Monograph:
Hong, X. and Billings, S.P. (1997) Parameter Estimation Based on Stacked Regression and Evolutionary Algorithms. Research Report. ACSE Research Report 692 . Department of Automatic Control and Systems Engineering

---

Parameter Estimation Based on Stacked Regression and Evolutionary Algorithms

X. Hong and S. A. Billings
Department of Automatic Control and Systems Engineering,
University of Sheffield, Mappin Street, Sheffield S1 3JD

Abstract — A new parameter estimation algorithm which minimises the cross-validated prediction error for linear-in-the-parameter models is proposed based on stacked regression and an evolutionary algorithm. It is initially shown that cross-validation is very important for prediction in linear-in-the-parameter models using a criterion called the mean dispersion error (*MDE*). Stacked regression, which can be regarded as a sophisticated type of cross-validation, is then introduced based on an evolutionary algorithm to produce a new parameter estimation algorithm which preserves the parsimony of a concise model structure that is determined using the forward orthogonal least squares (OLS) algorithm. The PRESS prediction errors are used for cross-validation, and the sunspot and Canadian lynx time series are used to demonstrate the new algorithms.

Keywords — linear-in-the-parameters models, cross-validation, mean dispersion error, evolutionary algorithms

Research Report No. 692

Oct 1997

1 Introduction

Two important aspects of system identification are the determination of the structure of the model and the estimation of the model parameters. Least squares parameter estimation has been widely used for linear-in-the-parameters models. However in certain circumstances the models estimated using the least squares method can be prone to overfitting which can lead to a deterioration in the generalisation performance of the model. Fundamental to the evaluation of the generalisation properties of a model is a procedure called cross-validation (Stone, 1974). Model selection criteria can be derived based on information theoretic principles (Ljung and Glad, 1994) and these can be used as measures of the generalisation property of the resulting model on future data. This can be used as the criteria in subset selection such as the forward orthogonal least squares (OLS) and backward elimination methods (Billings, *et al*, 1988, Hong and Billings, 1997).

A criterion called the mean dispersion error (MDE) for parameter estimates and the MDE-superior concept were introduced by Rao and Toutenburg (1995). The MDE is a distance measurement between the parameter estimates and the true values. If the least squares parameter estimates have very large variance, they can be much further away from the true value than some biased estimates in spite of the unbiasedness. It may be shown that regularisation, which is also known as ridge regression estimation in statistics, and some other biased parameter estimates including the Stein-Rule shrinkage estimator are MDE-superior compared to the unbiased least squares parameter estimates (Rao and Toutenburg, 1995, Vinod and Ullah, 1981). Consequently a combination of regularisation and forward orthogonal least squares (OLS) methods have been introduced to improve model generalisation performance with RBF networks (M. J. L. Orr, 1995, S. Chen, *et al*, 1996).

Not only has cross-validation been directly used in the identification of dynamic system structures (Myers, R. H., 1990, Wang and Cluett, 1996) but this has also been indirectly applied in parameter estimation to select the regularization parameter (M. J. L Orr, 1995). A more sophisticated version of cross-validation, which is called stacked generalization, is an idea put forward by Wolpert (1992). The aim of stacking is to get a generalization accuracy as high as possible. One application is stacked regression which is a method of constructing linear combinations of different predictors under some constraints in order to give improved prediction accuracy (L. Breiman, 1996). Although stacked regression, which can be regarded as a sophisticated type of cross-validation, will also generally provide better predictions, the resulting predictor does not necessarily result in a parsimonious model structure which is often desired in system identification.

The present work focuses on the parameter estimation problem in the framework of linear-in-the-parameters time series models. Initially it is theoretically shown that a model with MDE-superior parameter estimates will have smaller prediction error if this is used in prediction. A new parameter estimation algorithm is then proposed based on stacked regression and an evolutionary algorithm. Different predictors consisting of subsets of a concise model structure which is determined using a forward orthogonal least squares (OLS) algorithm, are stacked through the optimal combination coefficients with the objective of minimising the cross-validated prediction error which is easily computed using the PRESS prediction errors without actually sequentially splitting the estimation data set (Myers, 1990). An evolutionary algorithm is formulated to find the optimal combination coefficients under some specific constraints which were introduced and studied by Breiman (1996). The new parameter estimates for the concise model structure selected using the forward OLS procedure are defined by combining the least squares parameter estimates of each predictor through optimal combination coefficients. In this mode model parsimony is achieved. The whole procedure constitutes a novel method for the parameter estimation of time series predictors. The effectiveness of the new approach is demonstrated with numerical examples.

2 Mean Dispersion Error and Prediction Error

A criterion called the mean dispersion error (MDE) for parameter estimates and the MDE-superior concept were introduced by Rao and Toutenburg (1995). The MDE is a distance measurement between the parameter estimates and the true values. If the least squares parameter estimates tend to have a very large variance they may be much further away from the true value than some biased estimates in spite of the unbiasedness. By introducing a small bias MDE-superior estimates can be achieved.

Denote the mean dispersion error (MDE) of $\hat{\Theta}_b$, a biased estimate of the parameter vector Θ as

$$\begin{aligned} \mathrm{MDE}(\hat{\Theta}_b, \Theta) &= E\left[(\hat{\Theta}_b - \Theta)(\hat{\Theta}_b - \Theta)^T\right] \\ &= E\left[[(\hat{\Theta}_b - E(\hat{\Theta}_b)) + (E(\hat{\Theta}_b) - \Theta)][(\hat{\Theta}_b - E(\hat{\Theta}_b)) + (E(\hat{\Theta}_b) - \Theta)]^T\right] \\ &= V(\hat{\Theta}_b) + [\mathrm{Bias}(\hat{\Theta}_b, \Theta)][\mathrm{Bias}(\hat{\Theta}_b, \Theta)]^T \end{aligned} \tag{1}$$

where $E(\bullet)$ denotes expectation, and $V(\hat{\Theta}_b)$ denotes the covariance matrix of $\hat{\Theta}_b$

$$V(\hat{\Theta}_b) = E\left[[\hat{\Theta}_b - E(\hat{\Theta}_b)][\hat{\Theta}_b - E(\hat{\Theta}_b)]^T\right] \tag{2}$$

and

$$\text{Bias}(\hat{\Theta}_b, \Theta) = E(\hat{\Theta}_b) - \Theta \tag{3}$$

is the bias of $\hat{\Theta}_b$. Let $\hat{\Theta}_{bj}, j = 1, 2$ be two estimators of Θ, then $\hat{\Theta}_{b2}$ is called MDE-superior to $\hat{\Theta}_{b1}$ if the difference in the MDE matrix is nonnegative definite (Rao and Toutenburg, 1995), that is, if

$$\triangle(\hat{\Theta}_{b1}, \hat{\Theta}_{b2}) = \text{MDE}(\hat{\Theta}_{b1}, \Theta) - \text{MDE}(\hat{\Theta}_{b2}, \Theta) \geq 0 \tag{4}$$

The MDE measures the distance between the parameter estimates and the real parameters. Heuristically, if a set of parameter estimates that are closer to the real parameters are used in prediction, then better generalisation properties should be expected. In the following it is shown through theoretical analysis that a model with MDE-superior parameter estimates will have a smaller prediction error if this model is used in prediction.

