Sparse Model Identification Using a L_1 Regularized Orthogonal Forward Regression Algorithm with a Bootstrap Covariance Criterion

Pan, Y. and Billings, S. A.

Department of Automatic Control and Systems Engineering
University of Sheffield
Sheffield, S1 3JD
UK

Research Report No. 936
August 2006

Sparse Model Identification Using a L_1 Regularized Orthogonal Forward Regression Algorithm with a Bootstrap Covariance Criterion

Pan, Y. and Billings, S.A.
Department of Automatic Control & Systems Engineering
Sheffield University,
Sheffield, S1 3JD, UK

August, 2006

Abstract

This paper presents a new model identification method for parsimoniously selecting model terms and estimating the corresponding parameters of nonlinear dynamical systems. The generalization and prediction capability of the final identified model with the smallest model size is ensured by optimizing the model prediction error over an unseen data set using parametric bootstrap covariance estimates. The bootstrap estimates can be thought of as smoothed versions of cross-validation estimates and do not suffer high variability. The sparseness of the final model is also improved by performing l_1 norm regularisation over the model parameters, integrated into the stepwise orthogonal forward regression algorithm to improve the efficiency of the model selection procedure. The optimal values of the l_1 regularisation associated with different model terms is achieved by optimizing the model evidence in the framework of Bayesian learning theory in a computationally efficient way. The new identified algorithm which combines the l_1 regularized orthogonal forward regression algorithm and the parameteric bootstrap covariance criterion can provide an efficient and data adaptive identification method for modeling nonlinear dynamical systems. Experimental results demonstrate the efficiency and practicability of the new procedure.

1 Introduction

A basic principle in practical system identification problems is the parsimonious principle that the final model should be just large enough to explain the underlying dynamics. Model selection is a central and fundamental issue in ensuring the sparseness in system identification and is especially important in nonlinear system model building. Selecting a simple model structure with good generalization capabilities from a large number of possible model terms is a critically important problem. Various methods of estimating the generalization or prediction capability of identified models have been studied because of the close relationship with model selection procedures. It is widely accepted that the measure or estimate of the prediction error/generalizaiton error over an unseen data set is a good criterion for assessing the final identified model. For example, when a model of the form

$$y = \hat{f}(y, u) + \varepsilon = \hat{y} + \hat{\varepsilon} \tag{1}$$

is estimated based on training set data of input u and output y, it is important to assess how well $\hat{f}(\cdot)$ will predict over future unseen data which is independent of the training data set. The traditional and widely used method for estimating the prediction error over an unseen data set is cross-validation. It is well known that the cross-validation estimate can be nearly unbiased but also highly variable in some situations. Cross-validation can also be computationally costly.

As alternatives to cross-validation, various methods have been proposed to improve the efficiency of model selection criteria. As summarized in [1], most of these selection procedures choose the optimal model $\hat{f}$ by minimizing the following model selection criterion with respect to all possible candidate subsets f, where Y and $\hat{Y}$ are the training data vectors of the output and one step ahead predicted output respectively.

$$(Y - \hat{Y})'(Y - \hat{Y}) + \lambda|\hat{f}|\sigma^2 \tag{2}$$

These selection criteria are composed of a training error term and a covariance penalty term. In (2), λ is the penalty coefficient that controls the degree of penality according to different selection criteria and $|\hat{f}|$ is the dimension or the size of model $\hat{f}$. Most model selection criteria differ only in the choice of the value of λ. This is true for example, in Akaike's information criterion (AIC) [2] and its equivalent form in linear regression, for Mallows's C_p [3] $\lambda = 2$, while in the Bayesian information criterion (BIC) [4] which is based on an asymptotic Bayes factor, $\lambda = \log(n)$ (n is the number of data). Modified and generalized forms of these model selection criteria include the risk inflation criterion (RIC) [5] with $\lambda = 2\log(p)$ (p is the number of candidate regressors) which also has a close relationship with the covariance inflation criterion [6] with $\lambda = 4\sum_{j=1}^{|\hat{f}|}\log(n/j)/|\hat{f}|$ based on permuted versions of the data set, and Stein's unbiased risk estimate (SURE) [7].

In many practical situations, the performance of these model selection criteria with fixed covariance penalties λ varies according to different situations despite their computational efficiencies. For example, when p is large, these criteria may yield a large selection bias. Moreover, for a large λ, the model selection criterion may produce an optimal model whose size is small but which may perform poorly in other situations when the size of the optimal model is large, and vise versa [1]. Another important feature of these model selection criteria are the assumptions regarding the probability distribution of the model residuals. Usually, the distribution of the model residuals is assumed to be

Gaussian, which might be a big disadvantage in some situations especially for nonlinear systems. It is therefore of great importance to use a data adaptive model selection criterion to improve the applicability and reduce the selection bias. Cross-validation yields a nearly unbiased estimation of the model prediction error. However, the low bias of cross-validation is achieved at the cost of high variance. The bootstrap method which can be thought of as a smoothed version of cross-validation [8] [9] [10] can substantially reduce the variance of the estimation of the prediction error and improve the computational efficiency. Besides providing more accurate point estimation for the prediction errors, the bootstrap approach has other important advantages. Bootstrap replications can provide a direct assessment of the variance of a point estimate of the prediction errors. Furthermore, the bootstrap method is nonparametric and can be applied to various identification problems including regression problems providing the model residuals under study are identical and independently distributed (i.i.d.).

