer of

$$\hat{\boldsymbol{\theta}}_{NEN} = \arg\min_{\boldsymbol{\theta}} \{L(\lambda_1, \lambda_2, \boldsymbol{\theta})\} \tag{10}$$

This can be transformed into an equivalent LASSO problem on an augmented data, based on which the LAR procedure is applicable, referred to as LARS-EN in Zou and Hastie [2005]. The EN has some desirable properties, as it maintains the model sparsity of LASSO, but not as aggressive as LASSO in excluding correlated terms in the model. This is because these terms tend to be in or out of the model together as a result of the l^2 norm regularization, see Zou and Hastie [2005]. Note that there is no analytical solution to (10) unless the model terms are orthogonal.

The key to the proposed concept of ENOFR is to consider the following orthogonal elastic net (NEN) criterion based on (8)

$$L_e(\lambda_1, \lambda_2, \mathbf{g}) = \|\mathbf{y} - \mathbf{W}\mathbf{g}\|^2 + \lambda_2 \|\mathbf{g}\|^2 + \lambda_1 \|\mathbf{g}\|_1 \tag{11}$$

The naive elastic net solution for $\mathbf{g}$ is obtained by setting the subderivatives $\frac{\partial L_e}{\partial \mathbf{g}} = \mathbf{0}$, that is,

$$\mathbf{W}^T\mathbf{y} - \frac{\lambda_1}{2}\text{sign}(\mathbf{g}) = \left(\mathbf{W}^T\mathbf{W} + \lambda_2 \mathbf{I}\right)\mathbf{g}\ . \tag{12}$$

where $\mathbf{I}$ is an identity matrix of appropriate dimension and $\text{sign}(\mathbf{g}) = [\text{sign}(g_1), ..., \text{sign}(g_{n_M})]^T$, where

$$\text{sign}(s) \begin{cases} = & 1 \quad \text{if} \quad s > 0 \\ = & -1 \quad \text{if} \quad s < 0 \\ \in & [-1,\ 1] \quad \text{if} \quad s = 0 \end{cases} \tag{13}$$

Multiplying $2\mathbf{g}^T$ to both sides of (12) yields

$$2\mathbf{g}^T\mathbf{W}^T\mathbf{y} - \lambda_1\|\mathbf{g}\|_1 = 2\mathbf{g}^T\left(\mathbf{W}^T\mathbf{W} + \lambda_2\mathbf{I}\right)\mathbf{g}\ . \tag{14}$$

Substitute (14) into (11) to yield

$$\begin{aligned} L_e(\lambda_1, \lambda_2, \mathbf{g}) &= \mathbf{y}^T\mathbf{y} - 2\mathbf{g}^T\mathbf{W}^T\mathbf{y} + \mathbf{g}^T\mathbf{W}^T\mathbf{W}\mathbf{g} \\ &\quad + \lambda_2\|\mathbf{g}\|^2 + \lambda_1\|\mathbf{g}\|_1 \\ &= \mathbf{y}^T\mathbf{y} - \mathbf{g}^T\mathbf{W}^T\mathbf{W}\mathbf{g} - \lambda_2\|\mathbf{g}\|^2 \end{aligned} \tag{15}$$

Normalizing by $\mathbf{y}^T\mathbf{y}$,

$$\begin{aligned} & L_e(\lambda_1, \lambda_2, \mathbf{g})/(\mathbf{y}^T\mathbf{y}) \\ &= 1 - \sum_{i=1}^{n_M} \left(\mathbf{w}_i^T\mathbf{w}_i + \lambda_2\right)(g_i^{(NEN)})^2/(\mathbf{y}^T\mathbf{y}). \end{aligned} \tag{16}$$

where the superscript $^{(NEN)}$ denotes the naive elastic net solution. The elastic net error reduction ratio is defined by

$$[\text{eNerr}]_i = \frac{\left(\mathbf{w}_i^T\mathbf{w}_i + \lambda_2\right)(g_i^{(NEN)})^2}{(\mathbf{y}^T\mathbf{y})}, \quad i = 1, \cdots, n_M \tag{17}$$

