Article:
Billings, S.A. and Wei, H.L. (2007) Sparse model identification using a forward orthogonal regression algorithm aided by mutual information. IEEE Transactions on Neural Networks, 18 (1). pp. 306-310. ISSN: 1045-9227

https://doi.org/10.1109/TNN.2006.886356

Sparse Model Identification Using a Forward Orthogonal Regression Algorithm Aided by Mutual Information

Stephen A. Billings and Hua-Liang Wei

***Abstract*—A sparse representation, with satisfactory approximation accuracy, is usually desirable in any nonlinear system identification and signal processing problem. A new forward orthogonal regression algorithm, with mutual information interference, is proposed for sparse model selection and parameter estimation. The new algorithm can be used to construct parsimonious linear-in-the-parameters models.**

***Index Terms*—Model selection, mutual information, orthogonal least squares (OLS), parameter estimation.**

I. Introduction

The central task in learning from data is how to identify a suitable model from the observational data set. One solution is to construct nonlinear models using some specific types of basis functions, aided by various state-of-the-art techniques [1]–[5]. Among the existing sparse modeling techniques, linear-in-the-parameters regression models, which will be considered in this letter, are an important class of representations for nonlinear function approximation and signal processing. A general routine for linear-in-the-parameters modeling often starts by constructing a model term dictionary, whose elements are the candidate model terms. The task of system identification involves two aspects: the selection of the significant model terms and the determination of the number of model terms involved in the final identified model. The objective is to obtain a satisfactory sparse representation that involves only a small number of model terms by making a compromise between the approximation accuracy and the model complexity (model size). Notice that the objective of dynamical modeling is not merely data fitting. In dynamical modeling, the resulting sparse model should fit the observational data accurately, but at the same time the model should be capable of capturing the underlying system dynamics carried by the observational data, so that the resulting model can be used in simulation, analysis, and control studies.

Many approaches have been proposed to address the model structure selection problem; most of these focus on which bases are significant and should be included in the model. The orthogonal least squares (OLS) algorithm [2], [6], [7], which was initiated for nonlinear system identification, has become popular and has been widely used for sparse data modeling. This type of algorithm is simple and is very efficient at producing parsimonious linear-in-the-parameters models with good generalization performance [8]. An advantage of the OLS-type algorithms is that commonly used model selection and regularization techniques, for example the Akaike information criterion (AIC), Bayesian information criterion (BIC), and generalized cross validation (GCV) [9]–[11], can easily be adopted and incorporated into the model structure selection algorithms to yield compact linear-in-the-parameters regression models with good generalization properties [12]–[14].

Manuscript received March 30, 2006; revised July 10, 2006. This work was supported by the U.K. Engineering and Physical Sciences Research Council (EPSRC-UK).

The authors are with the Department of Automatic Control and Systems Engineering, the Univerity of Sheffield, Sheffield, S1 3JD, U.K. (e-mail: s.billings@sheffield.ac.uk; w.hualiang@sheffield.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNN.2006.886356

In the OLS-type algorithms, the criterion that is used to measure the significance of the candidate bases (model terms) is the error reduction ratio (ERR), which is equivalent to the squared correlation coefficient and is similar to the commonly used Pearson correlation function. Experience has shown that the OLS algorithms interfered by the ERR criterion can usually produce a satisfactory sparse model with good generalization performance. The adoption and the domination of the ERR criterion, however, does not exclude other criteria. It follows from practical experience that the selected model subsets are often criterion-dependent.

In this letter, a new criterion, derived from mutual information, is adopted into the OLS algorithm to measure the significance of candidate bases and to interfere with the model subset selection. The motivation of the adoption of a mutual information criterion is based on the following considerations. It is known that the task of modeling from data is generally structure-unknown and the model term dictionary is often predetermined and thus fixed. For this case, the selected model structures are usually criterion-dependent. This implies that the mutual information criterion and the ERR criterion may or may not produce exactly the same model structure given the same modeling problem. The two criteria can be used in parallel, and the performance of the resultant models can then be compared. The model with the better performance will be chosen as the final model. In this manner, the two criteria will complement each other and thus produce a better model.