Assume that a given time series is modeled using a linear-in-the-parameters model with a correct and parsimonious model structure with M regressors

$$\begin{aligned} y(t) &= \sum_{i=1}^{M} p_i(t)\theta_i + \xi(t) \\ &= \mathbf{p}(t)^T\Theta + \xi(t) \end{aligned} \tag{5}$$

where $\xi(t)$ is white noise sequence with the variance $\sigma^2 = E[\xi(t)^2]$. $\Theta^T = [\theta_1, \cdots, \theta_M]$ is the parameter vector, and $\mathbf{p}(t)^T = [p_1(t), p_2(t), \cdots, p_M(t)]$ is the regression vector, in which, $p_i(t)$ are regressors, or basis functions, which are some fixed non-linear functions of lagged outputs $y(t)$, so that

$$p_i(t) = p_i(\mathbf{x}(t)) \tag{6}$$

where

$$\mathbf{x}(t) = [y(t-1), \cdots, y(t-n_y)] \tag{7}$$

The model structure can be determined using the forward orthogonal least squares (OLS) procedure. Eq.(5) can also be written in a matrix form as

$$\mathbf{y} = \mathbf{P}\Theta + \Xi \tag{8}$$

where $\mathbf{y}^T = [y(1), \cdots, y(N)]$ is the output vector, $\Xi^T = [\xi(1), \cdots, \xi(N)]$ is the system noise vector, and

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

is known as the regression matrix.

Usually, an estimate $\hat{\Theta}_0 \in \Re^M$ of Θ can be found using a least squares algorithm

$$\hat{\Theta}_0 = (\mathbf{P}^T\mathbf{P})^{-1}\mathbf{P}^T\mathbf{y} \tag{9}$$

Consider two models with the parameter estimates $\hat{\Theta}_{bj}, j = 1,2$ that are used for prediction. The one-step ahead predictions are

$$\hat{y}_{bj}(t|t-1) = \mathbf{p}(t)^T\hat{\Theta}_{bj}, \quad j = 1,2 \tag{10}$$

respectively. Taking into account that the noise $\xi(t)$ is uncorrelated to all the regressors, the variance of the one-step ahead prediction errors, from Eq.(5) and Eq.(10), are

$$E\left[[y(t) - \hat{y}_{bj}(t|t-1)]^2\right] = E\left[(\hat{\Theta}_{bj} - \Theta)^T\mathbf{p}(t)\mathbf{p}(t)^T(\hat{\Theta}_{bj} - \Theta)\right] + \sigma^2, \quad j = 1,2 \tag{11}$$

Theorem 1: (C. M. Theobald, 1974) Define $m_j = E\left[(\hat{\Theta}_{bj} - \Theta)^T\mathbf{D}(\hat{\Theta}_{bj} - \Theta)\right]$. The following conditions are equivalent:

$$(a) \qquad \triangle(\hat{\Theta}_{b1}, \hat{\Theta}_{b2}) = \mathrm{MDE}(\hat{\Theta}_{b1}, \Theta) - \mathrm{MDE}(\hat{\Theta}_{b2}, \Theta) \geq 0$$

$$(b) \qquad m_1 - m_2 \geq 0$$

for all non-negative matrices $\mathbf{D}$.

The proof of Theorem 1, which was originally introduced in a pure mathematical context, can be found in Theobald (1974). In the following it will be shown that useful insights can be obtained using Theorem 1 by setting the non-negative matrix $\mathbf{p}(t)\mathbf{p}(t)^T$ as the matrix $\mathbf{D}$ in Eq.(11). This yields

$$E\left[[y(t) - \hat{y}_{b1}(t|t-1)]^2\right] - E\left[[y(t) - \hat{y}_{b2}(t|t-1)]^2\right] \geq 0 \tag{12}$$

if and only if

$$\triangle\left(\hat{\Theta}_{b1}, \hat{\Theta}_{b2}\right) = \mathrm{MDE}(\hat{\Theta}_{b1}, \Theta) - \mathrm{MDE}(\hat{\Theta}_{b2}, \Theta) \geq 0 \tag{13}$$

This means that the variance of the prediction error is smaller if the parameter estimates are MDE-superior.

MDE is a useful criterion for selecting biased parameters when the least squares estimates tend to have a very large variance. The above analysis illustrates that to achieve smaller prediction errors the model should use MDE-superior parameter estimates and this implies that cross-

validation can be used in the estimation of parameters with MDE-superior properties. A simple version of cross-validation is to split the data set into an estimation set, which is used to estimate the parameters, and a testing set, which is used to judge the predictive capability of the model. The procedure using the PRESS statistic for cross-validation can be used in the estimation in this way (Myers, 1990). Use the estimation data set and sequentially set aside each data point in turn, estimate a model using the remaining data, and then evaluate the prediction error using only the data point that was removed. Finally the average of all these prediction errors is computed. The prediction error thus computed employs data that were not used in the identification and will be interpreted as the prediction error that the model is expected to have when applied in prediction over new data. In the following section a new parameter estimation method is introduced which is based on stacked regression and an evolutionary algorithm with the criterion to minimise the PRESS statistic in the estimation data set.

3 A New Parameter Estimation Algorithm using Stacked Regression and an Evolutionary Algorithm

3.1 Stacked Regression

Consider a non-linear stationary time series which is described in terms of some nonlinear expansion of lagged outputs,

$$y(t) = f(y(t-1), y(t-2), \cdots, y(t-n_y)) + \xi(t) \tag{14}$$

where $t = 1, 2, 3, \cdots, N$, and N is the sample size of the estimation set, n_y is the maximum lag of the system, $\{\xi(t)\}$ is an additive white noise, and $f(\bullet)$ is some nonlinear function. A set of predictors can be identified based on a variety of model structures and it is a common practice that only one model is selected according to model selection criterion. The model selection critera are often based on cross-validation which is recommended as a means to determine model structure (Ljung, 1987). Justification for cross-validation was provided in Section 2 where it was shown that a model with smaller prediction errors can produce MDE-superior parameter estimates. There is an elegant way to generate the prediction error known as the PRESS statistic (Myers, 1990). The idea is to sequentially set aside each data point in the estimation data set in turn, estimate a model using the remaining data, and then evaluate the prediction error using only the data point that was removed. Finally the average of all these prediction errors is computed. If $f(\bullet)$ is modelled using linear-in-the-paramters models, it is advantageous that the PRESS errors can be computed without actually sequentially splitting the estimation data set

by using the Sherman-Morrison-Woodbury theorem (Myers, 1990). To select the best predictor from K candidate predictors $\hat{y}_k^{(-t)}(t|t-1), k = 1, \cdots, K$, the same modelling procedure is used to minimise

$$\sum_{t=1}^{N} \{y(t) - \hat{y}_k^{(-t)}(t|t-1)\}^2$$

where $\hat{y}_k^{(-t)}(t|t-1)$ is identified by leaving out the learning sample at time t and using the remaining $N-1$ data samples for the K candidate predictors. Assume the nth predictor

$$n = \arg\{\min\{\sum_{t=1}^{N} [y(t) - \hat{y}_k^{(-t)}(t|t-1)]^2, \forall k\}\} \tag{15}$$

is selected.