where $g_i^{(NEN)}$, $i = 1, ...n_M$ are the solution of (12), given by

$$g_i^{(NEN)} = \left(\frac{\mathbf{w}_i^T\mathbf{w}_i}{\mathbf{w}_i^T\mathbf{w}_i + \lambda_2} |g_i^{(LS)}| - \frac{\lambda_1/2}{\mathbf{w}_i^T\mathbf{w}_i + \lambda_2} \right)_+ \text{sign}(g_i^{(LS)}) \quad (18)$$

with $g_i^{(LS)} = \frac{\mathbf{w}_i^T\mathbf{y}}{\mathbf{w}_i^T\mathbf{w}_i}$ and

$$z_+ = \begin{cases} z \text{ if } & z > 0 \\ 0 \text{ if } & z \leq 0 \end{cases} \quad (19)$$

Based on this ratio, significant regressors can be selected in a forward regression procedure. The automatic model term selection property of naive elastic net is explained as follows. Note that for $\lambda_1 = 0$, $[\text{eNerr}]_i$ becomes the model term selective criterion, the regularized error reduction ratio $[\text{rerr}]_i$, as defined in Chen et al. [1996]. In order to produce a sparse model containing n_s ($\ll n_M$) significant regressors, a chosen tolerance ξ ($0 < \xi < 1$) needs to be preset, and the selection process is terminated at the n_s-th stage when

$$1 - \sum_{l=1}^{n_s} [\text{rerr}]_l < \xi \quad (20)$$

is satisfied Chen et al. [1996]. However using elastic net orthogonal forward regression ($\lambda_1 > 0$), there is no need of setting ξ. This is because the cost function contains sparsity inducing l^1 norm so that some parameters will be zeros and $[\text{eNerr}]_i$ can return exact zero values during the selection process. The model selection is terminated at the $(n_s + 1)$-th stage when $[eNerr]_{n_s+1} = 0$, producing a sparse model containing n_s ($\ll n_M$) significant regressors automatically. The naive elastic net orthogonal forward regression (ENOFR) algorithm based on the modified Gram-Schmidt (MGS) scheme for a given $\boldsymbol{\lambda} = [\lambda_1, \lambda_2]^T$ can be implemented by modifying the algorithm of Chen et al. [1996], Chen and Wigger [1995].

Finally the elastic net (EN) parameter estimate is defined by

$$g_i^{(EN)} = \left(|g_i^{(LS)}| - \frac{\lambda_1/2}{\mathbf{w}_i^T\mathbf{w}_i} \right)_+ \text{sign}(g_i^{(LS)}) \quad (21)$$

which produces the elastic net parameter estimates for the n_s term model selected using the algorithm of Appendix A. This step inflates $g_i^{(NEN)}$ by the original shrinkage amount $\frac{\mathbf{w}_i^T\mathbf{w}_i + \lambda_2}{\mathbf{w}_i^T\mathbf{w}_i}$ and aims to overcome the double shrinkage problem of naive elastic net estimator, see Zou and Hastie [2005]. This means that the effect of l^2 norm regularization to parameter estimation is undone by this step, which is helpful to reduce bias in the naive elastic net estimator which could be too large.

We point out that as this rescaling step happens after the model terms selection so the existence of λ_2 has an impact on model structure compared with the case of $\lambda_2 = 0$. For example, because the model term selective criterion (17) is dependent on λ_2, so for a three term candidate set of $\{\boldsymbol{\phi}_1, \boldsymbol{\phi}_3, \boldsymbol{\phi}_3\}$, a two term model could be composed by $\boldsymbol{\phi}_1$ and $\boldsymbol{\phi}_2$ for $\lambda_2 = 0$, but by $\boldsymbol{\phi}_1$ and $\boldsymbol{\phi}_3$ for $\lambda_2 \neq 0$. The effect of l^2 norm regularization in selecting groups (correlated terms) was analyzed Zou and Hastie [2005]. For our proposed algorithm, the analysis to the first two regression steps can be extended to any regression steps. As a result of combined effect of λ_1 and λ_2, the explained output variance by selected regressors at earlier regression steps are reduced in comparison with a model using the least square parameter $g_i^{(LS)}$. Effectively this would allow the model output to be further explained by other regressors, that are correlated to previously selected regressors, to enter the model at later stages. Therefore the proposed algorithm has a similar effect to the original elastic net, of keeping correlated terms in the model, which is advantageous in that less variable models could be produced to provide physical insights on the causal relationships of the systems from large data sets, see Zou and Hastie [2005].