II. Linear-in-the-Parameters Representation

Consider the identification problem for nonlinear systems given N pairs of input–output observations $\{u(t), y(t)\}_{t=1}^{N}$. Under some mild conditions, a discrete-time nonlinear system can be described by the following nonlinear autoregressive with exogenous inputs (NARX) model [1]

$$y(t) = f(y(t-1), \ldots, y(t-n_y), u(t-1), \ldots, u(t-n_u)) + e(t) \quad (1)$$

where $u(t)$, $y(t)$, and $e(t)$ are the system input, output, and noise variables, n_u and n_y are the maximum lags in the input and output, respectively, and f is some unknown nonlinear mapping. It is generally assumed that $e(t)$ is an independent identical distributed noise sequence.

The central task of system identification is to find a suitable approximator $\hat{f}$ for the unknown function f from the observational data. One solution is to construct nonlinear models using some specific types of basis functions including polynomials, kernel basis functions, and multiresolution wavelets [3]–[6], [15]. Among these existing modeling techniques, linear-in-the-parameters regression models, which will be considered in this letter, is an important class of representations for nonlinear system identification, because compared to nonlinear-in-the-parameters models, linear-in-the -parameters models are simpler to analyze mathematically and quicker to compute numerically.

Let $d = n_y + n_u$ and $\mathbf{x}(t) = [x_1(t), \ldots, x_d(t)]^T$ with

$$x_k(t) = \begin{cases} y(t-k), & 1 \le k \le n_y \\ u(t-(k-n_y)), & n_y+1 \le k \le n_y+n_u. \end{cases} \quad (2)$$

A general form of the linear-in-the-parameters regression model is given as follows:

$$\begin{aligned} y(t) &= \hat{f}(\mathbf{x}(t)) + e(t) = \sum_{m=1}^{M} \theta_m \phi_m(\mathbf{x}(t)) + e(t) \\ &= \boldsymbol{\varphi}^T(t)\boldsymbol{\theta} + e(t) \end{aligned} \quad (3)$$

where M is the total number of candidate regressors, $\phi_m(\mathbf{x}(t))$ $(m = 1, \ldots, M)$ are the model regressors and θ_m are the model parameters, and $\boldsymbol{\varphi}(t) = [\phi_1(\mathbf{x}(t)), \ldots, \phi_M(\mathbf{x}(t))]^T$ and $\boldsymbol{\theta}$ are the associated regressor and parameter vectors, respectively.

III. Mutual Information Interference for Model Structure Selection

In the standard OLS algorithm [2], [6], [7], the significance of candidate model terms is measured using the values of ERR, which is defined as the noncentralized squared correlation coefficient between two associated vectors. This coefficient between two given vectors $\mathbf{x}$ and $\mathbf{y}$ of size N is defined as

$$C(\mathbf{x}, \mathbf{y}) = \frac{(\mathbf{x}^T\mathbf{y})^2}{(\mathbf{x}^T\mathbf{x})(\mathbf{y}^T\mathbf{y})} = \frac{(\sum_{i=1}^{N} x_i y_i)^2}{\sum_{i=1}^{N} x_i^2 \sum_{i=1}^{N} y_i^2}. \quad (4)$$

Similar to the commonly used standard Pearson correlation coefficient, the function in (4) reflects the linear relationship between two vectors $\mathbf{x}$ and $\mathbf{y}$. Both the standard Pearson correlation coefficient and the squared correlation coefficient in (4) have wide application in various fields.

Another useful criterion, derived from mutual information, can be used to measure the relationship of two random variables by calculating the amount of information that the two variables share with each other. Mutual-information-based algorithms have in recent years been widely applied in various areas including feature selection [16]–[20]. In this letter, mutual information will be introduced to form a complementary criterion to the ERR criterion to interfere with the model structure selection procedure.

A. Mutual Information

Following [21], mutual information is defined as follows. Consider two random discrete variables $\mathbf{x}$ and $\mathbf{y}$ with alphabet $\mathcal{X}$ and $\mathcal{Y}$, respectively, and with a joint probability mass function $p(x, y)$ and marginal probability mass functions $p(x)$ and $p(y)$. The mutual information $I(\mathbf{x}, \mathbf{y})$ is the relative entropy between the joint distribution and the product distribution $p(x)p(y)$, given as

$$\begin{aligned} I(\mathbf{x}, \mathbf{y}) &= E\left\{\log\left(\frac{p(\mathbf{x}, \mathbf{y})}{p(\mathbf{x})p(\mathbf{y})}\right)\right\} \\ &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log\left(\frac{p(x, y)}{p(x)p(y)}\right). \end{aligned} \quad (5)$$

The mutual information $I(\mathbf{x}, \mathbf{y})$ is the reduction in the uncertainty of $\mathbf{y}$ due to some knowledge of $\mathbf{x}$ and vice versa. Mutual information provides a measure of the amount of information that one variable shares with another. If $\mathbf{y}$ is chosen to be the system output (the response), and $\mathbf{x}$ is one regressor in a linear model, $I(\mathbf{x}, \mathbf{y})$ can be used to measure the coherency of $\mathbf{x}$ with $\mathbf{y}$ in the model.

