REDUCING THE COMPUTATIONAL REQUIREMENT OF THE ORTHOGONAL LEAST SQUARES ALGORITHM

E.S.Chng[1] *S.Chen*[2] *B.Mulgrew*[1]

[1] Department of Electrical Engineering, The University of Edinburgh, UK
[2] School of Systems Engineering, The University of Portsmouth, UK

ABSTRACT

The orthogonal least squares (OLS) algorithm is an efficient implementation of the forward regression procedure for subset model selection. The ability to find good subset parameters with only linear increase in computational complexity makes this method attractive for practical implementations. In this paper, we will examine the computation requirement of the OLS algorithm to reduce a model of K terms to a subset model of R terms when the number of training data available is N. We will show that in the case where $N \gg K$, we can reduce the computation requirement by introducing an unitary transformation on the problem.

1. INTRODUCTION

Most nonlinear predictors created using radial basis functions (RBF) [2, 5] and Volterra expansion [6] results in a very large initial model that has the *linear-in-parameter* characteristic (figure 1). We can normally reduce this big model to a parsimonious one without significant degradation in performance if the subset model's parameters are chosen carefully. In fact, parsimonious models are sometimes preferred as they have better generalisation characteristic. This is especially true when models are used for time-series prediction. Large models tend to over-fit in the training phase and thus have poor prediction performance in the testing phase.

To find the optimum R parameter subset model from a original K parameter model, we must find the performance of models using all combinations of R parameters from the full set of K parameters and choose the best one. This requires exponential computation power and is thus prohibitively expensive. In the case of using the OLS algorithm to select subset model, the model found is not guaranteed to be optimum [3, 4]. However, we do normally get good selection with only linearly increasing computation complexity to find additional parameters for the subset model.

2. OLS ALGORITHM

Let us represent these nonlinear predictors that have the linear in parameter structure as a linear regression model.

$$\mathbf{y} = \mathbf{X}\mathbf{h} + \mathbf{e} \tag{1}$$

where $\mathbf{y}$ is the desired signal vector, $\mathbf{X}$ is the information matrix of size $N \times K$, $\mathbf{h}$ is the parameter vector of the model to be found and $\mathbf{e}$ the error vector of approximating $\mathbf{y}$ by $\mathbf{Xh}$. The column vector $\mathbf{y}$ and $\mathbf{e}$ contain N elements. i.e., the N test data and the N values of error in prediction.

The original $\mathbf{X}$ matrix will have K columns. To create a parsimonious model which has R parameters, we are actually trying to pick R columns from the input matrix $\mathbf{X}$ to form a subset input matrix $\mathbf{X}s$. The OLS algorithm selects columns from the input matrix sequentially. At each selection, all the unused columns are studied to find out how each column will contribute to fit the desired vector $\mathbf{y}$ with the current subset $\mathbf{X}s$. The column that provides the best combination with $\mathbf{X}s$ to model $\mathbf{y}$ will be picked to form the new $\mathbf{X}s$. The above selection technique is repeated until the number of columns in $\mathbf{X}s$ equals to R. The details of the algorithm can be found in Chen *et al* [1, 2].

3. REDUCED OLS

The amount of computation used by the OLS algorithm to find a R subset model from the initial model that has N test data and K parameters can be calculated with the following equation:

$$\text{No. of multiplications (OLS)} = \sum_{i=1}^{R}(3N(K-i-1)) + \sum_{i=1}^{R-1}(2N(K-i)) \tag{2}$$

The above equation ignores the addition/subtraction operations required by the OLS method.

If $N \gg K$, it may be possible to save computation by introducing an invariant transformation on the $\mathbf{X}$ matrix. This can be accomplished by pre-multiplying equation (1) with an orthonormal matrix $\mathbf{Q}^T$ where

0-7803-1775-0/94 $3.00 © 1994 IEEE

the columns of $\mathbf{Q}$ spans the column space of $\mathbf{X}$. This operation is called the unitary transformation [7, 8]. We can think of an unitary transformation as a rotation and/or reflection operation. As such operation preserves the length of each (column) vector and the angle between two vectors in the transformed matrix, we have not lost or created any new information.

The following sub-sections show how to create the orthogonal matrix $\mathbf{Q}$ by using the Gram-Schmidt (GS) [7, 8] method and the single value decomposition (SVD) [7, 8] method.