3.2 Choosing regularization parameters by optimizing the LOOMSE using PSO

Cross validation criteria are metrics that measure a model's generalisation capability. To optimize the model generalization capability, the model selection criteria are often based on cross-validation, see Stone [1974], Ljung [1987]. Due to its simplicity, a popular version of cross-validation is the so called leave one out (LOO) cross validation. If $f(\bullet)$ is modelled using linear models via least square method, there is an elegant way to generate LOOMSE , without actually sequentially splitting the estimation data set by using the Sherman-Morrison-Woodbury theorem, see Myers [1990]. In the following we show that LOOMSE based on the proposed ENOFR estimator can also be evaluated efficiently without actually sequentially splitting the estimation data set.

From (12) and (21), the elastic net parameter estimator based on a specified $\boldsymbol{\lambda}$ using N data points can be represented by

$$\mathbf{g}^{(EN)} = \mathbf{H}^{-1} \left(\mathbf{W}^T\mathbf{y} - \frac{\lambda_1}{2} \text{sign}(\mathbf{g}^{(EN)}) \right) \quad (22)$$

where $\mathbf{H} = \mathbf{W}^T\mathbf{W}$. The model residual is

$$\begin{aligned} e(k) &= y(k) - (\mathbf{g}^{(EN)})^T\mathbf{w}(k) \\ &= y(k) - \left(\mathbf{y}^T\mathbf{W} - \frac{\lambda_1}{2} [\text{sign}(\mathbf{g}^{(EN)})]^T \right) \mathbf{H}^{-1}\mathbf{w}(k) \end{aligned} \quad (23)$$

If the data sample indexed at k is removed from estimation data set, the leave one out elastic net parameter estimator obtained by using only $(N - 1)$ data points is given by

$$\begin{aligned} \mathbf{g}^{(EN,-k)} &= [\mathbf{H}^{(-k)}]^{-1} \\ &\times \left([\mathbf{W}^{(-k)}]^T\mathbf{y}^{(-k)} - \frac{\lambda_1}{2} \text{sign}(\mathbf{g}^{(EN,-k)}) \right) \end{aligned} \quad (24)$$

in which $\mathbf{H}^{(-k)} = [\mathbf{W}^{(-k)}]^T\mathbf{W}^{(-k)}$, $\mathbf{W}^{(-k)}$ and $\mathbf{y}^{(-k)}$ denote the resultant regression matrix and output vector respectively. The leave one out error evaluated at k is given by

$$\begin{aligned} & e^{(-k)}(k) \\ &= y(k) - [\mathbf{g}^{(EN,-k)}]^T\mathbf{w}(k) \end{aligned}$$

$$= y(k) - \left([\mathbf{y}^{(-k)}]^T \mathbf{W}^{(-k)} - \frac{\lambda_1}{2}[\mathrm{sign}(\mathbf{g}^{(EN,-k)})]^T \right) \times [\mathbf{H}^{(-k)}]^{-1}\mathbf{w}(k) \qquad (25)$$

It can be shown that

$$\mathbf{H}^{(-k)} = \mathbf{H} - \mathbf{w}(k)\mathbf{w}^T(k) \qquad (26)$$

$$[\mathbf{y}^{(-k)}]^T \mathbf{W}^{(-k)} = \mathbf{y}^T\mathbf{W} - y(k)\mathbf{w}^T(k) \qquad (27)$$