B. Model Structure Selection With Interference of Mutual Information

Let $\mathbf{y} = [y(1), \ldots, y(N)]^T$ be a vector of measured outputs at N time instants, and $\boldsymbol{\varphi}_m = [\phi_m(1), \ldots, \phi_m(N)]^T$ be a vector formed by the mth candidate model term, where $m = 1, 2, \ldots, M$. Let $\mathcal{D} = \{\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_M\}$ be a dictionary composed of the M candidate bases. From the viewpoint of practical modeling and identification, the finite dimensional set $\mathcal{D}$ is often redundant. The model term selection problem is equivalent to finding a full dimensional subset $\mathcal{D}_n = \{\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_n\} = \{\boldsymbol{\varphi}_{i_1}, \ldots, \boldsymbol{\varphi}_{i_n}\}$ of n $(n \le M)$ bases, from the library $\mathcal{D}$, where $\boldsymbol{\alpha}_k = \boldsymbol{\varphi}_{i_k}$, $i_k \in \{1, 2, \ldots, M\}$ and $k = 1, 2, \ldots, n$, so that $\mathbf{y}$ can be satisfactorily approximated using a linear combination of $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_n$ as

$$\mathbf{y} = \theta_1\boldsymbol{\alpha}_1 + \cdots + \theta_n\boldsymbol{\alpha}_n + \mathbf{e} \quad (6)$$

or in a compact matrix form

$$\mathbf{y} = \mathbf{A}\boldsymbol{\theta} + \mathbf{e} \quad (7)$$

where the matrix $\mathbf{A} = [\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_n]$ is assumed to be of full column rank, $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_n]^T$ is a parameter vector, and $\mathbf{e}$ is the approximation error. The model structure selection procedure starts from (3). Let $\mathbf{r}_0 = \mathbf{y}$, and

$$\ell_1 = \arg \max_{1 \le j \le M} \{I(\mathbf{r}_0, \boldsymbol{\varphi}_j)\} \tag{8}$$

where $I(\cdot, \cdot)$ is the mutual information defined by (5). The first significant basis can thus be selected as $\boldsymbol{\alpha}_1 = \boldsymbol{\varphi}_{\ell_1}$, and the first associated orthogonal basis can be chosen as $\mathbf{q}_1 = \boldsymbol{\varphi}_{\ell_1}$. Set

$$\mathbf{r}_1 = \mathbf{r}_0 - \frac{\mathbf{r}_0^T \mathbf{q}_1}{\mathbf{q}_1^T \mathbf{q}_1} \mathbf{q}_1. \tag{9}$$

In general, the mth significant model term can be chosen as follows. Assume that at the $(m-1)$th step, a subset $\mathcal{D}_{m-1}$, consisting of $(m-1)$ significant bases, $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_{m-1}$, has been determined, and the $(m-1)$ selected bases have been transformed into a new group of orthogonal bases $\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_{m-1}$ via some orthogonal transformation. Let

$$\mathbf{q}_j^{(m)} = \boldsymbol{\varphi}_j - \sum_{k=1}^{m-1} \frac{\boldsymbol{\varphi}_j^T \mathbf{q}_k}{\mathbf{q}_k^T \mathbf{q}_k} \mathbf{q}_k \tag{10}$$

$$\ell_m = \arg \max_{j \ne \ell_k, 1 \le k \le m-1} \left\{ I\left(\mathbf{r}_{m-1}, \mathbf{q}_j^{(m)}\right) \right\} \tag{11}$$

where $\boldsymbol{\varphi}_j \in \mathcal{D} - \mathcal{D}_{m-1}$, and $\mathbf{r}_{m-1}$ is the residual vector obtained in the $(m-1)$th step. The mth significant basis can then be chosen as $\boldsymbol{\alpha}_m = \boldsymbol{\varphi}_{\ell_m}$ and the mth associated orthogonal basis can be chosen as $\mathbf{q}_m = \mathbf{q}_{\ell_m}^{(m)}$. The residual vector $\mathbf{r}_m$ is given by

$$\mathbf{r}_m = \mathbf{r}_{m-1} - \frac{\mathbf{r}_{m-1}^T \mathbf{q}_m}{\mathbf{q}_m^T \mathbf{q}_m} \mathbf{q}_m. \tag{12}$$