In this paper, prediction error estimates using a parametric bootstrap method [11] is adapted to the orthogonal forward regression algorithm (OFR) [12] as a model selection criterion to obtain optimal model results in a computationally effective way. The OFR algorithm is a stepwise orthogonalization procedure of the candidate regressors which provides a forward selection of the significant terms from the candidate model terms based on the Error Reduction Ratio (ERR) criterion. But a practical problem for the OFR algorithm applications using the ERR criterion is that the number of model terms should be prior determined. With the assistance of the bootstrap covariance criterion, the smallest model size can be determined by assessing the estimate of the model prediction error. Therefore, the model identification procedure can be automatically terminated by the new model selection criterion.

It is well known that non-quadratic regularizers, in particular the l_1 norm regularizer which is also termed as a Laplace prior in the Bayesian framework [13], can yield sparse models that generalize well via parameter shrinkage. For some complex models involving high-dimensional real world data, l_1 regularised regression can also avoid over-fitting problems by adding penalty terms associated with the model parameters. Similarly, there is another kind of regularization which is also widely used, l_2 norm regularization [14][15]. In the Bayesian framework, the l_2 regularisers are called Gaussian priors. The main disadvantage of Gaussian priors is that they do not control the structural complexity of the learned function [16][17]. In this paper, l_1 regularization is integrated into the OFR algorithm where each regressor is associated with an individual regularizer to give approximately optimal values of the regularizers in the Bayesian evidence framework. The l_1 regularization provides another way to improve the sparseness of the final model.

The paper is arranged as follows. Section 2 describes the bootstrap covariance criterion for model selection. In section 3, the OFR algorithm is first introduced and the new model identification procedure combined with the bootstrap covariance criterion is presented. The l_1 norm regularized OFR algorithm is then proposed to improve the sparseness of the final model. Finally, three numerical examples in different situations are included in section 5 to illustrate the application of the new identification methods and to demonstrate the efficiency and feasibility of the new algorithm in modeling nonlinear dynamical systems.

2 Bootstrap Covariance Criterion

Assume that a nonlinear dynamical system can be described by the following function

$$y = f(u, y) + e = \mu + e \tag{3}$$

where e is an identical and independently distributed noise sequence.

Consider a following model which is fitted to y,

$$y = \hat{f}(y, u) + \varepsilon = \hat{y} + \hat{e} \tag{4}$$

where the training error of the model $\hat{f}$ is measured by the squared one step ahead error over the training data set.

$$TE = \sum_i TE_i = \sum_i (y_i - \hat{y}_i)^2 \tag{5}$$

Without any assumptions about the model f, the expectation of the prediction errors is given by [11],

$$E\{PE_i\} = E[(y_i - \mu_i)^2 + (\mu_i - \hat{y}_i)^2] = E[(y_i - \hat{y}_i)^2 + 2(\hat{y}_i - \mu_i)(y_i - \mu_i)] \tag{6}$$

Since $TE_i = (y_i - \hat{y}_i)^2$, (6) can also be written as

$$E\{PE_i\} = E\{TE_i + 2cov(\hat{y}_i, y_i)\} \tag{7}$$

where $2cov(\hat{y}_i, y_i)$ is the covariance penalty term added to the training error TE_i to yield an unbiased estimate of the prediction error PE_i. As mentioned in Section 1, the covariance penalty can be estimated in various ways according to different model selection criteria. In this paper, we use the parametric bootstrap method to estimate the covariance penalty which is almost unbiased. This is similar to the cross-validation method but the computational efficiency is enhanced and the variance of the estimate is also reduced.

Let $\hat{f}$ be an assumed prediction of y, and $y = \hat{f}(u, y) + \hat{e}$ where $\hat{e}$ is a random residual sequence. Then a large number "B" of the simulated observations and estimates from $\hat{y}$ can be generated in the following way.

$$\hat{y} \rightarrow y^* \rightarrow \hat{y}^* = \hat{f}(u, y^*) \tag{8}$$

From the bootstrap replications y^* and $\hat{y}^*$, the covariance penalty $cov_i = cov(\hat{y}_i, y_i)$ can be estimated using the following equation [11]

$$\widehat{cov}_i = \sum_{b=1}^{B} \hat{y}^{*b}(y_i^{*b} - \overline{y^*}_i)/(B-1) \tag{9}$$

where $\overline{y^*}_i = \sum_b y_i^{*b}/B$ is the mean of the bootstrap replications at point i. Thus, the bootstrap covariance criterion can be obtained using the estimate of the prediction error.

$$PE = \sum_i (y_i - \hat{y}_i)^2 + 2\sum_i \widehat{cov}_i \tag{10}$$

It can easily be concluded that a smaller estimate of the prediction error yields a better model. Therefore, in terms of generalization error, the model selection criterion based on the bootstrap covariance estimate can be used to assess the model performance with different model sizes, where the smallest model size can consequently be determined. Notice that the bootstrap method dose not require any assumptions on the model except that the model residual is identical and independently distributed. The bootstrap criterion is also data-adaptive where the covariance penalty is directly estimated from the training data. In this paper, the bootstrap estimate will be applied to the identification algorithm to choose an adequate model with smaller model terms.

3 Sparse Model Identification

In this section, sparse model identification algorithms are discussed. In section 3.1, the orthogonal forward regression algorithm is first briefly introduced and a new model selection procedure is proposed by adapting the bootstrap covariance criterion to the OFR algorithm. Finally, a l_1 regularization method is exploited to further improve the sparseness of the final model.