Applying the matrix inversion lemma to (26), yields

$$[\mathbf{H}^{(-k)}]^{-1} = [\mathbf{H} - \mathbf{w}(k)\mathbf{w}^T(k)]^{-1} = \mathbf{H}^{-1} + \frac{\mathbf{H}^{-1}\mathbf{w}(k)\mathbf{w}^T(k)\mathbf{H}^{-1}}{1 - \mathbf{w}^T(k)\mathbf{H}^{-1}\mathbf{w}(k)} \qquad (28)$$

and

$$[\mathbf{H}^{(-k)}]^{-1}\mathbf{w}(k) = \frac{\mathbf{H}^{-1}\mathbf{w}(k)}{1 - \mathbf{w}^T(k)\mathbf{H}^{-1}\mathbf{w}(k)} \qquad (29)$$

Substituting (27) and (29) into (25), yields

$$\begin{aligned} e^{(-k)}(k) &= y(k) - \left(\mathbf{y}^T\mathbf{W} - y(k)\mathbf{w}^T(k) - \frac{\lambda_1}{2}[\mathrm{sign}(\mathbf{g}^{(EN,-k)})]^T \right) \times \frac{\mathbf{H}^{-1}\mathbf{w}(k)}{1 - \mathbf{w}^T(k)\mathbf{H}^{-1}\mathbf{w}(k)} \\ &= \frac{y(k) - (\mathbf{y}^T\mathbf{W} - \frac{\lambda_1}{2}[\mathrm{sign}(\mathbf{g}^{(EN,-k)})]^T)\mathbf{H}^{-1}\mathbf{w}(k)}{1 - \mathbf{w}^T(k)\mathbf{H}^{-1}\mathbf{w}(k)} \end{aligned} \qquad (30)$$

The leave one out mean square error (LOOMSE) can be calculated as

$$J(\boldsymbol{\lambda}) = \frac{1}{N}\sum_{k=1}^{N}[e^{(-k)}(k)]^2 \qquad (31)$$

$$\approx \frac{1}{N}\sum_{k=1}^{N}[\frac{e(k)}{1 - \mathbf{w}^T(k)\mathbf{H}^{-1}\mathbf{w}(k)}]^2$$

$$= \frac{1}{N}\sum_{k=1}^{N}[\frac{e(k)}{1 - \sum_{i=1}^{n_s}[w_i(k)]^2/(\mathbf{w}_i^T\mathbf{w}_i)}]^2 \qquad (32)$$

by making use of (23) and assuming that $\mathrm{sign}(\mathbf{g}^{(EN,-k)}) = \mathrm{sign}(\mathbf{g}^{(EN)})$ holds for most k. This assumption is mild because only one data sample is removed at a time, based on significant regressors selected in a forward regression manner.

It is simple to evaluate $J(\boldsymbol{\lambda})$ as a result of the following reasons.

- Firstly the proposed elastic net cost function is based on parameter regularization within an orthogonal space, making it possible to derive an closed form expression for the parameters of the elastic net.
- Secondly we provide the above original derivation to show that the LOOMSE based on models using elastic net estimator can be analytically approximately evaluated without actually splitting the data by making use of the matrix inversion lemma and a mild assumption.
- Thirdly as the byproduct of the orthogonalization procedure $\mathbf{H}$ is diagonal, so that the evaluation of $e^{(-k)}(k)$ does not involve any matrix inversion and has a very small computational cost (see (30)).

The PSO constitutes a population based stochastic optimisation technique, which was inspired by the social behaviour of bird flocks or fish schools, see Kennedy and Eberhart [1995, 2001]. The algorithm commences with random initialisation of a swarm of individuals, referred to as particles, within the specific problem's search space. It then endeavours to find a globally optimum solution by gradually adjusting the trajectory of each particle towards its own best location and towards the best position of the entire swarm at each optimisation step. The PSO method is popular owing to its simplicity in implementation, ability to rapidly converge to a "reasonably good" solution and to "steer clear" of local minima. It has been successfully applied to wide-ranging optimisation problems, see van der Merwe and Engelbrecht [2003], Ratnaweera et al. [2004], Omran [2005], Guru et al. [2005], Soo et al. [2007]. We apply the PSO algorithm to find the minimum of (31). The complete algorithm can be illustrated with reference to the schematic diagram of Figure 1. The algorithm has a two layer structure. The upper level is the PSO with population size of S. It learns the two optimal regularization parameters based on the LOOMSE values provided by the lower level of S particles. At the lower level, each particle performs the ENOFN algorithm over the iterations, with each iteration consisting of two stages; (i) select a subset model based on the naive elastic net parameter estimator using the MGS algorithm in Appendix A; and (ii) determine the elastic net model parameters for the selected model terms using (21) and then calculate the associated LOOMSE using (30) & (31).