Subsequent significant bases can be selected in the same way step by step. From (12), the vectors $\mathbf{r}_m$ and $\mathbf{q}_m$ are orthogonal, thus

$$\|\mathbf{r}_m\|^2 = \|\mathbf{r}_{m-1}\|^2 - \frac{\left(\mathbf{r}_{m-1}^T \mathbf{q}_m\right)^2}{\mathbf{q}_m^T \mathbf{q}_m}. \tag{13}$$

By respectively summing (12) and (13) for m from 1 to n, yields

$$\mathbf{y} = \sum_{m=1}^{n} \frac{\mathbf{r}_{m-1}^T \mathbf{q}_m}{\mathbf{q}_m^T \mathbf{q}_m} \mathbf{q}_m + \mathbf{r}_n \tag{14}$$

$$\|\mathbf{r}_n\|^2 = \|\mathbf{y}\|^2 - \sum_{m=1}^{n} \frac{\left(\mathbf{r}_{m-1}^T \mathbf{q}_m\right)^2}{\mathbf{q}_m^T \mathbf{q}_m}. \tag{15}$$

Notice that if the function $I(\cdot, \cdot)$ in (8) and (11) is replaced by the squared correlation coefficient defined by (4), the above algorithm then belongs to the class of OLS-type algorithms [2], [6], [7]. The forward orthogonal regression algorithm interfered with mutual information will be referred to as the FOR-MI algorithm. The residual sum of squares, $\|\mathbf{r}_n\|^2$, which is also known as the sum-squared-error, or its variants including the mean-square-error (mse), can be used to form criteria for model selection. There are many criteria used for model selection include the AIC, BIC, and GCV [9]–[11], [13]. One popular version for each of the three criteria is

$$\text{AIC}(n) = N \log[\text{mse}(n)] + 2n \tag{16}$$

$$\text{BIC}(n) = N \log[\text{mse}(n)] + n \log(N) \tag{17}$$

$$\text{GCV}(n) = \left(\frac{N}{N-n}\right)^2 \text{mse}(n) \tag{18}$$

where $\text{mse}(n) = \|\mathbf{r}_n\|^2 / N$.

C. Parameter Estimation

It is easy to verify that the relationship between the selected original bases $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_m$, and the associated orthogonal bases $\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_m$ is given by

$$\mathbf{A}_m = \mathbf{Q}_m \mathbf{R}_m \tag{19}$$

where $\mathbf{A}_m = [\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_m]$, $\mathbf{Q}_m$ is an $N \times m$ matrix with orthogonal columns $\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_m$, and $\mathbf{R}_m$ is an $m \times m$ unit upper triangular matrix whose entries $u_{ij} (1 \le i \le j \le m)$ are calculated during the orthogonalization procedure. The unknown parameter vector, denoted by $\boldsymbol{\theta}_m = [\theta_1, \theta_2, \ldots, \theta_m]^T$, for the model with respect to the original bases [similar to (6)], can be calculated from the triangular equation $\mathbf{R}_m \boldsymbol{\theta}_m = \mathbf{g}_m$ with $\mathbf{g}_m = [g_1, g_2, \ldots, g_m]^T$, where $g_k = \left(\mathbf{r}_{k-1}^T \mathbf{q}_k\right) / \left(\mathbf{q}_k^T \mathbf{q}_k\right)$.

Note that some tricks can be used to avoid selecting strongly correlated model terms. Assume that at the $(m-1)$th step, a subset $\mathcal{D}_{m-1}$, consisting of $m-1$ significant bases $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_{m-1}$ has been determined. Also assume that $\boldsymbol{\varphi}_j \in \mathcal{D} - \mathcal{D}_{m-1}$ is strongly correlated with some bases in $\mathcal{D}_{m-1}$, that is $\boldsymbol{\varphi}_j$ is a linear combination of $\boldsymbol{\alpha}, \ldots, \boldsymbol{\alpha}_{m-1}$. Thus, $\left(\mathbf{q}_j^{(m)}\right)^T \mathbf{q}_j^{(m)} = 0$. In the implementation of the algorithm, the candidate basis $\boldsymbol{\varphi}_j$ will be automatically discarded if $\left(\mathbf{q}_j^{(m)}\right)^T \mathbf{q}_j^{(m)} < \delta$, where δ is a positive number that is sufficiently small. In this way, any severe mullticolinearity or ill-conditioning can be avoided.