An alternative approach is stacked regression which is a linear combination of a set of predictors. The predictors are stacked through a set of combination coefficients. The idea of stacking was shown to be a more sophisticated version of cross-validation in Wolpert (1992). Following the idea of Wolpert, Leo Breiman (1996) proposed the idea of least squares delete-1 cross-validated prediction errors under the constraints that the combination coefficients are non-negative and sum to one. The justification for this latter constraint is rather mathematical and is described in detail in Breiman (1995, 1996), where it has been shown that such a non-negative constraint is comparable to ridge regression (1995).

According to Leo Breiman (1996), instead of selecting one predictor from the predictors as in Eq.(15), in stacked regression a linear combination of K predictors $\hat{y}_k^{(-t)}(t|t-1), k = 1, \cdots, K$ is formed as a new predictor

$$\hat{y}^{(-t)}(t|t-1) = \sum_{k=1}^{K} \beta_k \hat{y}_k^{(-t)}(t|t-1) \tag{16}$$

where the combination coefficients $\beta_k, k = 1, \cdots, K$ are obtained by minimising

$$J = \sum_{t=1}^{N} \{y(t) - \hat{y}^{(-t)}(t|t-1)\}^2 \tag{17}$$

subject to the constraints $\beta_k \geq 0, k = 1, \cdots, K$, and $\sum_{k=1}^{K} \beta_k = 1$. Usually, a better prediction accuracy than any single predictor $\hat{y}_k^{(-t)}(t|t-1), k = 1, \cdots, K$ can be achieved. The procedure can be shown to find the best 'interpolating' predictor (Breiman, 1996).

3.2 New Parameter Estimates

In system identification it is often desirable that the model should be parsimonious. Although stacked regression will generally provide better predictions, the predictor which is obtained may result in a complex structure. It has been shown that the forward orthogonal least squares (OLS) procedure can be used to select a concise model structure (Billings, *et al*, 1988). In the present work, the subsets of the selected model structure, determined using the forward OLS algorithm constitute different predictors. In this mode, model parsimony is achieved because the final predictor is still based on the selected models which are parsimonious. Assume that the selected model consisting of M regressors is described by Eq.(5). The subset number of the selected model structure can be large and to reduce computations only $K = M + 1$ predictors are used to be stacked, including the model Eq.(5) itself and M models which are obtained by sequentially removing each of the M regressors in turn from this model. These predictors are formed and stacked following the determination of the model structure using the forward OLS algorithm. The new parameter estimates for the model structure are defined by combining the least squares parameter estimates of each predictor through the optimal combination coefficients. An evolutionary algorithm is then formulated with the objective of minimising the cross-validated prediction error to find the optimal combination coefficients under the specific constraint, which was introduced and studied by Breiman (1996), that the combination coefficients are nonnegative and sum up to one. The PRESS errors are used as the cross-validated prediction errors in the estimation data set and are easily computed using the Sherman-Morrison-Woodbury theorem (Myers, 1990).

Based on the model Eq.(5), a set of M predictors which have been obtained by deleting the kth regressor from the model Eq.(5) are stacked with the original model Eq.(5). Denote initially the original model as the first predictor $\hat{y}_0^{(-t)}(t|t-1)$. The kth regressor is then deleted in turn from the M regressors, $k = 1, \cdots, M$ to form the other M predictors $\hat{y}_k^{(-t)}(t|t-1), k = 1, \cdots, M$ using the remaining $(M-1)$ regressors. Consequently in Eq.(16) $\sum_{k=1}^{K}$ is replaced by $\sum_{k=0}^{M}$. In the computation of the cost function J in Eq.(17) the PRESS errors are used. Denote the PRESS errors by

$$\xi_0^{(-t)}(t) = y(t) - \hat{y}_0^{(-t)}(t|t-1) \tag{18}$$

$$\xi_k^{(-t)}(t) = y(t) - \hat{y}_k^{(-t)}(t|t-1), \quad k = 1, 2, \cdots, M \tag{19}$$

where $t = 1, 2, \cdots, N$.

Substituting Eq.(18), Eq.(19) and Eq.(16) (with $\sum_{k=1}^{K}$ replaced by $\sum_{k=0}^{M}$) into Eq.(17) and

applying the constraints of $\sum_{k=0}^{M} \beta_k = 1$ yields

$$J = \sum_{t=1}^{N} \{ \sum_{k=0}^{M} \beta_k \xi_k^{(-t)}(t) \}^2 \tag{20}$$

The computation of the PRESS error is an elegant way to generate the delete-1 cross-validated prediction errors which can be regarded as the prediction error the model is expected to have when applied to new data. If the $\hat{y}_k^{(-t)}(t|t-1)$'s are modelled using linear-in-the-paramters models the PRESS errors can be computed without actually sequentially splitting the estimation data set by using the Sherman-Morrison-Woodbury theorem (Myers, 1990). The procedure consists initially of computing the least squares parameter estimates for each predictor. The first predictor is the model Eq.(5) with the regression matrix $\mathbf{P} \in \Re^{N \times M}$. For each of M subset predictors that contain $(M-1)$ regressors, due to the deletion of the kth regressor $(k = 1, \cdots, M)$ from M regressors of the model Eq.(5), the regression matrix $\mathbf{P}_k \in \Re^{N \times (M-1)}, k = 1, \cdots, M$ is formed by deleting the kth regressor from $\mathbf{P}$.

The regression matrix is denoted as

$$\mathbf{P}_k = [\mathbf{p}_k(1), \mathbf{p}_k(2), \cdots, \mathbf{p}_k(N)]^T \tag{21}$$

where

$$\begin{aligned}
\mathbf{p}_1(t) &= [p_2(t), p_3(t), \cdots, p_{M-1}(t), p_M(t)]^T \\
\mathbf{p}_k(t) &= [p_1(t), \cdots, p_{k-1}(t), p_{k+1}(t), \cdots, p_M(t)]^T, k = 2, \cdots, M-1 \\
\mathbf{p}_M(t) &= [p_1(t), p_2(t), \cdots, p_{M-2}(t), p_{M-1}(t)]^T
\end{aligned}$$

are obtained by sequentially removing one regressor in turn from $\mathbf{p}(t)$. The least squares parameter estimates $\hat{\Theta}'_k = [\hat{\theta}'_{k,1}, \cdots, \hat{\theta}'_{k,k-1}, \hat{\theta}'_{k,k}, \cdots, \hat{\theta}'_{k,M-1}]^T \in \Re^{M-1}, k = 1, \cdots, M$ are computed from

$$\hat{\Theta}'_k = (\mathbf{P}_k^T \mathbf{P}_k)^{-1} \mathbf{P}_k^T \mathbf{y} \tag{22}$$

Note that the dash $'$ was used to represent the reduced dimension parameter vector, and

$$\hat{y}_0^{(-t)}(t|t-1) = \mathbf{p}(t)^T \hat{\Theta}_0^{(-t)} \tag{23}$$

$$\hat{y}_k^{(-t)}(t|t-1) = \mathbf{p}_k(t)^T \hat{\Theta}_k^{'(-t)}, \quad k = 1, 2, \cdots, M \tag{24}$$

where $t = 1, 2, \cdots, N$. $\hat{\Theta}_0^{(-t)} \in \Re^M$ are estimated using least squares by sequentially leaving out $y(t)$ and $\mathbf{p}(t)^T$, the data sample at t, in turn and using only the remaining data. $\hat{\Theta}_k^{'(-t)} \in \Re^{M-1}$ are estimated using least squares by sequentially leaving out $y(t)$ and $\mathbf{p}_k(t)^T$ in turn and using

only the remaining data.