The computation cost of the PSO is dominated by that of the cost function evaluation. Let I_{max} denote the total number of iterations in PSO. The total computational complexity of the proposed two-level learning scheme is determined by the total number of function evaluations of PSO ($S \times I_{max}$), multiplying the average computation cost of each particle, i.e, that of the elastic forward regression. The latter is in the order of $O(N)$, which is further scaled by the product of candidate and final model size $n_s \times n_M$. Note that n_M can be set much lower than N if the latter is too large in order to save computation cost. The computational cost of the proposed algorithm is much smaller than conventional cross validation approaches of grid search over a two-dimensional space. For example if the ten-fold cross validation is used for a very coarse grid search of 3 by 3 on $\boldsymbol{\lambda}$, its computation cost is roughly the same as the proposed algorithm with $S = 9$ and $I_{max} = 9$ which is found to be appropriate from our experience. However the grid search of 3 by 3 on $\boldsymbol{\lambda}$ is likely to be too coarse to produce reasonably solutions.

4. AN ILLUSTRATIVE EXAMPLE

Consider using a RBF network to approximate an unknown scalar function

$$f(x) = \frac{\sin(x)}{x} \qquad (33)$$

A data set of two hundred points was generated from $y = f(x) + \xi$, where the input x was uniformly distributed

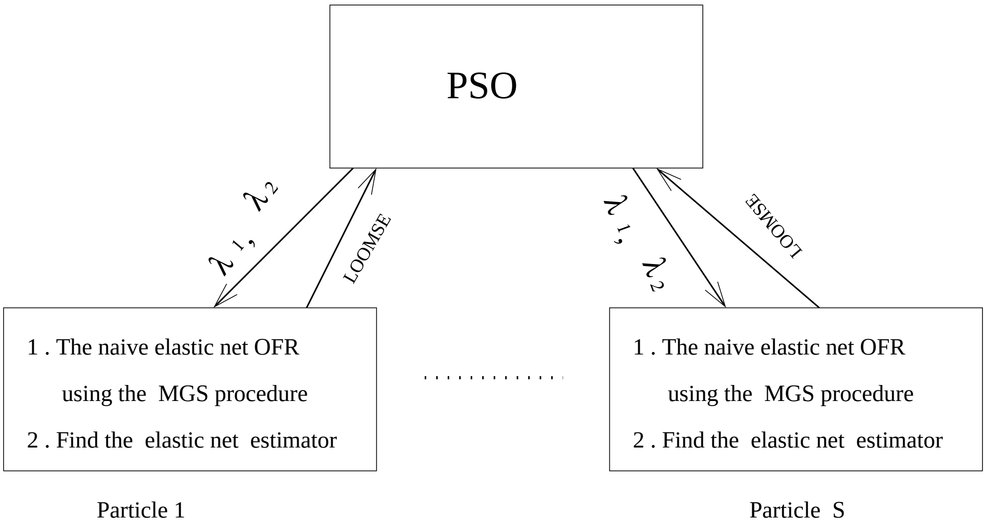

PSO
λ_1, λ_2
LOOMSE
λ_1, λ_2
LOOMSE
1 . The naive elastic net OFR using the MGS procedure
2 . Find the elastic net estimator
. . . .
1 . The naive elastic net OFR using the MGS procedure
2 . Find the elastic net estimator
Particle 1
Particle S

Fig. 1. A schematic diagram of the proposed ENOFR using PSO.