3.1 Orthogonal Forward Regression Algorithm with the Bootstrap Covariance Criterion

A wide class of nonlinear dynamical systems can be described by the NARX (Nonlinear AutoRegressive withe eXogenous inputs) model

$$y(t) = f(y(t-1), ..., y(t-n_y), u(t), ..., u(t-n_u)) + e(t) \tag{11}$$

where $f(\cdot)$ is an unknown mapping, $u(t)$, $y(t)$ and $e(t)$ are the input, output and noise sequences respectively, and n_u, n_y are the maximum input and output lags. The mapping $f(\cdot)$ can be constructed or regressed using a variety of local or global basis functions including polynomials, neural networks, kernel functions and wavelets. In this paper the nonlinear polynomial model is used. In summary, the NARX model can be expressed in a linear-in-the-parameters form

$$y(t) = \sum_{k=1}^{M_s} \theta_i \varphi_k(t) + e(t) \tag{12}$$

where $\varphi_k(t)$ is the model term in a polynomial form generated from the model regressors $\{y(t-1), ..., y(t-n_y), u(t), ..., u(t-n_u)\}$.

The OFR algorithm involves a stepwise orthogonalization of the regressors $\{\varphi_k(t)\}$ and a forward selection of the significant terms based on the Error Reduction Ratio (ERR) criterion [12]. The significant model terms are selected step by step by comparing the ERRs of all possible model terms. This algorithm also computes the optimal least square estimates of the term coefficients $\Theta = \{\theta_k\}$.
Write (12) in the vector format

$$\mathbf{Y} = \Phi\Theta + \mathbf{e} \tag{13}$$

where Φ is the regression matrix or the design matrix, Θ is the coefficient vector and $\mathbf{e}$ is the residual sequence. After orthogonalization, (13) is converted into

$$\mathbf{Y} = \mathbf{W}\mathbf{g} + \mathbf{e} \tag{14}$$

where

$$\mathbf{A} = \begin{pmatrix} 1 & a_{1,2} & \dots & a_{1,M_s} \\ 0 & 1 & \vdots & a_{2,M_s} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix} \tag{15}$$

and

$$\Phi = \mathbf{WA}, \mathbf{A}\Theta = \mathbf{g} \tag{16}$$

In (14), $\mathbf{W}$ is the corresponding orthogonal matrix and M_s denotes the number of terms in the final model.

The OFR algorithm is an effective and practical learning procedure for identifying nonlinear models. An important feature of the OFR algorithm is the capability to select the model terms according to the contribution of each term to the overall model accuracy and to eliminate redundant terms in a computationally effective way. However, the OFR algorithm only involves comparisons between different model terms in the selection procedure and the optimal model size can be determined in another auxiliary way. In this paper, the bootstrap covariance criterion is adapted to OFR algorithm to compare model prediction capabilities with different model sizes and determine the optimal model size step by step. The model one step ahead prediction error associated with the selected model $\{\varphi_k\}_{k=1}^{M_0}$ can be computed using the following equation

$$PE = \sum_t \left(y^2(t) - \sum_k^{M_0} g_k^2 w_k^2(t) \right) + 2 \sum_t \widehat{cov}(t) \tag{17}$$

where M_0 is the number of terms in the forward selection procedure, w_k is the selected orthogonal regression vector and the covariance penalty term $\sum_t \widehat{cov}(t)$ is estimated using the bootstrap method (9). Equation (10) also shows that the bootstrap covariance criterion also represents the prediction error of the selected model.

For a given candidate set of regressors $G = \{\varphi_k\}_{k=1}^{M}$ where M is the number of candidate regressors, the new model identification procedure which combines the OFR algorithm and the bootstrap covariance criterion can be briefly outlined as follows.

Step1: Select the first model term with the highest ERR

$$I_1 = I_M = \{1, 2, ..., M\} \tag{18}$$

$$w_k(t) = \varphi_k(t), \hat{b}_k = \frac{[w_k, Y]}{[w_k, w_k]} \tag{19}$$

$$l_1 = arg \max_{k \in I_1} \hat{b}_k^2 \frac{[w_k, Y]}{[Y, Y]} = arg \max_{k \in I_1} (ERR_k) \tag{20}$$

$$w_1^0 = w_{l_1}, c_1^0 = \frac{[w_1^0, Y]}{[w_1^0, w_1^0]}, a_{1,1} = 1 \tag{21}$$

Estimate the current model one step ahead prediction error using the bootstrap method (9) and (10)

$$PE_1 = \sum_t (y^2(t) - (c_1^0)^2 (w_1^0(t))^2) + 2 \sum_t \widehat{cov}^{(1)}(t) \tag{22}$$

Step j, $j = 2, 3, ...$: Iteratively orthogonalize the remaining regressors one by one to select the next model term with highest ERR among the remaining candidate terms.

$$I_j = I_{j-1} \setminus l_j - 1 \tag{23}$$

$$w_k(t) = \varphi_k(t) - \sum_{k=1}^{j-1} \frac{[w_k^0, Y]}{[w_k^0, w_k^0]}, \hat{b}_k = \frac{[w_k, Y]}{[w_k, w_k]} \tag{24}$$

$$l_j = arg \max_{k \in I_j} \hat{b}_k^2 \frac{[w_k, Y]}{[Y, Y]} = arg \max_{k \in I_j} (ERR_k) \tag{25}$$

$$w_j^0 = w_{l_j}, c_j^0 = \frac{[w_j^0, Y]}{[w_j^0, w_j^0]}, a_{k,j} = \frac{[w_k^0, \varphi_{l_j}]}{[w_k^0, w_k^0]}, k = 1, 2, ..., j-1 \tag{26}$$

Estimate the new model prediction error using the bootstrap method

$$PE_j = \sum_t \left(y^2(t) - \sum_{n=1}^{j} (c_n^0)^2 (w_n^0(t))^2 \right) + 2 \sum_t \widehat{cov}^{(j)}(t) \tag{27}$$

This procedure is terminated at the M_s-th step when the estimate of the new model prediction error PE_j at the j step l_j is bigger than PE_{j-1} at the $j-1$ step. The estimated coefficients $\Theta = \{\theta_k\}_{k=1}^{k=M_s}$ associated with the selected terms$\{\varphi_{l_k}\}_{k=1}^{k=M_s}$ are computed in the following equation

$$\Theta = \mathbf{A}^{-1} C, \tag{28}$$

where $\mathbf{A}$ is upper-triangular matrix which is defined in (15)and $C = (c_1^0, c_2^0, ..., c_{M_s}^0)$ is the coefficient vector associated with the orthogonalised terms$\{w_k^0\}_{k=1}^{M_s}$.

