Stability Orthogonal Regression for System Identification

DOI:
10.1016/j.sysconle.2018.05.002

Document Version
Accepted author manuscript

Stability Orthogonal Regression for System Identification

Xiaoquan Tang

School of Automation, Huazhong University of Science and Technology, Wuhan, 430074, China

Long Zhang

School of Electrical and Electronic Engineering, University of Manchester Manchester, M13 9PL, United Kingdom

Abstract

Variable selection methods have been widely used for system identification. However, there is still a major challenge in producing parsimonious models with optimal model structures as popular variable selection methods often produce suboptimal model with redundant model terms. In the paper, stability orthogonal regression (SOR) is proposed to build a more compact model with fewer or no redundant model terms. The main idea of SOR is that multiple intermediate models are produced by orthogonal forward regression (OFR) using sub-sampling technique and then the final model is a combination of these intermediate model terms but does not include infrequently selected terms. The effectiveness of the proposed methods is analysed in theory and also demonstrated using two numerical examples in comparison with some popular algorithms.

Keywords: Orthogonal forward regression, Stability selection, Stability orthogonal regression, System identification

1. Introduction

The main objective of system identification is to establish a mathematical model for a system using system input and output observations. The widely used linear models include auto-regressive with eXogenous input (ARX), auto-

regressive moving average with eXogenous input (ARMAX), Box Jenkins and state space models [1]. If the performance of the linear models is not satisfied, the nonlinear ARX (NARX) is an alternative option.

The most popular structure for the NARX model is a sum of nonlinear functions whose parameters are given a priori. The nonlinear functions with pre-set parameters are also referred to as terms in some literatures [2]. However, the prefixed values for these nonlinear parameters are not optimal, and therefore their corresponding nonlinear functions are often redundant. The simple option is to use ordinary least square methods to estimate all the coefficients of these nonlinear functions. For these redundant functions, their correct coefficients should be zeros. However, due to the noise effect and correlations between redundant and important functions, the estimated coefficients of redundant functions are often not zeros. In other words, the redundant functions are included into the estimated models, leading to unsatisfactory model performance. Alternatively, regularized least squares algorithms, such as l_1 or l_2 regularization can be used to penalize the coefficients and therefore to produce more compact models. For regularized methods, some additional parameters need to be tuned carefully [3].

Another popular option of building a nonlinear model is to select representative nonlinear functions and then determine their coefficients. The process for selecting nonlinear functions is also referred to as subset or term selection [2]. The predetermined model set may include a huge number of terms and most of terms should not be included into the final model. Therefore, it is important to determine which terms to be included into the final model. The principle of subset selection is to build a parsimonious model with as few redundant model terms as possible [2]. The ideal case is to produce an optimal model without any redundant model term. The orthogonal forward regression (OFR) is one of the most well known subset selection methods. A good review for these existing term selection and their modifications can be found in literatures [1, 3, 4, 5]. This paper focuses on the subset selection which is a hard problem in the NARX model [6].

The OFR and their modifications have been successfully used in many ap-

plications and well studied within system identification community. In most applications, they can produce a parsimonious model. However, a suboptimal solution can be obtained in some applications, in particular when the following conditions happen:

- **Insufficient input-output data and non-persistent excitation**: Most existing methods are based on least square principle and they are asymptotically optimal. The training data length is too short to incorporate all the useful information, which may lead to an inaccurate model. Non-persistent input is another proper problem relating to system input data. Non-persistent excitation can cause regression matrix being ill-conditioning, which may result in poor estimation of the parameters and also poor long term prediction [7].

- **Highly correlated terms**: The adjacent lagged system inputs or outputs could be very similar in their values and therefore their corresponding nonlinear terms are highly correlated, which causes difficulty in selecting the correct terms from the similar alternatives.

- **No optimal criteria**: Most methods have to rely on the information based criteria to determine the model structure. Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) and other statistical criteria are popular options [1]. These criteria are simple to use but they may not produce optimal model sizes.