in [-10,10] and the noise ξ was Gaussian with zero mean and standard deviation 0.2. The data were very noisy. The Gaussian function

$$\phi_i(x) = \exp(-\frac{(x-c_i)^2}{2\tau^2}) \tag{34}$$

was used as the basis function to construct a RBF model, with a kernel width $\tau^2 = 10$. All the two hundred data points were used as the candidate RBF centre set for c_i. The search space of PSO were set $[10^{-7}, 0.1]$ for λ_1, and $[10^{-7}, 1]$ for λ_2. $S = 5$, $I_{max} = 5$ were predetermined. The proposed algorithm automatically selects a final model with only 7 terms, produced by regularization parameters $\lambda_1 = 0.0465$, $\lambda_2 = 0.145$. These were automatically determined by the PSO based on the LOOMSE criterion without using another validation data set. Figure 2(a) depicts $[eNerr]_j$ values against the forward regression process, which automatically terminated at the $8th$ step when $[eNerr]_8 = 0$. Figure 2 (b) depicts the model prediction of the resultant 7-term model in comparison to the noisy data used for training and the unknown true function. The resultant 7-term model produces a mean square error of 0.0015 with respect to the true function, illustrating the excellent model generalization capability of the model in this particular problem.

For comparison we construct models using ENOFR algorithm introduced in the paper, except that for selecting $\boldsymbol{\lambda}$ ten fold cross validation was used, rather than LOOMSE with PSO. By setting a grid of $\lambda_1 = [10^{-7}, 10^{-5}, 10^{-4}, 10^{-3}, 0.1]$ and λ_2=$[10^{-7}, 10^{-5}, 10^{-3}, 0.1, 1]$, 25 settings of $\boldsymbol{\lambda}$ are evaluated using ten fold cross validation. We used the same kernel width $\tau^2 = 10$, and for each fold all resultant 180 training data points were used as the candidate RBF centre set. The estimated computational cost is roughly nine times of using LOOMSE with PSO in terms of how many times the MGS algorithm is applied. We also assume that, due to the reduction of 10% in training data set size for ten fold cross validation, there is also 10% computational cost reduction. The best $\boldsymbol{\lambda}$ is found to be $\lambda_1 = 0.1$, $\lambda_2 = 0.001$. For each fold, a 7-term model was produced. With respect to the true function, the resultant mean square error for all data points over ten models is 0.0023 ± 0.0003 (mean ± standard deviation), illustrating that selecting $\boldsymbol{\lambda}$ using ten fold cross validation does not offer superior performance to the proposed algorithm for this particular problem.

5. CONCLUSIONS

Aiming at maximizing a model's generalisation capability, this paper has proposed an efficient two-level model identification method for the linear-in-the-parameters models. At the lower level is the proposed ENOFR algorithm that is able to perform simultaneous model selection and elastic net parameter estimation for a given pair of regularization parameters. At the upper level these regularization parameters are optimized using a particle swarm optimization (PSO) algorithm by minimizing the leave one out (LOO) mean square error (LOOMSE). As a result a fully automated procedure is achieved without resort to any other validation data set for iterative model evaluation. An illustrative example is included to demonstrate the effectiveness of the proposed algorithm.

REFERENCES

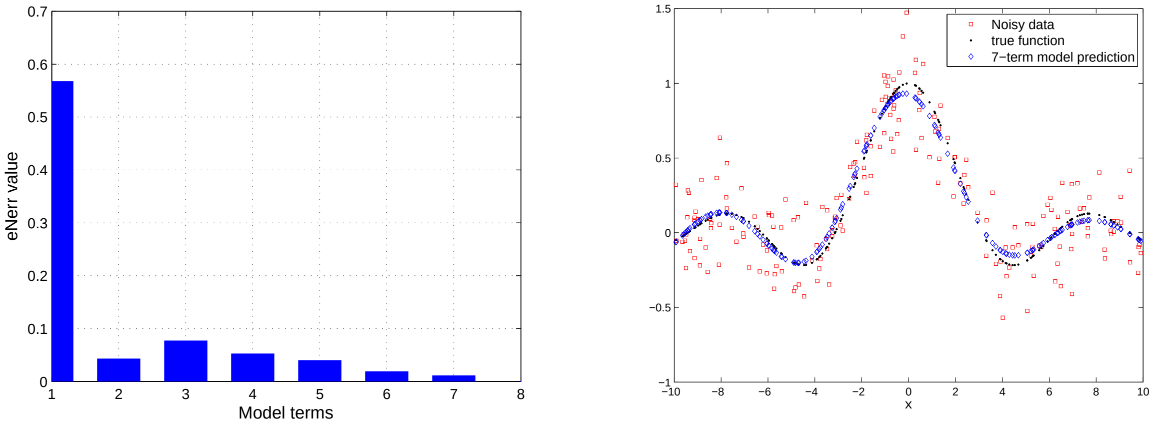
S. Chen. Locally regularised orthogonal least squares algorithm for the construction of sparse kernel regression models. In *Proceedings of 6th Int. Cof. Signal Processing*, pages 1229–1232, Beijing, China, 2002.