3.2 l_1 Regularization Regression

Regularization is performed by introducing a kind of penalty function with some hyper parameters associated with the model parameters. An important problem for regularised regression methods is how to find the appropriate values of the hyper parameters without any subjective work to achieve a better model approximation. These problems have been extensively studied in [18]. Solutions include the Discrepancy Principle [19], generalized cross-validation [20] and the L-curve method[21]. But these methods are all computationally costly. In this section, a new method is proposed to find appropriate values of the hyper parameters in the Bayesian framework, which are defined in terms of the noise variance and a measurement of smoothness of the model fit. The Bayesian method allows values to be objectively assigned to the tuning parameters which are commonly unknown a priori. A typical advantage of the Bayesian learning method is that it can quantitatively rank a whole class of models by evaluating the corresponding evidence and the hyper parameters are consequently tuned to maximize the evidence.

In this paper, l_1 regularization is adapted to the OFR algorithm for nonlinear system identification. The optimal values of the l_1 regularisers are given using Bayesian learning theory. This new method has two main advantages, one is that the effective model term selection procedure of the OFR algorithm is maintained and the contribution of the individual regularisers to evidence of the regression model can be evaluated by orthogonalizing the candidate regressors, the other advantage is that the optimal regularisers can be inferred by maximizing the evidence in the Bayesian learning framework.

In the Bayesian framework, the optimal estimates of the parameters for the orthogonal regression model (14) are obtained by maximizing the posterior probability of the parameters $\mathbf{g} = (g_1, ..., g_{M_s})$ which is given by

$$P(\mathbf{g}|\mathbf{Y}, \mathbf{\Lambda}, \epsilon) = \frac{P(\mathbf{Y}|\mathbf{g}, \mathbf{\Lambda}, \epsilon)P(\mathbf{g}|\mathbf{\Lambda}, \epsilon)}{P(\mathbf{Y}|\mathbf{\Lambda}, \epsilon)} \tag{29}$$

where $P(\mathbf{Y}|\mathbf{g}, \mathbf{\Lambda}, \epsilon)$ is the likelihood function, $P(\mathbf{g}|\mathbf{\Lambda}, \epsilon)$ is the priori probability density with regularisers $\mathbf{\Lambda} = (\lambda_1, ..., \lambda_{M_s})$ and $\epsilon = 1/\sigma_{\mathrm{e}}^2$ denoting the smoothness of the fitted regression model and the noise model of data respectively, $P(\mathbf{Y}|\lambda, \epsilon)$ is the evidence of the regression model associated with the regularisers $\mathbf{\Lambda}$ and ϵ.

Here, it is assumed that the residual sequence $\mathbf{e}$ is zero mean gaussian noise with standard deviation σ_{e}. Following [22],the likelihood function can therefore be described as

$$P(\mathbf{Y}|\mathbf{g}, \mathbf{\Lambda}, \epsilon) = \left(\frac{\epsilon}{2\pi}\right)^{N/2} \exp\left(-\frac{\epsilon \mathbf{e}^T \mathbf{e}}{2}\right) \tag{30}$$

The density function of the parameters $\mathbf{g}$ with the l_1 regularisers can be written as follows.

$$P(\mathbf{g}|\mathbf{\Lambda}, \epsilon) = \prod_{i=1}^{M_s} \frac{\lambda_i}{2} \exp(-\lambda_i |g_i|) \tag{31}$$

Maximizing the log posterior probability with l_1 regularisers with respect to g is equivalent to minimizing the following cost function.

$$\mathbf{J}_{BL}(\mathrm{g}, \mathbf{\Lambda}, \epsilon) = \frac{1}{2}\epsilon \mathrm{e}^T \mathrm{e} + \mathrm{g}|\Lambda| \tag{32}$$

The optimal values of g_i to maximize the log posterior probability is obtained by setting $\partial \log(P(\mathbf{Y}|\mathrm{g}, \mathbf{\Lambda}, \epsilon)P(\mathrm{g}|\mathbf{\Lambda}, \epsilon)/\partial g_i = 0$,which yields (refer to Appendix B)

$$g_i = sgn(w_i^T \mathbf{Y}) \left(\frac{|w_i^T \mathbf{Y}|}{\|w_i\|^2} - \frac{\lambda_i}{\epsilon \|w_i\|^2} \right)_+ \tag{33}$$

where $\|\mathbf{v}\| = \sum_i v_i^2$ denotes the squared Euclidean norm, $(\cdot)_+$ is the *positive part operator* (defined as $(a)_+ = a$, if $a \geq 0$, and $(a)_+ = 0$, if $a < 0$), and $sgn(\cdot)$ is the sign function. Note that when the absolute value of $w_i^T \mathbf{Y}/\|w_i\|^2$ is below a threshold, the estimate of g_i is exactly zero; otherwise, the estimate is obtained by subtracting a threshold. This rule is also called a *soft threshold* [23].