From Eq.'s(18),(19),(23) and (24) and the Sherman-Morrison-Woodbury theorem the PRESS errors $\xi_0^{(-t)}(t)$ and $\xi_k^{(-t)}(t)$ can be calculated using (Myers, R. H.,1990)

$$\begin{aligned} \xi_0^{(-t)}(t) &= y(t) - \hat{y}_0^{(-t)}(t|t-1) \\ &= y(t) - \mathbf{p}(t)^T \hat{\Theta}_0^{(-t)} \\ &= \frac{\xi_0(t)}{1 - \mathbf{p}(t)^T [\mathbf{P}^T \mathbf{P}]^{-1} \mathbf{p}(t)} \end{aligned} \tag{25}$$

and

$$\begin{aligned} \xi_k^{(-t)}(t) &= y(t) - \hat{y}_k^{(-t)}(t|t-1) \\ &= y(t) - \mathbf{p}_k(t)^T \hat{\Theta}_k^{\prime(-t)} \\ &= \frac{\xi_k(t)}{1 - \mathbf{p}_k(t)^T [\mathbf{P}_k^T \mathbf{P}_k]^{-1} \mathbf{p}_k(t)} \end{aligned} \tag{26}$$

where

$$\xi_0(t) = y(t) - \mathbf{p}(t)^T \hat{\Theta}_0 \tag{27}$$

$$\xi_k(t) = y(t) - \mathbf{p}_k(t)^T \hat{\Theta}'_k, \quad k = 1, 2, \cdots, M \tag{28}$$

The predictors $\hat{y}_k^{(-t)}(t|t-1), k = 0, 1, \cdots, M$ can then be stacked to form a new predictor

$$\hat{y}^{(-t)}(t|t-1) = \sum_{k=0}^{M} \beta_k \hat{y}_k^{(-t)}(t|t-1) \tag{29}$$

The combination coefficients are obtained using an evolutionary algorithm through the optimisation of the cost function J in Eq.(20) under the constraints that $\beta_k \geq 0, k = 0, \cdots, M$, and $\sum_{k=0}^{M} \beta_k = 1$.

Finally new parameter estimates can be defined and computed by combining the least squares estimates of each predictor through the combination coefficients. As each predictor is a subset of the concise model structure selected using forward OLS procedure, the final predictor is still based on the selected model. A new set of parameter estimates $\hat{\Theta}_s \in \Re^M$ for the original model structure can now be formulated by stacking, where the subscript 's' denotes stacking. Initially introduce the auxiliary vectors $\hat{\Theta}_k, \hat{\Theta}_k^{(-t)}$, where $k = 1, 2, \cdots, M$, $t = 1, 2, \cdots, N$ and denote

$$\hat{\Theta}'_k = [\hat{\theta}'_{k,1}, \cdots, \hat{\theta}'_{k,k-1}, \hat{\theta}'_{k,k}, \cdots, \hat{\theta}'_{k,M-1}]^T \tag{30}$$

$$\hat{\Theta}_k^{\prime(-t)} = [\hat{\theta}_{k,1}^{\prime(-t)}, \cdots, \hat{\theta}_{k,k-1}^{\prime(-t)}, \hat{\theta}_{k,k}^{\prime(-t)}, \cdots, \hat{\theta}_{k,M-1}^{\prime(-t)}]^T \tag{31}$$

By inserting a zero term into the kth row of $\hat{\Theta}_k^{\prime}$ and $\hat{\Theta}_k^{\prime(-t)}$ respectively, define

$$\hat{\Theta}_k = [\hat{\theta}_{k,1}^{\prime}, \cdots, \hat{\theta}_{k,k-1}^{\prime}, 0, \hat{\theta}_{k,k}^{\prime}, \cdots, \hat{\theta}_{k,M-1}^{\prime}]^T \tag{32}$$

$$\hat{\Theta}_k^{(-t)} = [\hat{\theta}_{k,1}^{\prime(-t)}, \cdots, \hat{\theta}_{k,k-1}^{\prime(-t)}, 0, \hat{\theta}_{k,k}^{\prime(-t)}, \cdots, \hat{\theta}_{k,M-1}^{\prime(-t)}]^T \tag{33}$$

where $k = 1, 2, \cdots, M, t = 1, 2, \cdots, N$. Because $\mathbf{p}_k(t)$ was formed by deleting the kth row from $\mathbf{p}(t)$, using Eq.'s (24),(31),(33) yields

$$\hat{y}_k^{(-t)}(t|t-1) = \mathbf{p}_k(t)^T \hat{\Theta}_k^{\prime(-t)} = \mathbf{p}(t)^T \hat{\Theta}_k^{(-t)} \tag{34}$$

Substitute Eq.(23) and Eq.(34) into Eq.(29) to give

$$\hat{y}^{(-t)}(t|t-1) = \mathbf{p}(t)^T \sum_{k=0}^{M} \beta_k \hat{\Theta}_k^{(-t)} \tag{35}$$

and denote

$$\hat{\Theta}^{(-t)} = \sum_{k=0}^{M} \beta_k \hat{\Theta}_k^{(-t)} \tag{36}$$

By directly expanding Eq.(36) and omitting the superscript '$(-t)$', new parameter estimates for the stacked model can then be defined as

$$\hat{\Theta}_s = \sum_{k=0}^{M} \beta_k \hat{\Theta}_k \tag{37}$$

where $\beta_k \geq 0, k = 0, \cdots, M$, and $\sum_{k=0}^{M} \beta_k = 1$ are obtained using an evolutionary algorithm with the objective of minimising the cost function J in Eq.(20). It is important to note that the $\hat{y}_k^{(-t)}(t|t-1)$'s and $\hat{\Theta}_k^{(-t)}$'s were used to illustrate the idea but do not have to be explicitly computed in the algorithm which will be summarised and shown to be very simple in Section 3.4. These estimates cannot guarantee an optimal MDE in practice, but may lead to a suboptimal solution if the results lead to better generalisation properties. In the latter case a solution can be found that is at least MDE-superior. Experience shows that the parameter estimates thus defined will more often than not result in improved predictive performance over a test data set, especially when the estimation sample is small. Compared with other biased parameter estimates such as regularisation and the Stein-Rule shrinkage estimator, the new estimates require less *prior* subjective information because the process of estimating the regularisation parameter or the shrinkage parameter is avoided. Note that the new algorithm introduced above

is computationally intensive and this can be a disadvantage when the parameter estimates only result in a small amount of improvement in the prediction performance compared to the least squares estimates.