- **Mixed problems**. The above problems can be coupled, which makes it more difficult to build an accurate model, especially for nonlinear systems.

The above reasons can cause sub-optimal model structure with redundant terms. Generally, there are two types of redundant terms. The first type is that the terms are highly correctly with the useful terms and they represent the useful terms when entering the models. The second type is that the terms can generally reduce the model error but tend to approximating the noise. For sparse modeling problems where the number of the useful terms is much smaller

than that of the whole candidate terms, the second type, noise terms, could be more serious than the first type in terms of their number in the final model. Since the noise in the input-output data is usually unknown and very hard to estimate with a good accuracy, it is difficult to choose a proper stopping criterion or threshold to control the number of noise terms. Further, system identification usually use random data as the system input. When repeating the modelling using different input-output data, the models could be significantly different in terms of the number of redundant or noise terms even if the model stopping criteria or threshold is fixed. In other words, one main difficulty in choosing the model stopping criterion using OFR in practice is to limit the model redundant model terms. If a good criterion or threshold is chosen, the resultant model has fewer or no redundant terms. If not chosen well, the model could has a large number of redundant terms. Another difficulty is that, when repeating the modelling process but just using different input-output data, a number of different models may be generated and it is hard to determine which model should be chosen as the final one.

In this paper, the stability orthogonal regression (SOR) is proposed to build a more parsimonious model by reducing the redundant model terms. A main advantage of SOR is that it can produce an improved model with fewer redundant terms than the original OFR method, and further it may provide the chance to produce an optimal model without any redundant terms. This is achieved by introducing the stability selection scheme into the OFR method. The stability selection was introduced in [8] and mainly aims to produce a stable model with minimal redundant terms. The main principle of stability selection is that it produces multiple intermediate models using sub-sampling techniques. Then the final model is consistent of the most frequently selected terms in the intermediate models.

This paper is organized by starting to introduce the basic including the NARX model and OFR method, then propose the SOR method and analyze its properties in theory, followed by numerical examples.

2. Basics

2.1. NARX model

The linear-in-the-parameters NARX model can be written as the matrix form given as follow:

$$\mathbf{y} = \mathbf{P}\boldsymbol{\Theta} + \boldsymbol{\Xi} \tag{1}$$

where $\mathbf{y} = [y(1), \ldots, y(N)]^T$ is the output vector, $\boldsymbol{\Theta} = [\theta_1, \ldots, \theta_M]^T$ is the weight vector, $\boldsymbol{\Xi} = [e(1), \ldots, e(N)]^T$ is the residual vector. The matrix $\mathbf{P}$ is the whole candidate terms given by $\mathbf{P} = [\mathbf{p}_1, \ldots, \mathbf{p}_M]$, which is an $N \times M$ matrix with $\mathbf{p}_i = [p_i(1), \ldots, p_i(N)]^T$.

The main objective of the subset or term selection is to select the useful terms $\mathbf{P}_m = [\mathbf{p}_{i_1}, \ldots, \mathbf{p}_{i_m}]$ from the whole candidate term pool $\mathbf{P}$, where m denotes the number of selected terms and $[i_1, ..., i_m]$ are indexes. Then the coefficients of the selected terms can be written as $\boldsymbol{\Theta}_m = [\theta_{i_1}, \ldots, \theta_{i_m}]$. Using orthogonal least squares (OLS) method, equation (1) can be factorized as

$$\mathbf{y} = \mathbf{W}\mathbf{A}\boldsymbol{\Theta} + \boldsymbol{\Xi} \tag{2}$$

here matrix A is an $M \times M$ unit upper triangular matrix. $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_M]$ is an $N \times M$ matrix with orthogonal columns that satisfies $\mathbf{W}^T\mathbf{W} = diag[w_i^T w_i]$. For brief, equation (2) can be rewritten as

$$\mathbf{y} = \mathbf{W}\mathbf{A}\boldsymbol{\Theta} + \boldsymbol{\Xi} = \mathbf{W}\mathbf{g} + \boldsymbol{\Xi} \tag{3}$$

where $\mathbf{g} = [g_1, g_2, ..., g_M]^T = \mathbf{A}\boldsymbol{\Theta}$ is the orthogonal weight vector.