And using the Gaussian approximation method, the log evidence of the model with l_1 regularisers Λ and ϵ can be approximated as

$$\begin{aligned} \log(P(\mathbf{Y}|\Lambda, \epsilon)) \approx & \sum_{i=1}^{M_s} \log(\lambda_i)/2 - \frac{M_s}{2}\log(\pi) - \frac{N}{2}\log(2\pi) + \frac{N}{2}\log(\epsilon) \\ & - \sum_{i=1}^{M_s} (\lambda_i |g_i|) - \frac{1}{2}\epsilon \mathrm{e}^T \mathrm{e} - \frac{1}{2}\log(\det(\mathrm{B})) + \frac{M_s}{2}\log(2\pi) \end{aligned} \tag{34}$$

where g is set to be the optimal value of a posterior probability solution. The Hessian matrix $\mathrm{B} = \nabla_{\mathrm{g}}^2 (J_{(BL)}(\mathrm{g}, \mathbf{\Lambda}, \epsilon))$ is diagonal when $g_i \neq 0$ due to the orthogonalization of the design matrix which can easily be satisfied for practical regression problems. The corresponding Hessian matrix is given by

$$\mathrm{B} = \epsilon \mathbf{W}^T \mathbf{W} = diag\{\epsilon w_1^T w_1, ..., \epsilon w_{M_s}^T w_{M_s}\} \tag{35}$$

Setting $\partial \log(P(\mathbf{Y}|\mathbf{\Lambda}, \epsilon))/\partial \varepsilon = 0$ yields the computation formula for the optimal ε

$$\varepsilon = (N - M_s)/\mathrm{e}^T \mathrm{e} \tag{36}$$

Setting $\partial \log(P(\mathbf{Y}|\mathbf{\Lambda}, \epsilon))/\partial \lambda_i = 0$ yields the computation formula for the optimal λ_i

$$\lambda_i = \frac{1}{|g_i|} \tag{37}$$

For a large sample of data, the optimal estimate of the variance of the residual usually changes slightly, so the influence of the noise prior on the parameter g can be ignored. The optimal estimate of the parameter g_i with the optimal regulariser λ_i can be therefore written as

$$g_i^{(BL)} = sgn(w_i^T \mathbf{Y}) \left(\frac{|w_i^T \mathbf{Y}|}{\|w_i\|^2} - \frac{1}{\epsilon \|w_i\|_2 |g_i|} \right)_+ \tag{38}$$

As a summary, the new model identification method which adapts the bootstrap covariance criterion to the l_1 regularized OFR algorithm can be briefly described as follows.
(a) Use the OFR algorithm with bootstrap covariance criterion described in Section 3.1 to select significant model terms from the candidate terms with the smallest model size and give an initial least squares estimate of the parameter $\mathbf{g}$.
(c) Calculate the value of the l_1 regularisers $\mathbf{\Lambda}$ and the noise prior ϵ using (37) and (36) using the paramter $\mathbf{g}$.
(d) Calculate the optimal regularized estimate of the parameter $\mathbf{g}^{(BL)}$ using (38) with the optimal value of l_1 regularisers.
(e) Remove the corresponding model terms if the associated parameters are zero after regularization shrinkage.

4 Numerical Examples

In this section, two numerical examples are included to illustrate the new identification method proposed in this paper. The first example is a nonlinear time series simulated using a non-polynomial model. The other example concerns the real sampled data from monthly ozone time series where a nonlinear polynomial model is fitted.

In this section, model predicted output will be employed to test the prediction capability of the identified models over future unseen data. For an identified NARX model $y(t) = \hat{f}(u(t-1), ..., u(t-n_u), y(t-1), ..., y(t-n_y)) + e(t)$, the many step ahead predicted output is defined in the following way.

$$y^{(msa)}(t) = \hat{f}(u(t-1), ..., u(t-n_u), y^{(msa)}(t-1), ..., y^{(msa)}(t-n_y)) \qquad (39)$$

Notice that this is a much more severe test than the common approach of using one-step-ahead predictions.

4.1 Example1: A Simulated Nonlinear Time Series

Consider a nonlinear time series given by the following equation

$$\begin{aligned} y(t) = {} & (0.8 - 0.5\exp(-y^2(t-1)))y(t-1) - (0.3+ \\ & 0.9\exp(-y^2(t-1)))y(t-2) + 0.1\sin(\pi y(t-1)) + e(t), t = 1, 2, ... \qquad (40) \end{aligned}$$

where $e(t)$ is the random noise sequence with zero mean and standard deviation $\sigma = 0.1767$. The nonlinear time series were numerically simulated for 550 times from the initial conditions $y(0) = y(-1) = 0$. The first 500 noisy data were used for model identification while the remaining 50 data were used to test the model prediction capabilities.

The new model identification algorithm is applied and the model results are given in Table(1). Fig.(1) shows the estimates of the prediction error using the parametric bootstrap method in (9) and (10). It can be seen that when the model size is equal to 4, the prediction error for the identified model reaches the minimum value. Consequently, four model terms are selected in the final model. The one-step-ahead predicted output and error are plotted in Fig.(2). The many step ahead predicted output is plotted against the unseen test data in Fig.(3), from which it can be seen the identified model has a very good prediction capability for the future data set.