3.3 A New Evolutionary Algorithm for the Combination Coefficients

In recent years, there has been a growing interest in using genetic algorithms (GA) or evolutionary algorithms for parameter optimization problems (Z. Michalewicz, 1994). The aim of these algorithms applied to the parameter optimization problem is to search over a population of points to improve the solution through probabilistic transition rules and simple operations. Genetic algorithms and evolutionary algorithms are probabilistic algorithms which imitate the principles of natural evolution which maintains a population of individuals for a specific generation. Each individual represents a potential solution which is implemented as some data structure, called a chromosome representation. The population of individuals undergoes a sequence of transformations called a crossover and a mutation according to some probabilities to form new individuals. Each individual is evaluated by some measure, called a fitness function. A next generation is generated by selecting the fittest individuals and discarding the remainder. Evolutionary algorithms are genetic algorithms which have been modified to be more problem dependent so as to suit a specific requirement of the problem such as nontrival constraints. In this work, an evolutionary algorithm is applied to determine the combination coefficients $\beta_k, k = 0, \cdots, M$ so that the cost function J in Eq.(29) is minimised under the constraints that $\beta_k \geq 0, k = 0, \cdots, M$, and $\sum_{k=0}^{M} \beta_k = 1$. The evolutionary algorithm was modified from GENOCOP (Z. Michalewicz, 1994). The basic characteristic of the new evolutionary algorithm is that the constraints are applied upon the solution space where the chromosomes were generated. This means that in all operations, no chromosome that violates the constraints will be produced.

The procedure to compute the combination coefficients $\beta_k, k = 0, \cdots, M$ using the evolution algorithm are described below.

(i). Set the population size of the chromosome to N_{pop}. Set the probabilities of crossover and mutation as p_c and p_m respectively. Define each chromosome in the population as a vector $\mathbf{B}_i$ of length $(M+1)$, $\mathbf{B}_i = [\beta_{0i}, \beta_{1i}, \cdots, \beta_{Mi}]$, $i = 1, \cdots, N_{pop}$, with the constraints $\beta_{ki} \geq 0, k = 0, \cdots, M$, and $\sum_{k=0}^{M} \beta_{ki} = 1$. This means that the chromosomes are the combination parameters themselves.

(ii). N_{pop} chromosomes $\mathbf{B}_i, i = 1, 2, \cdots, N_{pop}$ for the initial population are generated as $(M+1) \times N_{pop}$ dimensional uniformly distributed numbers, and then these numbers are normalised

by setting $\mathbf{B}_i \leftarrow \mathbf{B}_i/\|\mathbf{B}_i\|$ so as to satisfy the constraints $\sum_{k=0}^{M} \beta_{ki} = 1$, where the $\| \bullet \|$ denotes the Euclidean norm.

(iii). A fitness function is defined as

$$fit(\mathbf{B}_i) = \frac{1}{J_i} \tag{38}$$

where J_i is the cost function J computed via Eq.(20) for the ith chromosome, $i = 1, 2, \cdots, N_{pop}$. Note that the cost function J is computed by sequentially using Eq's.(6), (22), (27), (28), (25) and (26).

(iv). The probability of reproduction of each chromosome is designated to be $fit(\mathbf{B}_i)/\sum_{i=1}^{N_{pop}} fit(\mathbf{B}_i)$. A new population of size N_{pop} is generated by selecting the chromosomes according to the probability of reproduction. In the new population, chromosomes are randomly selected to generate offspring via crossover and mutation operations with respective probabilities p_c and p_m. The crossover operation between two parent chromosomes $\mathbf{B}_i$ and $\mathbf{B}_j$ creates two offsprings $\mathbf{B}_i'$ and $\mathbf{B}_j'$, using two steps

$$(1) \begin{cases} \mathbf{B}_i' = \alpha \mathbf{B}_i + (1-\alpha)\mathbf{B}_j \\ \mathbf{B}_j' = (1-\alpha)\mathbf{B}_i + \alpha \mathbf{B}_j \end{cases}$$

$$(2) \begin{cases} \mathbf{B}_i' \leftarrow \mathbf{B}_i'/\|\mathbf{B}_i'\| \\ \mathbf{B}_j' \leftarrow \mathbf{B}_j'/\|\mathbf{B}_j'\| \end{cases}$$

where α is a (0,1) uniformly distributed number. It is clear that since α falls within $[0, 1]$, The resulting $\mathbf{B}_i'$ and $\mathbf{B}_j'$ after the first step will still be nonnegative as long as $\mathbf{B}_i$ and $\mathbf{B}_j$ are nonnegative. The constraint $\sum_{k=0}^{M} \beta_k = 1, for\ \mathbf{B}_i', \mathbf{B}_j'$ is realised via the second step.

The mutation operation on $\mathbf{B}_i$ is composed of three steps

$$(1)\ \ \mathbf{B}_i' = \mathbf{B}_i + N(0,1)$$
$$(2)\ \ \mathbf{B}_i' \leftarrow \mathbf{B}_i' - \text{round}(\mathbf{B}_i')$$
$$(3)\ \ \mathbf{B}_i' \leftarrow \mathbf{B}_i'/\|\mathbf{B}_i'\|$$

where $N(0, 1)$ is a length M vector of independent random Gaussian numbers with a mean of zero and standard deviation 1, and round($\bullet$) indicates rounding down to an integer. The first step is a basic stochastic mutation operation, but the resulting $\mathbf{B}_i'$ falls out of the constrained solution space. The resulting $\mathbf{B}_i'$ after the second step will be nonnegative. The constraint $\sum_{k=0}^{M} \beta_k = 1, for\ \mathbf{B}_i'$ is realised via the third step.

The crossover and mutation operations are defined such that the resulting offsprings still satisfy the constraints stated in (i).

(v). The new population after the crossover and the mutation plus the original population constitutes a whole population of size $2 \times N_{pop}$, which are evaluated according to the fitness function $fit(\bullet)$. Half of the weaker chromosomes are deleted from the whole population. Go to step (iii) until some stopping criterion is reached. In this procedure, we assume the algorithm converges after a suffcient number of generations.

After the fitness function $fit(\bullet)$ converges at a set of optimal combination coefficients, compute the parameter estimates $\hat{\Theta}_s$ using Eq.(37).

3.4 The Algorithms

The new parameter estimation algorithm based on stacked regression and the evolution algorithm are summarised in the following:

(i). Use the forward OLS algorithm to construct a concise model Eq.(5) with a predetermined number of regressors M. The model is used as the first predictor denoted as $\hat{y}_0^{(-t)}(t|t-1)$ in the stacked regression. Compute the least squares estimates $\hat{\Theta}_0$ using Eq.(9).

(ii). Based on the order of regressor positions in the model Eq.(5) delete the kth regressor, $k = 1, 2, \cdots, M$, to form one predictor $\hat{y}_k^{(-t)}(t|t-1)$ using the remaining $(M-1)$ regressors. M predictors are constructed in this way. Compute the least squares estimates $\hat{\Theta}'_k$, $k = 1, 2, \cdots, M$ using Eq.(22).