2.2. Orthogonal forward regression (OFR)

OFR is one of most well-known term selection methods and it mainly involves a series of orthogonal composition using OLS method. OFR begins with an empty model without any terms in it and then gradually builds a model by adding one term that gives the largest decrease or increase in the cost function at a time until the model performance is met under some stopping criterion.

The first important task in OFR is to choose a cost function for determining which term is included to a resultant model. The error reduction ratio (ERR) is a popular criterion for term selection, and its value is derived from the sum of squares of the model output. More specifically, the sum of squares of the output variables $\mathbf{y}$ is

$$\mathbf{y}^T\mathbf{y} = \sum_{i=1}^{m} g_i^2 \mathbf{w}_i^T \mathbf{w}_i + \mathbf{\Xi}^T \mathbf{\Xi} \tag{4}$$

It can be seen that $g_i^2 \mathbf{w}_i^T \mathbf{w}_i$ is the contribution of the term $\mathbf{w}_i$ to the sum of squares of the output. The ERR value due to $\mathbf{w}_i$ is defined as [2]

$$[err]_i = g_i^2 \mathbf{w}_i^T \mathbf{w}_i / (\mathbf{y}^T\mathbf{y}) = g_i \mathbf{w}_i^T \mathbf{y} / (\mathbf{y}^T\mathbf{y}). \tag{5}$$

The details of the OFR procedure using the ERR criterion are summarized as follows [2, 9]:

At the kth step, for $1 \le i \le M, i \ne i_1, ..., i \ne i_{k-1}$ the following procedure are calculated:

$$\left.\begin{array}{l} if \quad k = 1 \\ \qquad \mathbf{w}_1^{(i)} = \mathbf{p}_i \\ else \\ \quad a_{jk}^{(i)} = \mathbf{w}_j^T \mathbf{p}_i / \mathbf{w}_j^T \mathbf{w}_j, 1 \le j < k \\ \quad \mathbf{w}_k^{(i)} = \mathbf{p}_i - \sum_{j=1}^{k-1} a_{jk}^{(i)} \mathbf{w}_j \end{array}\right\} \tag{6}$$

and

$$\left.\begin{array}{l} g_k^{(i)} = (\mathbf{w}_k^{(i)})^T \mathbf{y} / (\mathbf{w}_k^{(i)})^T \mathbf{w}_k^{(i)}, \\ err_k^{(i)} = g_k^{(i)} (\mathbf{w}_k^{(i)})^T \mathbf{y} / \mathbf{y}^T \mathbf{y} \end{array}\right\} \tag{7}$$

The largest ERR value is calculated using $err_k^{(i_k)} = \max\left\{err_k^{(i)}, i \ne i_1, ..., i_{k-1}\right\}$ and the term related to the number i_k is used as

$$\left.\begin{array}{l} \mathbf{w}_k = \mathbf{w}_k^{(i_k)} = \mathbf{p}_{i_k} - \sum_{j=1}^{k-1} a_{jk}^{(i_k)} \mathbf{w}_j \\ g_k = \mathbf{w}_k^T \mathbf{y} / \mathbf{w}_k^T \mathbf{w}_k \end{array}\right\} \tag{8}$$

This procedure will be terminated at the mth step when

$$1 - \sum_{k=1}^{m} err_k^{(i_k)} < \rho \tag{9}$$

where $\rho \in [0, 1]$ is a preset tolerance. Alternatively, information based criteria can be used to stop the selection, for example

$$AIC = Nlog\frac{1}{N}\boldsymbol{\Xi}^T\boldsymbol{\Xi} + 2m \tag{10}$$

is minimized or under a preset threshold. The model parameters $\boldsymbol{\Theta}$ can then be computed with backward substitution

$$\left.\begin{aligned} &\theta_m = g_m \\ &\theta_j = g_j - \sum_{k=j+1}^{m} a_{jk}\theta_k, j = m-1, ...1 \end{aligned}\right\} \tag{11}$$