Table 1: Terms and parameters of the identified model for the simulated time series in Example 1

Model terms	Estimated parameters without regularization	Estimated parameters with l_1 regularization	ERR	PE
$y(t-3)$	-0.0036	-0.0055	0.6392	132.25
$y(t-2)$	-1.095	-1.0917	0.0854	101.46
$y(t-1)$	0.6554	0.65366	0.0915	68.556
$y^2(t-1)y(t-2)$	0.33507	0.33375	0.0621	44.544

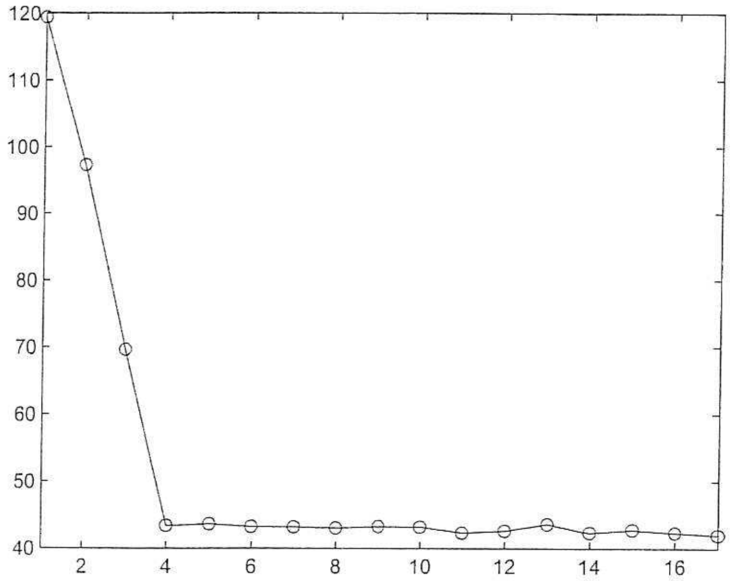

Figure 1: The estimates of the model prediction errors for different model sizes of the simulated time series.

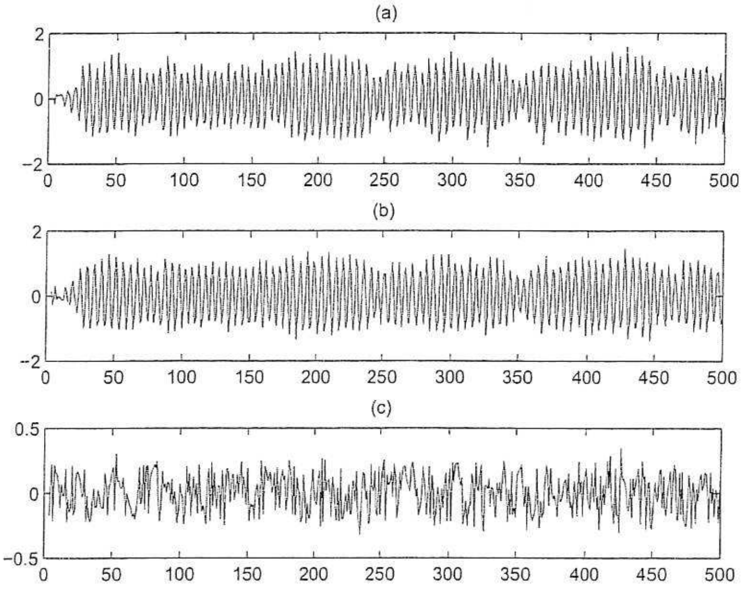

Figure 2: One-step-ahead prediction of the simulated time series in Example 1. Figure (a) indicates the system measurements, Figure (b) indicate the one-step-ahead predicted outputs for the final model with l_1 regularization and Figure (c) indicates the one-step-ahead prediction error.

4.2 Example2: Ozone Concentration Time Series

The ozone concentration has been widely studied because of the close link to the environment and the effects on human beings. In this example, the monthly time series of the mean concentration of the ozone layer in Dobson units at Arosa, Switzerland, from 1930 to 1971 [24] (the ozone data can be found in the following website http://www-personal.buseco.monash.edu.au/ hyndman/TSDL/) were identified using the new proposed modelling algorithm. A few sampling points that are missing in the original data were amended using a linear interpolation method. The maximum time lag of the candidate model terms was set to be 12 and the nonlinear degree to 2. Actually, it can be easily found from the sampling curve that the monthly ozone time series have a rough period of 12.

The ozone concentration time series were divided into two groups where the first 455 samples were used for model identification and the remaining 31 samples were used to test the many step ahead prediction capability of the identified model. The model identification results are given in Table(2), where 10 model terms are selected using the bootstrap covariance criterion. Fig.(4) shows the estimates of the prediction error associated with different model sizes.

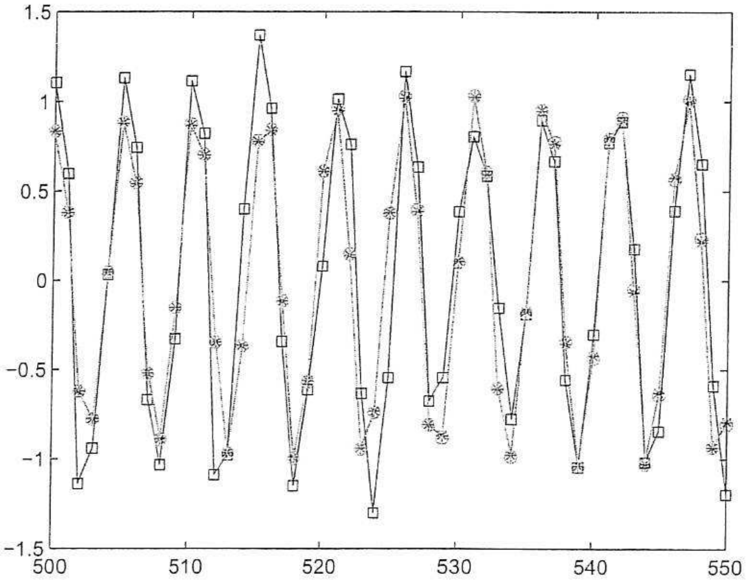

Figure 3: Many step ahead predicted output of the simulated time series in Example 1. Solid line with squares indicates the system measurements , the lines with stars and with circles indicate the model predicted outputs for the final model with l_1 and without l_1 regularization respectively.