3.1. Gram-Schmidt Approach

The Gram-Schmidt decomposition on the $N \times K$ matrix $\mathbf{X}$ results in the product of an $N \times K$ orthonormal matrix $\mathbf{Q}$ and a $K \times K$ upper triangle matrix $\mathbf{B}$, where $\mathbf{Q}^T\mathbf{Q} = \mathbf{I}$, $\mathbf{I}$ is an identity matrix of appropriate size and $\mathbf{B}$ a $K \times K$ upper triangle matrix. That is

$$\mathbf{X} = \mathbf{QB} \tag{3}$$

The transposed $\mathbf{Q}$, i.e. $\mathbf{Q}^T$, will be used to pre-multiply equation (1) to get

$$\mathbf{Q}^T\mathbf{y} = \mathbf{Bh} + \mathbf{Q}^T\mathbf{e} \tag{4}$$

If we introduce $\tilde{\mathbf{y}} = \mathbf{Q}^T\mathbf{y}$, $\tilde{\mathbf{X}} = \mathbf{B}$ and $\tilde{\mathbf{e}} = \mathbf{Q}^T\mathbf{e}$, we can rewrite equation (4) as

$$\tilde{\mathbf{y}} = \tilde{\mathbf{X}}\mathbf{h} + \tilde{\mathbf{e}} \tag{5}$$

where $\tilde{\mathbf{y}}$ and $\tilde{\mathbf{e}}$ are $K \times 1$ vectors and $\tilde{\mathbf{X}}$ is a $K \times K$ matrix. We can apply the OLS algorithm to perform subset selection based on $\tilde{\mathbf{y}}$ and $\tilde{\mathbf{X}}$. The columns selected to form the subset model $\mathbf{X}s$ using $\tilde{\mathbf{y}}$ and $\tilde{\mathbf{X}}$ is identical to that of using $\mathbf{y}$ and $\mathbf{X}$.

To decompose $\mathbf{X}$ into $\mathbf{QB}$ using Gram-Schmidt method, we require approximately $N \times K^2$ multiplications [9]. If savings in computation by using the OLS algorithm on $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{y}}$ offsets this additional pre-processing computation, this reduced OLS method should be applied.

The number of multiplications to perform the pre-processing using the Gram-Schmidt decomposition is calculated using

$$\text{No. multiplications (GS decomposition)} = NK^2 + NK \tag{6}$$

Figure 2 shows the number of multiplications required by the OLS to perform subset selection from an information matrix of size 500 × 84. The x-axis indicates the number of parameters required to form the subset model. For this example, we can see that when a subset model of size greater than 24 parameters is desired, we should implement the unitary transformation on the problem.

3.2. SVD Approach

To further reduce the computational load of the OLS, we can use an approximated matrix $\hat{\mathbf{X}}$ to represent $\mathbf{X}$. We define $\mathbf{X}$ and $\hat{\mathbf{X}}$ as

$$\mathbf{X} = \mathbf{U\Lambda V}^T \tag{7}$$

$$\hat{\mathbf{X}} = \mathbf{U}_R\mathbf{\Lambda}_R\mathbf{V}_R^T \qquad R < K \tag{8}$$

where the columns of $\mathbf{U}$ are the left eigenvectors, $\mathbf{\Lambda}$ is the diagonal matrix containing the singular values and the rows of $\mathbf{V}^T$ are the right eigenvectors formed by using SVD on $\mathbf{X}$. The singular values in $\mathbf{\Lambda}$ are arranged such that $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_K$. The matrix $\hat{\mathbf{X}}$ is a rank R approximation of the matrix $\mathbf{X}$ created by the product of matrix $\mathbf{U}_R$, $\mathbf{\Lambda}_R$ and $\mathbf{V}_R^T$. The $N \times R$ matrix $\mathbf{U}_R$ is formed by using the first R columns of $\mathbf{U}$, the diagonal $R \times R$ matrix $\mathbf{\Lambda}_R$ is formed by using the first R rows and columns of $\mathbf{\Lambda}$, and the $R \times K$ matrix $\mathbf{V}_R^T$ is formed by using the first R rows of $\mathbf{V}^T$.

If $\hat{\mathbf{X}}$ is used to approximate $\mathbf{X}$, equation (1) will be written as

$$\mathbf{y} \approx \hat{\mathbf{X}}\mathbf{h} + \mathbf{e} \tag{9}$$

Pre-multiply the previous equation by $\mathbf{U}_R^T$, we get

$$\mathbf{U}_R^T\mathbf{y} \approx \mathbf{\Lambda}_R\mathbf{V}_R^T\mathbf{h} + \mathbf{U}_R^T\mathbf{e} \tag{10}$$

If we introduce the $R \times 1$ vectors $\mathbf{y}_R = \mathbf{U}_R^T\mathbf{y}$ and $\mathbf{e}_R = \mathbf{U}_R^T\mathbf{e}$, and the $R \times K$ matrix $\tilde{\mathbf{X}}_R = \mathbf{\Lambda}_R\mathbf{V}_R^T$, equation (10) can be written as

$$\tilde{\mathbf{y}}_R \approx \tilde{\mathbf{X}}_R\mathbf{h} + \tilde{\mathbf{e}}_R \tag{11}$$

Since the dimension of $\tilde{\mathbf{y}}_R$ and $\tilde{\mathbf{X}}_R$ is smaller then that of the vector $\tilde{\mathbf{y}}$ and matrix $\tilde{\mathbf{X}}$ in equation (5), the computation requirement is further reduced when the OLS algorithm is applied. This method is only appropriate when the approximation of $\mathbf{X}$, i.e $\hat{\mathbf{X}}$, is created by a sufficiently large rank R, otherwise the subset model found may not be good.