(iii). Form $\hat{\Theta}_k$, $k = 1, 2, \cdots, M$ using Eq.(32) by inserting a zero term into $\hat{\Theta}'_k$ in Eq.(30).

(iv). Use the evolutionary algorithm described in Section 3.3 to determine the combination coefficients β_k, $k = 0, 1, 2, \cdots, M$ for the $(M+1)$ predictors.

(v). The parameter estimates $\hat{\Theta}_s$ are computed using Eq.(37).

4 Numerical Examples

Example 1: Consider the annual sunspot time series for the years 1700-1955. The data is plotted in Fig.1. The first 221 observations were used to fit a NAR polynomial model, and the last 35 points were used as a test data set. Note that the data splitting technique employed here is for comparison and validation of the algorithms only and is obviously different from the sequential splitting of the estimation data set by deteting one data sample in turn.

The parsimonious model structure with 11 terms and $n_y = 9$ was used (Chen and Billings, 1989).

$$\begin{aligned} \mathbf{p}(t) &= [y(t-1), y(t-2), y(t-9), y^3(t-1), y(t-1)y(t-8), \\ &\quad y(t-2)y(t-5)y(t-8), y^2(t-9), y(t-1)y(t-5)y(t-8), \\ &\quad y(t-5)y^2(t-7), y(t-3), y(t-2)y(t-3)y(t-4)]^T \end{aligned} \tag{39}$$

The ordinary least squares method produced the model

$$\begin{aligned} \hat{\Theta}_0 &= [1.1171, -0.0920, 0.3830, -0.000028084, 0.0044, -0.00016203, \\ &\quad -0.0025, 0.000095842, 0.000011992, -0.2181, 0.0000095906]^T \end{aligned} \tag{40}$$

Each of the 11 regressors was in turn deleted from the model to form 11 distinctive predictors. Each of the 11 predictors consists of 10 regressors. The parameter estimates were identified by using ordinary least squares Eq.(22). The estimates are written in the form of auxiliary vectors, Eq.(32), where the zero terms are inserted at the k'th row in $\hat{\Theta}_k$ as

$$\begin{aligned} \hat{\Theta}_1 &= [0, 0.8597, 0.7420, -0.000015587, 0.0058, -0.00033601, \\ &\quad -0.0053, 0.00030596, 0.0000069030, -0.3052, -0.0000035588]^T \\ \hat{\Theta}_2 &= [1.0666, 0, 0.3945, -0.000028232, 0.0042, -0.00017343, \\ &\quad -0.0026, 0.00011101, 0.000012464, -0.2639, 0.0000095990]^T \\ \hat{\Theta}_3 &= [1.3266, -0.2186, 0, -0.000033644, 0.0059, -0.00012926, \\ &\quad -0.00056097, 0.000025866, 0.000017981, -0.1787, 0.000011780]^T \\ \hat{\Theta}_4 &= [0.9449, -0.1303, 0.5143, 0, -0.0011, -0.00020229, \\ &\quad -0.0029, 0.00018832, 0.000011219, -0.1148, 0.0000036443]^T \\ \hat{\Theta}_5 &= [1.1413, -0.0239, 0.4287, -0.000020858, 0, -0.00018349, \\ &\quad -0.0024, 0.00016753, 0.000011433, -0.2872, 0.0000077866]^T \\ \hat{\Theta}_6 &= [1.3307, -0.3555, 0.3141, -0.000031671, 0.0059, 0, \\ &\quad -0.0020, -0.000047341, 0.0000014209, -0.1619, 0.0000084848]^T \\ \hat{\Theta}_7 &= [1.2505, -0.1538, 0.1298, -0.000029495, 0.0042, -0.00014066, \\ &\quad 0, 0.000056175, 0.000015430, -0.2103, 0.000011218]^T \\ \hat{\Theta}_8 &= [1.2233, -0.2364, 0.3224, -0.000031475, 0.0065, -0.00010309, \\ &\quad -0.0021, 0, 0.000012355, -0.1784, 0.000012228]^T \\ \hat{\Theta}_9 &= [1.0946, -0.1313, 0.4284, -0.000027836, 0.0043, -0.00012388, \\ &\quad -0.0028, 0.000099023, 0, -0.1498, 0.0000033528]^T \end{aligned}$$

$$\begin{aligned}\hat{\Theta}_{10} &= [1.1377, -0.2963, 0.3670, -0.000026306, 0.0053, -0.00015117,\\ &\quad -0.0025, 0.000077174, 0.0000083314, 0, 0.0000024945]^T\\ \hat{\Theta}_{11} &= [1.0806, -0.0924, 0.3934, -0.000026885, 0.0042, -0.00015953,\\ &\quad -0.0026, 0.00011034, 0.0000080798, -0.1350, 0]^T\end{aligned}$$

The parsimonious model with all 11 regressors, together with the above 11 distinctive predictors were then stacked through a set of combination coefficients $\beta_k, k = 0, \cdots, 11$. The optimal combination coefficients $\mathbf{B} = [\beta_0, \beta_1, \cdots, \beta_{11}]$ were determined using the evolutionary algorithm in Section 3.3. The probability of crossover and mutation were set to be $p_c = 0.25$ and $p_m = 0.10$ respectively, the population size was $N_{pop} = 100$, and the number of generations was set to be 1000. The objective is to minimise J in Eq.(20). The evolution of the cost function J of the fittest chromosome for each generation is plotted in Fig.2. The combination coefficients converged to

$$\begin{aligned}\mathbf{B} &= [0.1272, 0.0705, 0.0378, 0.0454, 0.1911, 0.0618, 0.0248,\\ &\quad 0.1582, 0.0422, 0.0499, 0.1550, 0.0363]\end{aligned}$$

and the parameter estimates obtained from the stacked regression were computed through Eq.(37) as

$$\begin{aligned}\hat{\Theta}_s &= [1.1067, -0.0878, 0.3648, -0.000026264, 0.0040, -0.00015466,\\ &\quad -0.0022, 0.000093727, 0.000011119, -0.2063, 0.0000080371]^T \qquad (41)\end{aligned}$$

The two sets of parameters $\hat{\Theta}_0$ and $\hat{\Theta}_s$ were used to calculate the predictions over the estimation set and the test set. The MSE of the one-step ahead prediction errors in the estimation and test sets were computed, and the results are shown in Table.1. The result indicates that the original model based on the least squares (LS) estimates fits well in the estimation data set, but the prediction deteriorates when the model is used on a new data set. In the estimation set, the MSE of one-step ahead prediction errors based on the new estimates increased by 0.45% compared with that of the LS estimates, the change is not significant. It is seen that both parameter estimates provide a good fit to the estimation set. However, in the test data set, the MSE of the one-step ahead prediction errors based on the new estimates decreases by 3.33% compared with that of the LS estimates, which is a useful improvement in the accuracy of the predictions. The new estimates are therefore considered to be superior.