From Table(2), it can be seen that the differences of the term parameters after l_1 regularization is trivial in this example. Fig.(5) shows the one-step-ahead predicted output and the training error for the ozone concentration time series. The model predicted outputs of the identified models in Table(2) are plotted against the test data in Fig.(6). It can be seen that the identified model has a fairly good prediction capability for the ozone concentration time series.

5 Conclusions

The problem of identifying sparse models for nonlinear dynamical systems has been studied. The sparseness of the final model is ensured in two ways using the new proposed identification algorithm. The optimal model size is initially determined by the bootstrap covariance criterion with the smallest estimates of prediction error, which is more computationally efficient than cross validation. Complexities of the final models can also be reduced in some situations by the l_1 norm regularization method using the parameters shrinkage approach.

Table 2: Terms and parameters of the identified model for the ozone concentration time series in Example 2

Model terms	Estimated parameters without regularization	Estimated parameters with l_1 regularization	ERR	PE
$y(t-12)$	0.7487	0.7838	0.99505	2.5471e5
$y(t-1)$	1.2837	1.181	0.00115	1.9532e5
$y(t-10)y(t-11)$	0.0006	0.0006	0.00094	1.4604e5
$y(t-2)$	-0.7613	-0.7269	0.00015	1.3794e5
$y(t-4)y(t-12)$	0.0021	0.0020	0.00008	1.3396e5
$y(t-2)y(t-9)$	0.0030	0.0029	0.00005	1.3215e5
$y^2(t-12)$	-0.0009	-0.0010	0.00017	1.2851e5
$y(t-4)y(t-9)$	-0.0025	-0.0024	0.00007	1.2089e5
$y(t-5)y(t-9)$	-0.0005	-0.0005	0.00004	1.1539e5
$y(t-1)y(t-12)$	-0.0026	-0.0023	0.0002	1.1301e5

References

[1] X. Shen and J. Ye, "Adaptive model seletion," *J. Am. Statist. Ass.*, vol. 97, pp. 210–221, 2002.

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

[3] C. Mallows, "Some comments on cp," *Techmetrics*, vol. 15, pp. 661–675, 1973.

[4] G. Schwarz, "Estimating the dimension of a model," *Ann. Statist.*, vol. 6, pp. 461–464, 1979.

[5] D. Foster and E. George, "The risk inflation criterion for multiple regression," *Ann. Statist.*, vol. 22, pp. 461–470, 1994.

[6] R. Tibshirani and K. Knight, "The covariance inflation criterion for adaptive model selection," *J. R. Statist. Soc. (B)*, vol. 61, pp. 529–546, 1999.

[7] C. Stein, "Estimation of the mean of a multivariate nomal distribution," *Ann. Statist.*, vol. 9, pp. 1135–1151, 1981.

[8] B. Efron, "Bootstrap method: another look at the jackknife," *Ann. Statist.*, vol. 7, pp. 1–26, 1979.

[9] ——, "Estimating the error rate of a prediction rule: some improvements on cross-validation," *J. Am. Statist. Ass*, vol. 78, pp. 338–361, 1983.

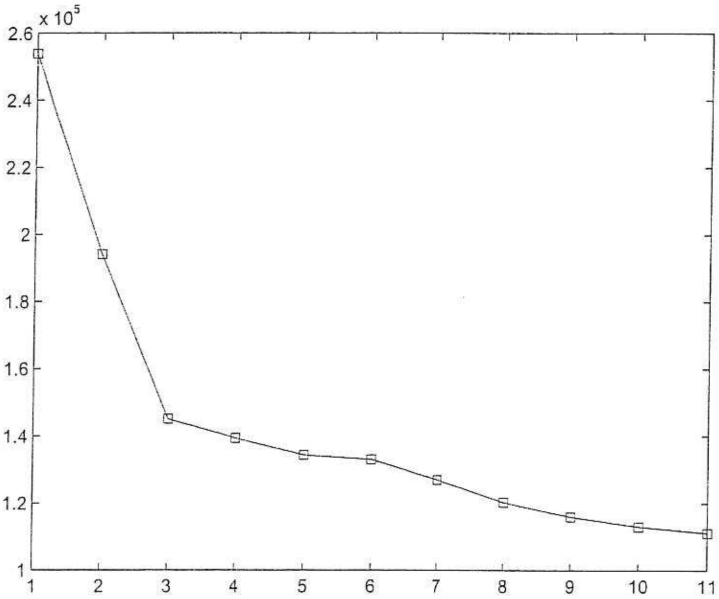

Figure 4: The estimates of the model prediction errors for different model sizes of the ozone concentration time series.

[10] B. Efron and R. Tibshirani, "Cross-validation and the bootstrap: Estimating the error rate of a prediction rule," *Technical Report 176, Dept of Statistics, Stanford University.*

[11] B. Efron, "Estimating the prediction error: covariance penalties and cross-validation," *J. Am. Statist. Ass*, vol. 99, pp. 619–632, 2004.

[12] S. Billings, S. Chen, and M. Kronenberg, "Identification of mimo non-linear systems using a forward regression orthogonal estimator," *Int. J. Control*, vol. 49, pp. 2157–2189, 1988.

[13] P. Williams, "Bayesian regularization and pruning using a laplace prior," *Neural computation*, vol. 7, pp. 117–143, 1995.

[14] S. Chen, X. Hong, and C. Harris, "Sparse kernel regression modelling using combined locally regularized orthogonal least squares and d-optimality experimental design," *IEEE Trans. Automatic Control*, vol. 48, pp. 1029–1036, 2003.

[15] M. Tipping, "Sparse bayesian learning and the relevance vector machine," *J. Machine Learning Reseach*, vol. 3, pp. 211–244, 2001.