3. Stability orthogonal regression (SOR)

OFR is a computationally efficient subset selection algorithm. However, the resultant model obtained by OFR may have some redundant model terms in some applications under the aforementioned conditions. In the present work, the SOR is proposed to reduce the redundant terms. This is achieved by introducing the stability selection scheme into the OFR method. Stability selection was proposed in [8] as a general technique and it mainly aims to aid existing subset selection methods to produce a parsimonious model with minimal redundant terms. The core idea of stability selection is to produce multiple intermediate models using sub-sampling techniques. Then the final model is consistent of the most frequently selected terms in the intermediate models. The redundant model terms can be included into the intermediate models but their included frequencies are much less than that of the important terms. Therefore, the redundant terms can be excluded from the final model due to their low selection probability. This paper borrows the important finding in the stability selection to improve the model sparseness of the well-known OFR method.

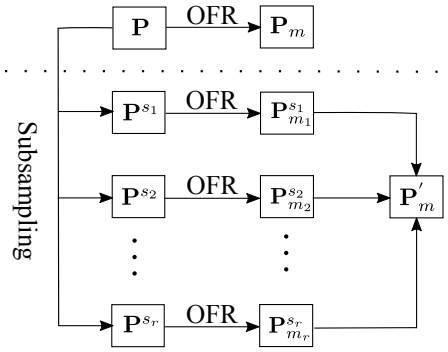

Figure 1: The idea of SOR

To illustrate procedure of the SOR well, the comparison with original OFR is shown in Fig. 1. The OFR method uses all the available training data $\mathbf{P}^s$ to produce a single model $\mathbf{P}_m$. However, SOR produces multiple intermediate models using sub-sampling data. More specifically, a random sub-sampling $\mathbf{P}^{s_1}$ from the whole data $\mathbf{P}^s$ is used for building an intermediate model $\mathbf{P}^{s_1}_{m_1}$ via the OFR method. The above process repeats until multiple, say r, intermediate models $[\mathbf{P}^{s_1}_{m_1}, \mathbf{P}^{s_2}_{m_2}, \ldots, \mathbf{P}^{s_r}_{m_r}]$ are produced. Then count how many times each term has been selected in the whole r models. Suppose $sn(\mathbf{p}_i), i = 1, \ldots, M$, is the times of the ith term being selected among r models. The selecting probability can then be defined as the selecting times over the r repetitions, i.e.

$$sf(\mathbf{p}_i) = \frac{sn(\mathbf{p}_i)}{r} \tag{12}$$

The largest number for $sn(\mathbf{p}_i)$ is r, which means the term $\mathbf{p}_i$ have been selected by all the r models. The smallest number for $sn(\mathbf{p}_i)$ is 0, which indicates the term $\mathbf{p}_i$ has not been selected by any model. If the maximal and minimal numbers are divided by the repetition times, the term selection probability range $sf(\mathbf{p}_i) \in [0, 1]$ can be obtained.

The final model $\mathbf{P}'_m$ includes the terms whose selecting probability is bigger than a preset threshold λ, which is given by

$$\mathbf{P}'_m = \{\mathbf{p}_i, \quad \text{where } sf(\mathbf{p}_i) > \lambda, i = 1, \ldots, M\} \tag{13}$$

Here, the sub-sampling scheme and selecting threshold have to be determined in advance. As $\mathbf{P}^s$ is an N-row matrix, the random sub-sampling, say $N_s = N/2$ out of N rows, are recommended in the literature [8]. As the re-sampling has to be repeated r times, the repetition number r is chosen as a fixed value. Empirically, the value of $r = 100$ have been shown to be sufficient in many cases. Alternatively, the trial-and-error method can be used for choosing the number of subsamplings. This can be achieved by starting from a fixed number, say 100, and gradually increasing the total subsamplings. The procedure stops when increasing subsamplings do not change the resultant models. Furthermore, the threshold value that is another tunable parameter within $[0, 1]$. However, it was shown in the literature [8] that its impact on the final model is rather small. When values are in the range of, say $[0.6, 0.9]$, the final models tend to be very similar. In a words, these tunable parameters are not critical to model performances and can be easily chosen.