Example 2: Consider the Canadian lynx time series which represents the annual numbers of Canadian lynx trapped in the Mackenzie River district of North-Western Canada. This data was collected over a period of 114 years between 1821-1934 (Rao and Gabr, 1984). The original

	MSE(estimation set)	MSE(test set)
LS estimates $\hat{\Theta}_0$	121.54	177.79
The new estimates $\hat{\Theta}_s$	122.09	171.70

Table 1: A comparison of MSE of one-step ahead prediction error in Example 1 (annual sunspot numbers)

and log-transformed data are shown in Fig.3. The log-transformed data was used throughout. The first 80 observations were used to fit a NAR polynomial model, and the last 34 points were used as a test data set. A parsimonious model structure with 15 terms, $n_y = 9$ and polynomial degree $n_l = 3$ was used. The model structure was identified using the forward Orthogonal Least Squares (OLS) method. Then, each of the 15 regressors was in turn deleted and the parameters re-estimated to form 15 additional predictors each consisting of 14 regressors. The parameter estimates for these predictors were identified using ordinary least squares through Eq.(22). The original model with all 15 regressors, together with the 15 subset predictors consisting of 14 regressors were then stacked through a set of combination coefficients $\beta_k, k = 0, 1, \cdots, 15$ which were determined using the evolutionary algorithm. In the algorithm, the probability of crossover and mutation were set to be $p_c = 0.25$ and $p_m = 0.10$ respectively, the population size was $N_{pop} = 100$, and the number of generations was set to be 1000. The objective was to minimise J in Eq.(29). The evolution of the cost function J of the fittest chromosome for each generation is plotted in Fig.4. The combination coefficients converged to

$$\begin{aligned} \mathbf{B} \;=\; & [0.0806, 0.0400, 0.0929, 0.0193, 0.0064, 0.0521, 0.1330, 0.1191 \\ & 0.1170, 0.0193, 0.0195, 0.1277, 0.0608, 0.0233, 0.0741, 0.0148] \end{aligned}$$

Finally, the new parameter estimates were computed through Eq.(37). The model structure and the two sets of parameters $\hat{\Theta}_0$ and $\hat{\Theta}_s$ are shown in Table.2. The two models were then used to calculate predictions over the estimation and test sets. The MSE of the one-step ahead prediction errors are listed in Table.3. It is seen that both parameter estimates provide equally good fits to the estimation data, but the original model based on the LS estimates shows a slightly larger deterioration in prediction compared with the new estimates when used on a new or test data set. In the test data set, the MSE of the one-step ahead prediction errors based on the new estimates decreased by 2.38% compared with the LS estimates, which is an improvement on the accuracy of the prediction. The new estimates are therefore preferred. Note that for both sets of parameter estimates the MSE in the test data set is significantly different to that of the estimation data set. This may be caused by the length of the data set, which is very short, or

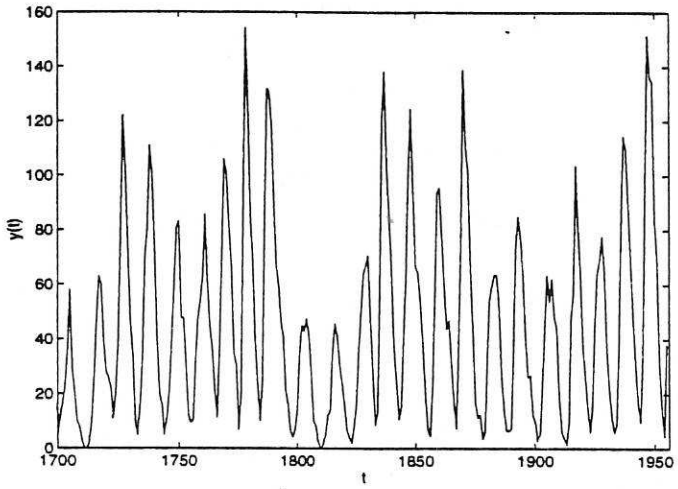

Figure 1: The annual sunspot time series.

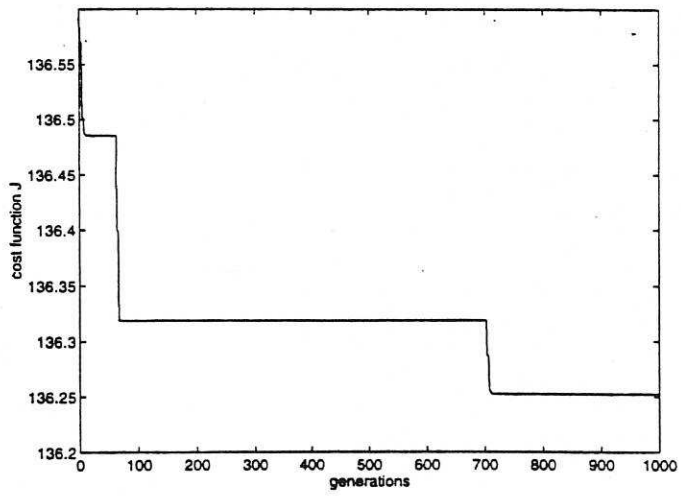

Figure 2: The evolution of the cost function Eq.(20) of the fittest chromosome in Example 1

terms $\mathbf{p}(t)$	estimates $\hat{\Theta}_0$	estimates $\hat{\Theta}_s$
$y(t-1)$	1.2395	1.1909
$y(t-1)y(t-2)y(t-3)$	−0.1210	−0.1005
$y(t-1)y(t-2)y(t-8)$	0.2246	0.2088
$y^3(t-2)$	−0.0134	−0.0190
$y^2(t-1)y(t-4)$	0.0413	0.0191
$y(t-1)y^2(t-8)$	−0.1373	−0.1398
$y^3(t-8)$	0.0621	0.0609
$y^2(t-1)y(t-7)$	−0.0641	−0.0381
$y(t-7)y^2(t-8)$	−0.0242	−0.0200
$y(t-1)y(t-5)$	−0.0134	0.0310
$y(t-2)y(t-5)y(t-8)$	0.1225	0.0831
$y^3(t-3)$	0.0463	0.0384
$y(t-2)y(t-4)y(t-8)$	−0.2367	−0.1810
$y(t-1)y(t-4)y(t-7)$	0.1342	0.0992
$y(t-1)y^2(t-5)$	−0.0566	−0.0429

Table 2: Terms and parameter estimates in Example 2

	MSE(estimation set)	MSE(test set)
LS estimates $\hat{\Theta}_0$	0.0231	0.0756
The new estimates $\hat{\Theta}_s$	0.0233	0.0738

Table 3: A comparison of MSE of one-step ahead prediction error in Example 2 (Canadian lynx time series)

estimation data set. This may be caused by the length of the data set, which is very short, or may be an indication of non-stationarity in the time series. The latter implies that the structure of the model should be adapted to achieve a better prediction.