4. RESULTS

The two reduced OLS algorithm was applied to find subset models of a 84 tap Volterra predictor created using a degree 3, embedding vector length 6 expansion. The Volterra predictor was created using the following expansion:

$$\begin{aligned}\hat{s}_i &= \sum_{j=1}^{M} h'_j s_{i-j} + \sum_{j \leq k}^{M} h'_{jk} s_{i-j} s_{i-k} + \cdots \\ &\quad + \sum_{j \leq k \ldots \leq L}^{M} h'_{jk \cdots L} s_{i-j} s_{i-k} \ldots s_{i-L}\end{aligned} \tag{12}$$

where $M = 6$ is the number of past signal samples used in the expansion and is the dimension of the embedding vector. $L = 3$ is the highest power of combination of past signal values and is called the degree of expansion. From figure 1, we can see that the Volterra kernels $h'_j, h'_{jk}, \cdots, h'_{jk\ldots L}$ correspond to the parameters $h_1, h_2 \cdots, h_{K-1}$ of the predictor, and the monomials $s_{i-j}, s_{i-j}s_{i-k}, \cdots, s_{i-j}s_{i-k}\ldots s_{i-L}$ are the corresponding transformed model inputs $x_{i-1}, x_{i-2}, \cdots, x_{i-(K-1)}$.

The predictor was used to perform single-step prediction on the chaotic time series (Figure 3) generated by the Duffing's equation. For the experiment, we had used 500 training data, i.e. $N = 500$. The information matrix, $\mathbf{X}$, is thus of size 500×84. The least squares criteria was used to find the parameters of the full model, i.e., $\mathbf{h} = \mathbf{X}^+\mathbf{y}$, where $\mathbf{X}^+$ is the pseudo-inverse of $\mathbf{X}$ [7]. If the least squares approximation of $\mathbf{y}$ using product of $\mathbf{Xh}$ is not perfect, we will have a modelling error. To measure the modelling quality of the predictor, the normalised mean square error ($NMSE$) is used:

$$NMSE = 10log_{10}\left(\frac{\sum_{i=1}^{N} e_i^2}{\sum_{i=1}^{N} s_i^2}\right) \qquad (13)$$

where s_i is the desired signal value at sample i, and $e_i = s_i - \hat{s}_i$. From equation (13), we can see that when we have perfect prediction, i.e. $e_i = 0$ for all i, the $NMSE$ will be $-\infty$ dB. When there is no prediction, i.e. $\hat{s}_i = 0, e_i = s_i$ for all i, the $NMSE$ will be 0 dB.

The result in figure 4 shows that we can get to within 0.5dB prediction performance of the full model by using a 20 parameter model selected using OLS GS. To use the OLS SVD approach to find a 20 parameter subset model with performance comparable to that found using OLS GS, we need to approximate the matrix $\mathbf{X}$ by a rank equal to or greater than 40.

The result shows that subset model performance can be trade for computation complexity.

ACKNOWLEDGEMENTS

We would like to thank Gary Ushaw for providing the Duffing's data used in the simulations. E.S. Chng would like to thank his parents for their understanding and support for the undertaking of the PhD studies.