A similar algorithm, called maximally informative dimensions (MID), has been proposed in [20], where the main objective was to find, in a high dimensional stimulus space, significant features of sensory stimulus ("inputs") that are most relevant to the measured neural responses ("outputs"). There is some similarity between FOR-MI and MID in that both algorithms employ the mutual information function to measure the dependency between two specified vectors. The implementation of the two algorithms, and the final objectives that the two algorithms aim to achieve, however, are different from each other. The FOR-MI algorithm deals with the linear-in-the-parameters regression problem, and involves a combination of a forward orthogonal regression procedure and the calculation of mutual information. Significant bases are selected in a stepwise way, one at a time. The final objective is to produce a sparse regression model, where both the model structure and the unknown model parameters need to be determined using the OLS algorithm. The MID algorithm, however, is a nonlinear optimization method that uses a combination of gradient ascent and simulated annealing algorithms. The MID algorithm thus involves the calculation of not only the mutual information function itself but also the associated gradient. The MID algorithm aims to find the maximally informative dimensions in an iterative way by increasing the dimensionality until the information is saturated (up to the noise level). The unknown parameters in the model were estimated using some statistical approach.

IV. Example

The magnetosphere is a complex input–output dynamical nonlinear system, where the solar wind and the associated parameters play the role of the inputs and the geomagnetic indices can be considered as the outputs. The *Dst* index is a key parameter to characterize the disturbance of the geomagnetic field in the magnetic storms. Modeling of the Dst index is thus very important for the analysis of the geomagnetic field. Fig. 1 presents 850 data points of the measurements for the solar wind parameter, VBs, and the Dst index of this dynamical process. The solar wind parameter VBs was treated to be the system

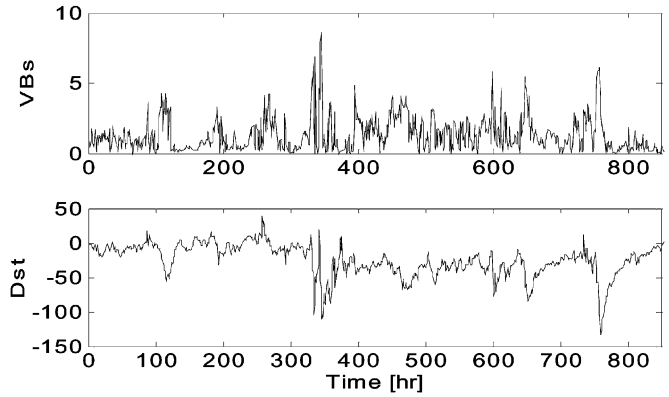

Fig. 1. Measurements of the solar wind parameter VBs and the Dst index for the terrestrial magnetospheric system.

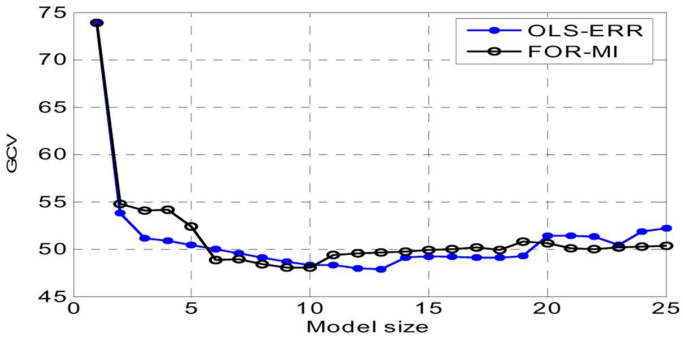

Fig. 2. GCV versus the number of regressors selected from the candidate model terms using both the OLS-ERR and the FOR-MI algorithms.

input, and the Dst index was treated to be the system output. The objective here was to identify a mathematical model to forecast the future behavior of the Dst index. The data set was partitioned into two parts. The first 500 points were used for model estimation and the remaining 350 points were used for model performance test.