To make the concept of stability selection clear, a simple example is presented here. Suppose five model terms $[\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3, \mathbf{p}_4, \mathbf{p}_5]$ are given using priori knowledge. The first task is to select the important terms from the five available ones. Here, five intermediate models are produced using different training data which are obtained from random sub-sampling. The resultant intermediate models and their terms are shown in Table 1. Then the second task is to count how many times each term has been selected and compute the term selection probabilities. The results are shown in Table 2. The terms with high selection

Table 1: Intermediate models of the simple example

	Model	Terms
Model 1	$\mathbf{P}_{m_1}^{s_1}$	$[\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_5]$
Model 2	$\mathbf{P}_{m_2}^{s_2}$	$[\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_4]$
Model 3	$\mathbf{P}_{m_3}^{s_3}$	$[\mathbf{p}_1, \mathbf{p}_2]$
Model 4	$\mathbf{P}_{m_4}^{s_4}$	$[\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_4]$
Model 5	$\mathbf{P}_{m_5}^{s_5}$	$[\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_4]$

Table 2: Term selection probability of the simple example

Term	$\mathbf{p}_1$	$\mathbf{p}_2$	$\mathbf{p}_3$	$\mathbf{p}_4$	$\mathbf{p}_5$
Times	5	5	0	3	1
Probability	1	1	0	0.6	0.2

probabilities will be included into the final model. In this example, if a selecting threshold is chosen as 0.6, terms $\mathbf{p}_1$, $\mathbf{p}_2$ and $\mathbf{p}_4$ can be included into the final while $\mathbf{p}_3$ and $\mathbf{p}_5$ are excluded due to their low selection probabilities.

4. Theoretical properties and discussions

It is worth pointing out that SOR may not be an optimal method and it is not guaranteed to produce a perfect model without any redundant terms. In other words, SOR can include some redundant terms. The number of redundant terms which are also referred to as falsely selected terms in the final model can be bounded, which is reported in [8]. To analyze the number of falsely selected terms, we give several following definitions.

Definition 1: $S = \{k : \theta_k \neq 0\}$ is the set of useful model term index and $Z = \{k : \theta_k = 0\}$ is the set of redundant model term index in the true model.

Definition 2: Given a threshold ρ (that determines when the procedure of OFR stops), suppose the resultant model is represented by $\widehat{S}^\rho$. A set of threshold, $\Lambda = \{\rho : \rho \in \mathbb{R}^+\}$, can be used to generate a set of models, $\widehat{S}^\Lambda = \cup_{\rho\in\Lambda}\widehat{S}^\rho$. Let q be the average of the model size (the number of model terms), namely, $q = E(|\widehat{S}^\Lambda|)$. Further, suppose $\widehat{S}^\Lambda$ is produced using the data sample I that is a random subsample of $\{1, \cdots, N\}$ with the length of data being $N/2$, then we rewrite $E(|\widehat{S}^\Lambda|) = E(|\widehat{S}^\Lambda(I)|)$.

Definition 3: Assume the size of random subsample I is $N/2$. The probability of model set $K \subseteq \{1, \cdots, M\}$ being selected in set $\widehat{S}^\rho(I)$ is $\widehat{\Pi}_K^\rho = P(K \subseteq \widehat{S}^\rho(I))$, where the probability P is with respect to random subsampling.

Definition 4: Assume I_1 and I_2 are two random subsets of $\{1, \cdots, N\}$, the size of the two subsets are $N/2$ and $I_1 \cap I_2 = \emptyset$. Then the simultaneously selected

set is defined as $\widehat{S}^{sim,\rho} = \widehat{S}^{\rho}(I_1) \cap \widehat{S}^{\rho}(I_2)$.