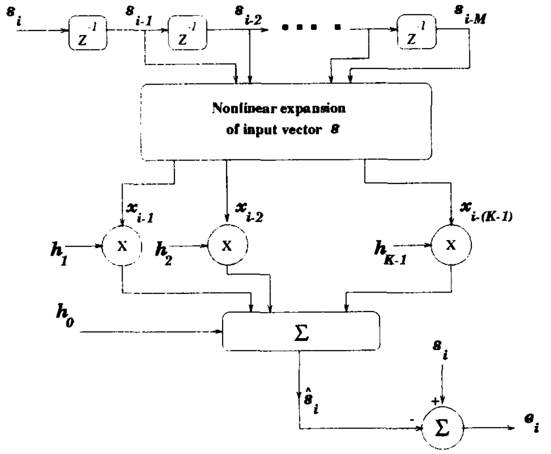

Figure 1: Nonlinear Predictor of Order K

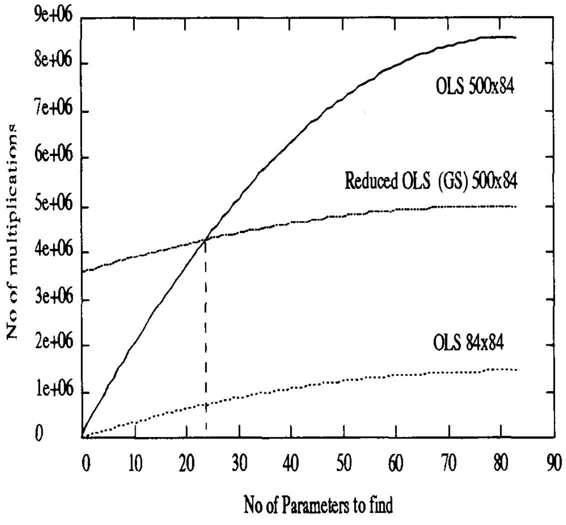

Figure 2: Computation requirement for X matrix of size 500x84

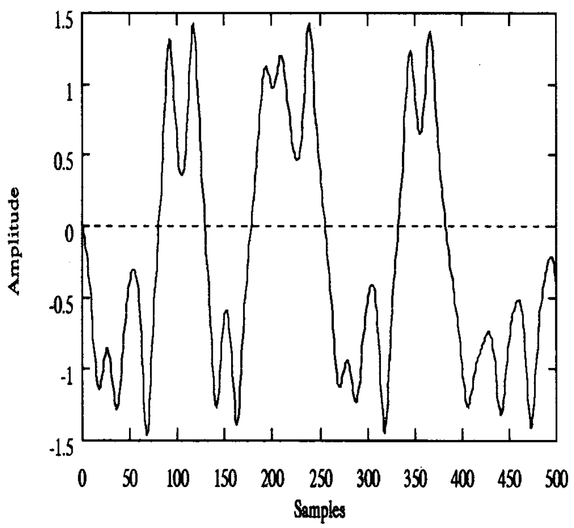

Figure 3: Duffing's chaotic time series

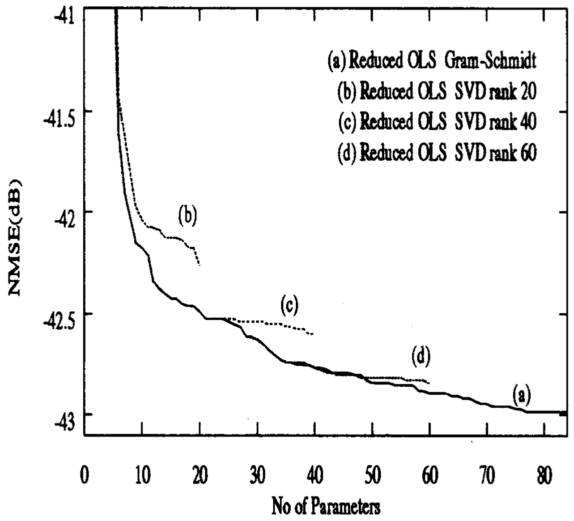

Figure 4: Performance of subset model found using reduced OLS

REFERENCES

[1] S.CHEN, S.A.BILLINGS, and W.LUO, "Orthogonal least squares methods and their application to nonlinear system identification", *Int. J. Control*, vol. 50, pp. 1873-1896, 1989.

[2] S.CHEN, C.F.COWAN, and P.M.GRANT, "Orthogonal least squares learning algorithm for radial basis function networks", *IEEE Trans. Neural Networks*, vol. 2, pp. 302-309, 1991.

[3] R.R.HOCKING, "The analysis and selection of variables in linear regression", *Biometrics*, vol. 32, pp. 1-49, 1976.

[4] A.J.MILLER, *Subset Selection in Regression*, Chapman and Hall, 1990.

[5] M.J.D.POWELL, "Radial basis functions for multivariable interpolation: a review", *Algorithms for Approximation*, pp. 143-167, J.C.MASON and M.G.COX (Eds), Oxford, 1987.

[6] P.ALPER, "A consideration of the discrete Volterra series", *IEEE Trans. AC*, vol. AC-10, pp. 322-327, 1965.

[7] G.H.GOLUB and C.REINSCH, "Singular value decomposition and least squares solutions", *Numerische Math.*, vol. 14, pp. 403-420, 1970.

[8] G.H.GOLUB and C.F.VAN LOAN, *Matrix Computations (2nd Edition)*, The John Hopkins University Press, Baltimore, MD, 1989.

[9] P.COMON and G.H.GOLUB, "Tracking a few extreme singular values and vectors in signal processing", *Proc. IEEE*, vol. 78, pp. 1327-1343, 1990.