The polynomial NARX model was employed to describe the magnetospheric system. Denote the system input and output using $u(t) = VBs(t)$ and $y(t) = Dst(t)$, respectively. The "input" vector for the model was chosen to be $\mathbf{x}(t) = [x_1(t), \ldots, x_{12}(t)]$ $[y(t-1), \ldots, y(t-5), u(t-1), \ldots, u(t-7)]$, and the candidate regressors $\phi_m(\mathbf{x}(t))$ in (3) are of the form $x_{j_1}^{i_1}(t)x_{j_2}^{i_2}(t)x_{j_3}^{i_3}(t)$, where $x_{j_k}^{i_k}(t) \in \{x_1(t), \ldots, x_{12}(t)\}$ $(k = 1, 2, 3)$, $i_k \in \{0, 1, 2, 3\}$, $0 \leq i_1 + i_2 + i_3 \leq 3$, and $j_k \in \{1, 2, \ldots, 12\}$. Thus, a total of 455 candidate model terms are involved.

Both the OLS-ERR algorithm and the FOR-MI algorithm were applied to the 455 candidate model terms. The associated criterion GCV is shown in Fig. 2, which suggests that the number of model terms included in the OLS-ERR and the FOR-MI identified NARX models should be 13 and 10, respectively. Note that the AIC and BIC produce similar results for this data set, and the curves for AIC and BIC were thus omitted. The selected model terms, along with the associated parameter estimates are reported in Table I, where model terms are listed in the order of their significance (the order that the terms entered into the model). The FOR-MI identified model for this data set is in structure simpler than the model produced by the OLS-ERR algorithm.

The performance of the two identified NARX models was inspected and compared by calculating both short-term and long-term predictions, over the validation data set. The performance of one-step-ahead (OSA) predictions and model predicted (MPO) outputs, calculated from the OLS-ERR and the FOR-MI identified models are presented

TABLE I
SELECTED MODEL STRUCTURE BY OLS-ERR AND FOR-MI

No	OLS-ERR		FOR-MI	
	Term	Parameter	Term	Parameter
1	y(t-1)	0.807507	y(t-1)	0.775816
2	u^2(t-1)	-1.822126	u(t-1)	-2.572108
3	y(t-4)	0.143209	y(t-4)	0.166492
4	u^3(t-1)	0.137199	y(t-2)	0.006133
5	u(t-5)	-0.293282	u(t-5)	1.004308
6	u(t-1)u(t-2) u(t-4)	-0.254211	u^2(t-1)	-0.399763
7	u(t-2)u(t-5)	1.026867	y(t-3)	-0.040214
8	u(t-2)u(t-5) u(t-6)	-0.193897	u(t-1)u(t-3) × u(t-4)	-0.113742
9	y^2(t-3)u(t-6)	0.000428	u(t-2)	0.788903
10	u(t-6)	0.913923	u(t-1)u(t-3)	0.252518
11	u(t-1)u(t-4)	0.765627		
12	y(t-2)u(t-1) u(t-5)	0.009401		
13	y(t-1)u^2(t-1)	-0.006575		
	Run time: 12.93s. MSE(OSA),over validation data set: 30.39. MSE(MPO), over the validation dataset: 156.64.		Run time: 22.51s. MSE(OSA),over validation data set: 29.53. MSE(MPO),over the validation dataset: 153.96	

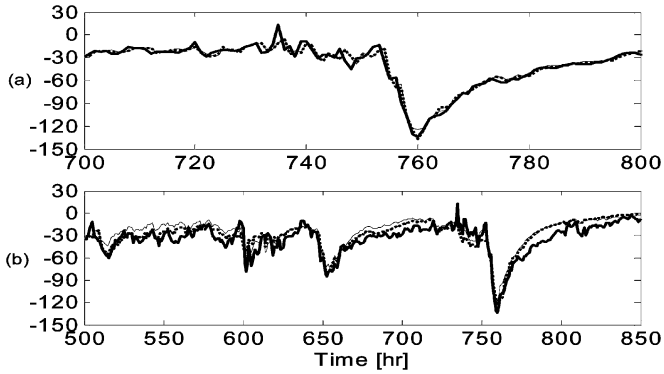

Fig. 3. Prediction performance of the OLS-ERR and the FOR-MI identified models, over the validation data set, for the terrestrial magnetospheric system. (a) OSA predictions. (b) MPO. Thick solid line shows the measurements, thin solid line shows predictions from the OLS-ERR identified model, and thick dotted line shows predictions from the FOR-MI identified model.

in Table I and Fig. 3. Clearly, if mse is used as the criterion to measure the model performance, the FOR-MI identified model for this data set will be prior to the model produced by the OLS-ERR algorithm.