Definition 5: For any set $K \subseteq \{1, \cdots, M\}$, the simultaneous selection probability is defined as $\widehat{\Pi}_K^{sim,\rho} = P(K \subseteq \widehat{S}^{sim,\rho})$, where probability P is with respect to random sample splitting.

Variable selection is conventionally addressed by selecting one model from the set of models

$$\{\widehat{S}^{\rho}; \rho \in \Lambda\} \tag{14}$$

here Λ is also the set of predetermined threshold parameter ρ. The true or optimal model may not be one element of set (14). The main reason may be that sometimes, even only with small difference on ρ which may result in quite different models for system identification. Therefore, if inappropriate ρ being included in the set Λ, obtaining an optimal model may be hard.

With stability selection, the final model includes those frequently selected terms in the intermediate models which are obtained using subsampling technique, rather than directly choose one model from set (14). Using the predetermined threshold λ with $0 \leq \lambda \leq 1$, the resultant model selected by stability selection can be written as

$$\mathbf{P}_m^{'} = \{k : \max_{\rho \in \Lambda} \widehat{\Pi}_k^{\rho} \geq \lambda\} = \{k : sf(\mathbf{p}_k) \geq \lambda\} = \widehat{S}^{SS} \tag{15}$$

It is worth noting that k represents a single model, $k = 1, \cdots, M$ and it is different from K that is a model set with $K \subseteq \{1, \cdots, M\}$. Only the terms with a high selection probability can be selected into the final model while the one with a relatively low selection probability will be excluded. However, some redundant terms with a low selection probability can still be chosen into the model as long as their selecting probability is larger than λ.

Using the above definitions, the falsely selected terms can be written as $V = \widehat{S}^{SS} \cap Z$. If $\lambda \in (0.5,1]$, the expected amount of falsely selected terms can be bounded by

$$E(|V|) \leq \frac{1}{2\lambda - 1} \frac{q^2}{M} \tag{16}$$

which is also called error control. The bounded theory is proved as follows.

Lemma 1 (Lower bound for simultaneous selection probability) [8]: For any set $K \subseteq \{1, \cdots, M\}$, a lower bound for the simultaneous selection probability is given by $\widehat{\Pi}_K^{sim,\rho} \geq 2\widehat{\Pi}_K^{\rho} - 1$.

Lemma 2: For a random subsamples, let $K \subset \{1, 2, ..., M\}$ and $\widehat{S}^{\rho}$ the set of selected terms. If $P(K \subseteq \widehat{S}^{\rho}) \leq \varepsilon$, then $P(\widehat{\Pi}_K^{sim,\rho} \geq \xi) \leq \varepsilon^2/\xi$. If $P(K \subseteq \cup_{\rho\in\Lambda}\widehat{S}^{\rho}) \leq \varepsilon$ for some $\Lambda \subseteq \mathbb{R}^+$, then $P(\max_{\rho\in\Lambda} \widehat{\Pi}_K^{sim,\rho} \geq \xi) \leq \varepsilon^2/\xi$.