5 Conclusions

This paper has focused on the parameter estimation problem in the framework of linear-in-the-parameters time series models. Least squares estimates have been widely used in parameter estimation of linear-in-the-parameter models. However in certain circumstances the model based on the least squares method can be prone to overfitting and this can lead to a deterioration in the generalisation performance. A criterion called a mean dispersion error (MDE) which is a

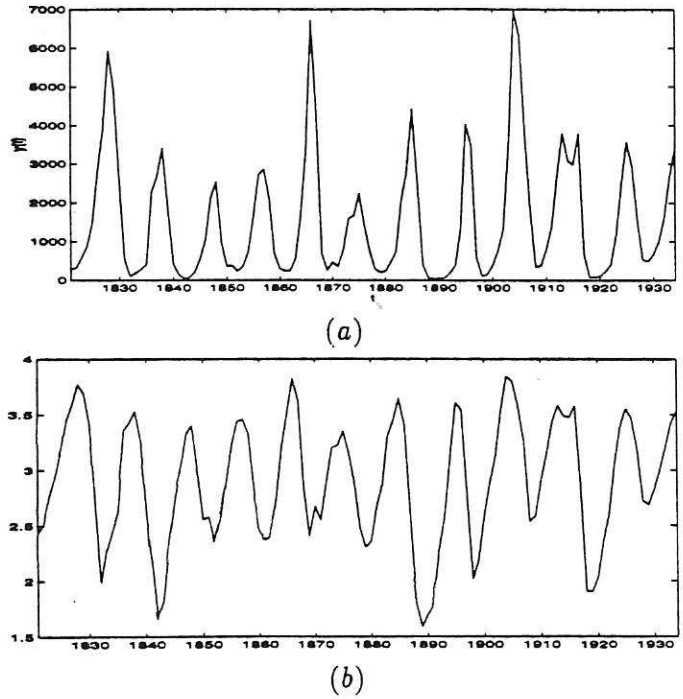

Figure 3: Canadian lynx data; (a) The original time series and (b) The log-transformed time series.

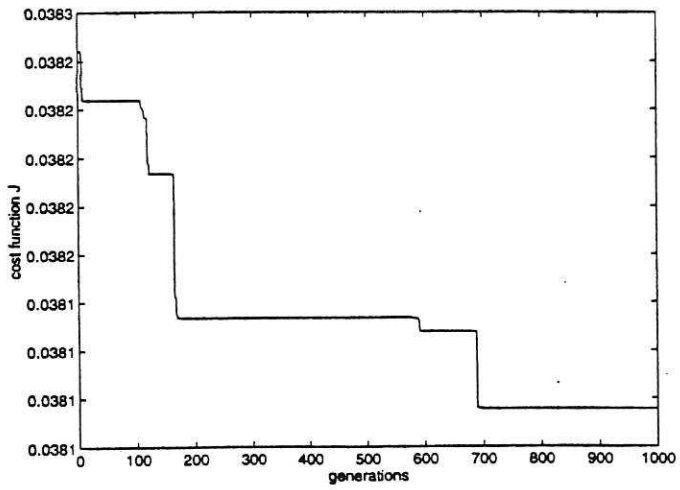

Figure 4: The evolution of the cost function Eq.(20) of the fittest chromosome in Example 2

concept were introduced by Rao and Toutenburg (1995). A theoretically analysis showed that a model with MDE-superior parameter estimates will have a smaller prediction error if this is used in prediction.

Cross-validation has been widely used in the identification of dynamic system structures and parameter estimation. A more sophisticated version of cross-validation which is called stacked generalization involves getting the generalization accuracy as high as possible (Wolpert, 1992). One realisation of stacked generalization is called stacked regression where different predictors are stacked with the objective of minimising the cross-validated prediction errors under the constraints that the nonnegative combination coefficients sum to one (L. Breiman, 1996). Although stacked regression will generally provide better predictions, the obtained predictor does not necessarily result in a parsimonious model structure which is often desired in system identification. In this work, a parameter estimation procedure was proposed which is based on stacked regression and an evolutionary algorithm. Different predictors consisting of subsets of a concise model structure are stacked following the determination of the pasimonious model structure using the forward orthogonal least squares (OLS) procedure. The new parameter estimates for the selected model structure were defined by combining the least squares parameter estimates of each predictor through the optimal combination coefficients which are computed using stacked regression. In the algorithm the cross-validated prediction error in the estimation was computed using the PRESS errors (Myers, 1990). In this mode model parsimony is achieved. The effectiveness of the new approach was demonstrated with numerical examples.

6 Acknowledgements

SAB gratefully acknowledges that part of this work was supported by EPSRC. XH expresses her thanks for the award of an ORS scholarship which made this study possible.

References

[1] Billings, S. A., Korenberg, M. J., and Chen, S. (1988). Identification of non-linear output-affine systems using an orthogonal least-squares algorithm. *Int. J. Systems Sci.*, Vol. 19, pp1559-1568.

[2] Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics.* Vol. 37, No. 4, pp373-384.

[3] Breiman, L. (1996). Stacked regression. *Machine Learning.* 24, pp49-64.

[4] Chen, S. and Billings, S. A. (1989). Modelling and analysis of non-linear time series. *Int. J. Control*, Vol. 50, No. 6, pp2151-2171.

[5] Chen, S., Chng, E. S. and Alkadhimi, K. (1996). Regularized orthogonal least squares algorithm for constructing radial basis function networks. *Int. J. Control*, Vol. 64, No. 5, pp829-837.

[6] Hong, X. and Billings, S. A. (1997). A Givens rotation based fast backward elimination algorithm for RBF neural network pruning. To appear in *IEE Proc. D, Control Theory and Applications*, Vol. 144, No. 5.

[7] Ljung, L. (1987). *System Identification: Theory for the User.*. Prentice-Hall.

[8] Ljung, L. and Glad, T. (1994). *Modelling of Dynamic Systems.* Information and Systems Sciences Series. Prentice Hall, Englewood Cliffs, NJ.

[9] Michalewicz, Z. (1994).*Genetic Algorithms + Data Structures = Evolution Programs.* 2nd ed. Springer-Verlag.

[10] Myers, R. H. (1990).*Classical and Modern Regression with Applications.* 2nd ed. PWS-KENT, Boston.

[11] Orr, M. J. L. (1989). Regularization in the selection of radial basis function centres. *Neural Computation*, 7(3), pp606-623.

[12] Rao, T. S. and Gabr, M. M. (1984). *An Introduction to Bispectral Analysis and Bilinear Time Series Models.* Lecture Notes in Statistics, Vol. 24. Springer-Verlag, New York.

[13] Rao, C. R. and Toutenburg, H. (1995). *Linear Models–Least Squares and Alternatives*, Springer.

[14] Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *J.Roy. Statist. Soc. Ser.B*, 36, pp111-147.

[15] Theobald, C. M. (1974). Generalizations of mean square error applied to ridge regression. *J. Royal Statist. Soc. B*, 36, pp103-106.

[16] Vinod, H. D. and Ullah, A. (1981). *Recent advances in Regression Methods.* STATISTICS: textbooks and monographs; v. 41. Marcel Dekker, Inc.

[17] Wang, L. and Cluett, W. R. (1996). Use of PRESS residuals in dynamic system identification. *Automatica*, Vol. 32, No 5, pp781-784.

[18] Wolpert, D. (1992). Stacked generalization. *Neural Networks*, Vol. 5, pp241-259.