V. CONCLUSION

Sparse modeling involves the determination of significant bases. ERR is an efficient index to measure the significance of candidate regressors in the widely used OLS-type algorithms for nonlinear model structure selection. The dominant adoption of ERR, however, does not exclude other criteria. It is observed that the selected model subsets are often criterion-dependent, that is, the OLS algorithms interfered with by different criteria may select different significant bases and thus produce different model subsets. Motivated by this observation, the new FOR-MI algorithm has been introduced as a complementary approach to the commonly used least-squares-type algorithms. Using the two criteria in a modeling problem may or may not produce exactly the same model structure. But by inspecting and comparing the performance of the resulting models, a more accurate

sparse representation can often be obtained. In this way, the accuracy of the identified sparse model will be improved compared with results based on any one single criterion. Notice, however, that the fact that the FOR-MI algorithm is superior to the OLS-ERR algorithm for the given szexample does not mean that FOR-MI is always superior to OLS-ERR for all cases. Conditions, under which one algorithm outperforms the other, or vice versa, have not been determined, and that is why we suggest using the two algorithms in parallel.

ACKNOWLEDGMENT

The authors would like to thank Dr. M. A. Balikhin, the University of Sheffield, Sheffield, U.K., for providing the magnetosphere data.

REFERENCES

[1] I. J. Leontaritis and S. A. Billings, "Input-output parametric models for non-linear systems," *Int. J. Control*, vol. 41, no. 2, pp. 303–344, 1985.
[2] S. A. Billings, S. Chen, and M. J. Korenberg, "Identification of MIMO non-linear systems suing a forward regression orthogonal estimator," *Int. J. Control*, vol. 49, no. 6, pp. 2157–2189, Jun. 1989.
[3] S. Chen and S. A. Billings, "Neural networks for nonlinear system modeling and identification," *Int. J. Control*, vol. 56, no. 2, pp. 319–346, Aug. 1992.
[4] V. Cherkassky and F. Mulier, *Learning from Data*. New York: Wiley, 1998.
[5] C. J. Harris, X. Hong, and Q. Gan, *Adaptive Modelling, Estimation and Fusion From Data: A Neurofuzzy Approach*. Berlin, Germany: Springer-Verlag, 2002.
[6] M. Korenberg, S. A. Billings, Y. P. Liu, and P. J. McIlroy, "Orthogonal parameter-estimation algorithm for non-linear stochastic-systems," *Int. J. Control*, vol. 48, no. 1, pp. 193–210, Jul. 1988.
[7] S. A. Billings, M. J. Korenberg, and S. Chen, "Identification of non-linear output-affine systems using an orthogonal least-squares algorithm," *Int. J. Syst. Sci.*, vol. 19, no. 8, pp. 1559–1568, Aug. 1988.
[8] S. Chen, X. Hong, and C. J. Harris, "Sparse kernel regression modeling using combined locally regularized orthogonal least squares and D-optimality experimental design," *IEEE Trans. Autom. Control*, vol. 48, no. 6, pp. 1029–1036, Jun. 2003.
[9] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Autom. Control*, vol. 19, no. 6, pp. 716–723, Dec. 1974.
[10] G. Schwarz, "Estimating the dimension of a model," *Ann. Statist.*, vol. 6, pp. 461–464, 1978.
[11] G. H. Golub, M. Heath, and G. Wahha, "Generalized cross-validation as a method for choosing a good ridge parameter," *Technometrics*, vol. 21, pp. 215–223, 1979.
[12] I. J. Leontaritis and S. A. Billings, "Model selection and validation methods for nonlinear-systems," *Int. J. Control*, vol. 45, no. 1, pp. 311–341, Jan. 1987.
[13] M. J. L. Orr, "Regularization in the selection of radial basis function centers," *Neural Comput.*, vol. 7, no. 3, pp. 606–623, May 1995.
[14] S. Chen, E. S. Chng, and K. Alkadhimi, "Regularized orthogonal least squares algorithm for constructing radial basis function networks," *Int. J. Control*, vol. 64, no. 5, pp. 829–837, Jul. 1996.
[15] S. A. Billings and H. L. Wei, "A new class of wavelet networks for nonlinear system identification," *IEEE Trans. Neural Netw.*, vol. 16, no. 4, pp. 862–874, Jul. 2005.
[16] R. Battiti, "Using mutual information for selecting features in supervised neural-net learning," *IEEE Trans. Neural Netw.*, vol. 5, no. 4, pp. 537–550, Jul. 1994.
[17] G. L. Zheng and S. A. Billings, "Radial basis function networks configuration using mutual information and the orthogonal least squares algorithm," *Neural Netw.*, vol. 9, pp. 1619–1637, Dec. 1996.
[18] V. Sindhwani, S. Rakshit, D. Deodhare, D. Erdogmus, and J. C. Principe, "Feature selection in MLPs and SVMs based on maximum output information," *IEEE Trans. Neural Netw.*, vol. 15, no. 4, pp. 937–948, Jul. 2004.
[19] T. W. S. Chow and D. Huang, "Estimating optimal feature subsets using efficient estimation of high- dimensional mutual information," *IEEE Trans. Neural Netw.*, vol. 16, no. 1, pp. 213–224, Jan. 2005.
[20] T. Sharpee, N. C. Rust, and W. Bialek, "Analyzing neural responses to nature signals: maximally informative dimensions," *Neural Comput.*, vol. 16, no. 2, pp. 223–250, Feb. 2004.
[21] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