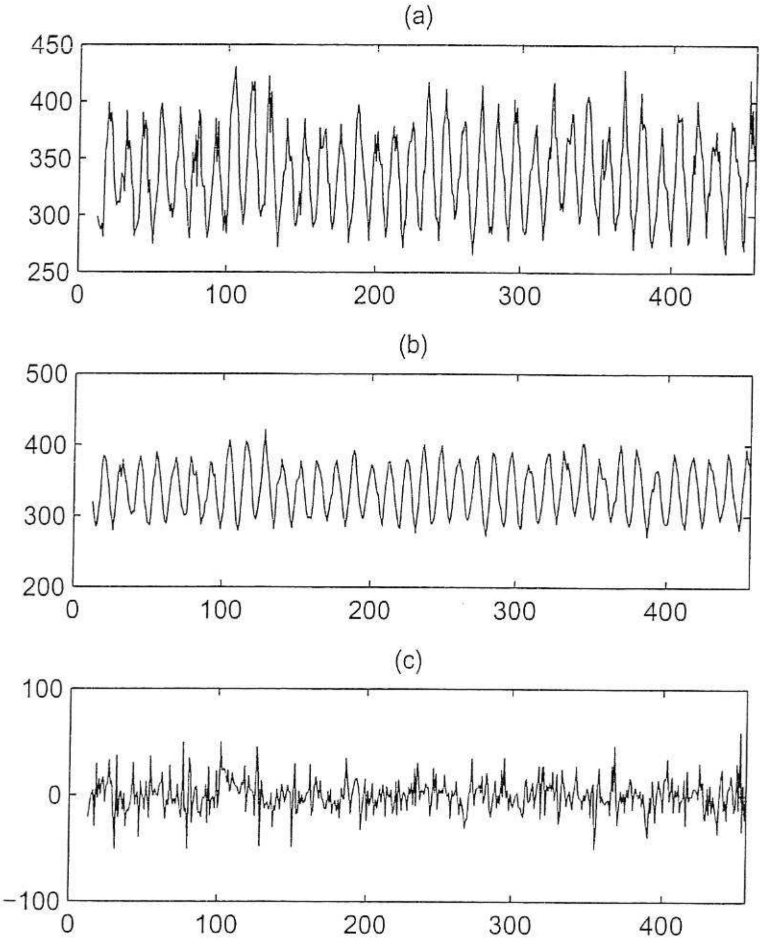

Figure 5: Model predicted output of the ozone concentration time series in Example 2. Figure (a) indicates the system measurements, Figure (b) indicate the one-step-ahead predicted outputs for the final model with l_1 regularization and Figure (c) indicates the one-step-ahead prediction error.

[16] M. Figueiredo, "Adaptive sparseness for supervised learning," *IEEE Trans. Pattern Analysis and Machine Interlligence*, vol. 25, pp. 1150–1159, 2003.

[17] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. R. Statist. Soc. (B)*, vol. 58, pp. 267–288, 1996.

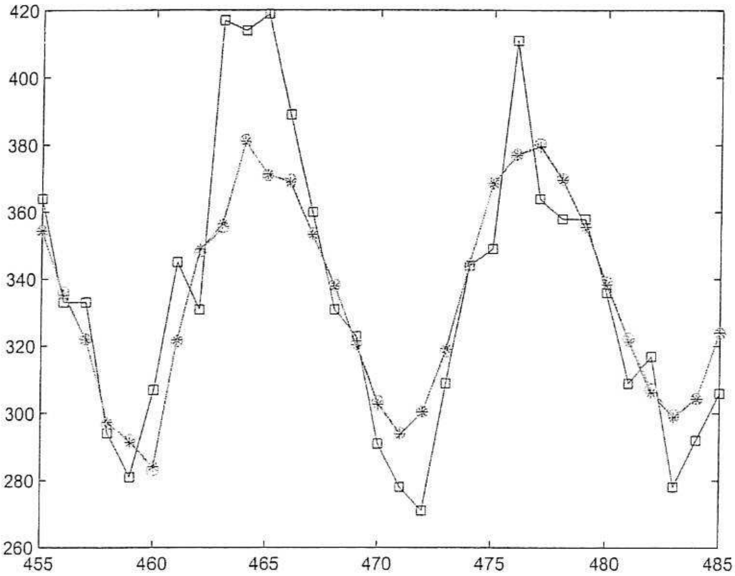

Figure 6: Many step ahead predicted output of the ozone concentration time series in Example 2. Solid line with squares indicates the ozone concentration measurements , the line with stars and the line with circles indicate the model predicted outputs for the final model with l_1 and without l_1 regularization respectively.

[18] M. Kilmer and D. O'Leary, "Choosing regularization parameters in iterative methods for ill-posed problems," *SIAM J. Matrix Analysis and Applications*, vol. 22, pp. 1204–1221, 2001.

[19] A. Frommer and P. Maass, "Fast cg-based methods for tikhonov-philips regularization," *SIAM J. Sci. Comput.*, vol. 20, pp. 1831–1850, 1999.

[20] G. Golub, M. Heath, and G. Wahba, "Generalized cross-validation as a method for choosing a good ridge parameter," *Technometrics*, vol. 21, pp. 215–233, 1979.

[21] P. Hansen, "Analysis of discrete ill-posed problems by means of the l-curve," *SIAM Rev.*, vol. 34, pp. 561–580, 1992.

[22] D. MacKay, "Bayesian interpolation," *Neural Computation*, vol. 4, pp. 417–447, 1992.

[23] D. Donoho and I. Johnstone, "Ideal spatial adaption via wavelet shrinkage," *Biometrika*, vol. 81, pp. 425–455, 1994.

[24] D. Andrews and A. Herzberg, *Data: A collection of problems from many fields for the student and research worker.* Springe, 1985.