S. Chen and S. A. Billings. Representation of nonlinear systems: The NARMAX model,. *International Journal of Control*, 49(3):1013–1032, 1989.

S. Chen and J. Wigger. Fast orthogonal least squares algorithm for efficient subset selection. *IEEE Transactions on Signal Processing*, 43(7):1713–1715, 1995.

S. Chen, E. S. Chng, and K. Alkadhim. Regularized orthogonal least squares algorithm for constructing radial basis function networks. *International Journal of Control*, 64(5):829–837, 1996.

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*, 48(6):1029–1036, 2003.



(a) (b)

Fig. 2. The modeling results of the simple scalar function problem by the selected model ($\lambda_1 = 0.0465$, $\lambda_2 = 0.145$);(a) 7 nonzero $eNerr_j$ values during the elastic net orthogonal forward regression steps; and (b) model predictions of the 7-term model.

S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61, 1998.

B. Efron, I. Johnstone, T. Hastie, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32:407–451, 2004.

S. M. Guru, S. K. Halgamuge, and S. Fernando. Particle swarm optimisers for cluster formation in wireless sensor networks. In *Proc. 2005 Int. Conf. Intelligent Sensors, Sensor Networks and Information Processing*, pages 319–324, Melbourne, Australia, Dec. 5-8, 2005.

J. Kennedy and R. Eberhart. Particle swarm optimization. In *Proc. of 1995 IEEE Int. Conf. Neural Networks*, volume 4, pages 1942–1948, Perth, Australia, Nov. 27-Dec. 1, 1995.

J. Kennedy and R. C. Eberhart. *Swarm Intelligence*. Morgan Kaufmann, 2001.

L. Ljung. *System Identification: Theory for the User*. New Jersey: Prentice Hall, 1987.

D. J. C. MacKay. *Bayesian Methods for Adaptive Models*. PhD thesis, California Institute of Technology, USA, 1991.

R. H. Myers. *Classical and modern regression with applications*. PWS-KENT, Boston, 2nd edn., 1990.

M. G. H. Omran. *Particle Swarm Optimization Methods for Pattern Recognition and Image Processing*. PhD thesis, University of Pretoria, Pretoria, South Africa, 2005.

A. Ratnaweera, S. K. Halgamuge, and H. C. Watson. Self-organizing hierarchical particle swarm optimizer with time-varying acceleration coefficients. *IEEE Trans. Evolutionary Computation*, 8:240–255, June 2004.

K. K. Soo, Y. M. Siu, W. S. Chan, L. Yang, and R. S. Chen. Particle-swarm-optimization-based multiuser detector for cdma communications. *IEEE Trans. Vehicular Technology*, 56:3006–3013, September 2007.

M. Stone. Cross validatory choie and assessment of statistical predictions. *Journal of the Royal Statistical Society, Series B*, 36:117–147, 1974.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of Royal Statistical Society. Series B*, 58 (1):267–288, 1996.

D. W. van der Merwe and A. P. Engelbrecht. Data clustering using particle swarm optimization. In *Proc. CEC 2003*, pages 215–220, Cabberra, Australia, Dec. 8-12, 2003.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *J. R. Stasti. Soc. B*, 67(2): 301–320, 2005.