New Delay-Dependent Stability Criteria for Neural Networks With Time-Varying Delay

Yong He, Guoping Liu, and D. Rees

***Abstract*—In this letter, a new method is proposed for stability analysis of neural networks (NNs) with a time-varying delay. Some less conservative delay-dependent stability criteria are established by considering the additional useful terms, which were ignored in previous methods, when estimating the upper bound of the derivative of Lyapunov functionals and introducing the new free-weighting matrices. Numerical examples are given to demonstrate the effectiveness and the benefits of the proposed method.**

***Index Terms*—Delay-dependent, neural networks (NNs), time-varying delay, linear matrix inequality (LMI), stability.**

I. INTRODUCTION

Neural networks (NNs) have been extensively studied over the past few decades and have found many applications in a variety of areas, such as signal processing, pattern recognition, static image processing, associative memory, and combinatorial optimization. Although considerable effort has been devoted to analyzing the stability of NNs without a time delay, in recent years, the stability of delayed NNs has also received attention [1]–[24] since time delay is frequently encountered in NNs, and it is often a source of instability and oscillations in a system. The stability criteria for delayed NNs can be classified into two categories, namely, delay-independent [1], [3], [5]–[12], [14], [15], [20] and delay-dependent [4], [16], [19], [24]. Since delay-independent criteria tend to be conservative, especially when the delay is small, much attention has been paid to the delay-dependent type.

As for the delay-dependent stability criteria, the free-weighting matrix approach proposed in [25]–[27] is very effective for time-delay systems since the bounding techniques on some cross-product terms are not involved in this approach. In [19], delay-dependent stability criteria are established for NNs with multiple time-varying delays using the free-weighting matrix approach. On the other hand, an alternative criterion is derived for NNs with single time-varying delay in [24] by introducing a new Lyapunov functional which is similar to [28]. However, there is room for further investigation when estimating the upper bound of the derivative of Lyapunov functional for systems with time-varying delay. For example, in [19] and [24], the derivative of $\int_{-h}^{0}\int_{t+\theta}^{t}\dot{z}^T(s)Z\dot{z}(s)\,ds\,d\theta$ is often estimated as $h\dot{z}^T(t)Z\dot{z}(t) - \int_{t-d(t)}^{t}\dot{z}^T(s)Z\dot{z}(s)\,ds$ and the term $-\int_{t-h}^{t-d(t))}\dot{z}^T(s)Z\dot{z}(s)\,ds$ is ignored, which may lead to considerable conservativeness.

In this letter, a new method that introduces the new free-weighting matrices is proposed to estimate the upper bound of the derivative of

Manuscript received June 22, 2006; revised September 10, 2006. This work was supported in part by the U.K. Leverhulme Trust, the National Science Foundation of China under Grants 60574014 and 60528002, and by the Doctor Subject Foundation of China under Grant 20050533015.

Y. He is with the Faculty of Advanced Technology, University of Glamorgan, Pontypridd CF37 1DL, U.K.and also with the School of Information Science and Engineering, Central South University, Changsha 410083, China (e-mail: heyong08@yahoo.com.cn).

G. Liu is with the Faculty of Advanced Technology, University of Glamorgan, Pontypridd CF37 1DL, U.K. and also with the Complex Systems and Intelligence Science (CSIS) Lab, Institute of Automation, Chinese Academy of Sciences, Beijing 100080, China.

D. Rees is with the Faculty of Advanced Technology, University of Glamorgan, Pontypridd CF37 1DL, U.K.

Digital Object Identifier 10.1109/TNN.2006.888373