Proof of the bounded theory (16) Define $Z_\Lambda = Z \cap \widehat{S}^{\Lambda}$ to be the set of redundant terms which are falsely selected into $\widehat{S}^{\Lambda}$ and analogously $U_\Lambda = S \cap \widehat{S}^{\Lambda}$ which represents the true model terms being selected into $\widehat{S}^{\Lambda}$. Then the expected number of redundant terms selected can be calculated with $E(|Z_\Lambda|) = E(|\widehat{S}^{\Lambda}|) - E(|U_\Lambda|) = q - E(|U_\Lambda|)$. In addition, assume that the original result is not worse than random guessing for any $\rho \in \Lambda$, namely $\frac{E(|S\cap\widehat{S}^{\rho}|)}{E(|Z\cap\widehat{S}^{\rho}|)} \geq \frac{|S|}{|Z|}$. With this assumption, we can obtain $E(|U_\Lambda|) \geq E(|Z_\Lambda|)|S|/|Z|$. According to the expression of $E(|U_\Lambda|)$ and $E(|Z_\Lambda|)$, we can get $(1 + |S|/|Z|)E(|Z_\Lambda|) \leq q$ and $|Z|^{-1}E(|Z_\Lambda|) \leq q/M$. Here, with the exchangeability assumption, $P(k \in \widehat{S}^{\Lambda}) = E(|Z_\Lambda|)/|Z|$ for all $k \in Z$. Therefore, $P(k \in \widehat{S}^{\Lambda}) \leq q/M$ holds as desired. Now, using **Lemma 2** above, $P(\max_{\rho\in\Lambda} \widehat{\Pi}_k^{sim,\rho} \geq \xi) \leq (q/M)^2/\xi$ for $k \in Z$ and $0 < \xi < 1$. Then $P(\max_{\rho\in\Lambda} \widehat{\Pi}_k^{\rho} \geq \lambda) \leq P((\max_{\rho\in\Lambda} \widehat{\Pi}_k^{sim,\rho} + 1)/2 \geq \lambda) \leq (q/M)^2/(2\lambda - 1)$ can be obtained with **Lemma 1**. Therefore, $E(|V|) = \sum_{k\in Z} P(\max_{\rho\in\Lambda} \widehat{\Pi}_k^{\rho} \geq \lambda) \leq \frac{1}{2\lambda-1}\frac{q^2}{M}$ is approved.

In the following part, we mainly focus on analyzing the impact of λ and q on the average number of falsely selected terms. If the final model has no redundant term, it means $E(|V|) < 1$. Here, considering the bounded theory shown in (16) and we could have two cases:

- $E(|\widehat{S}^{SS} \cap Z|) \leq \frac{q^2}{(2\lambda-1)M} < 1$**:** When $\frac{q^2}{(2\lambda-1)M} < 1$, then $\frac{1+\frac{q^2}{M}}{2} < \lambda$. Further, as $\lambda \in$(0.5,1], we have $\lambda \in (\frac{1+\frac{q^2}{M}}{2}, 1]$.
- $E(|\widehat{S}^{SS} \cap Z|) < 1 \leq \frac{q^2}{(2\lambda-1)M} < \infty$**:** When $1 \leq \frac{q^2}{(2\lambda-1)M}$, then $\lambda \leq \frac{1+\frac{q^2}{M}}{2}$. Further, as $\lambda \in$(0.5,1], we have $\lambda \in (0.5, \frac{1+\frac{q^2}{M}}{2}]$.

q can be used for determining the range for λ in theory. However, q is dependent on specific application and can be not easily given and therefore the above mentioned optimal range for λ is often impossible to obtain in practice. If the final model has some redundant terms, we have $b \leq E(|V|) \leq \frac{q^2}{(2\lambda-1)M}$ $(b \geq 1)$. As $b \geq 1$, $\lambda \leq \frac{1+\frac{q^2}{bM}}{2} \leq \frac{1+\frac{q^2}{M}}{2}$. In this case, increasing the threshold can reduce value $\frac{1+\frac{q^2}{M}}{2}$ and therefore limiting the amount of redundant terms. However, if λ is chosen as too big, say its upper limit 1, useful terms may be excluded from the model. In practice, in order to limit the number of falsely selected terms and avoid missing useful terms, $\lambda \in (0.6, 0.9)$ is recommended.

It is worth pointing out that the new algorithm tends to producing a more parsimonious model. But the resultant model may not be an optimal model. For example, a highly correlate term with high cross correlation coefficient may enter the model with a probability and may be included into the final model. In some cases, to produce an optimal model, the global search or Monte-carlo based methods have to be employed but they may require large computations. The new algorithm is capable of reducing the redundant terms by scarifying limited computations, which provides a good trade-off between model performance and computational requirements.

The differences with stability selection has been discussed as follows. First, stability selection is originally provided in statistics community and its effectiveness is demonstrated using static systems [8]. In this paper, we extended the idea to nonlinear dynamic systems. Dynamic system modelling is referenced to as system identification in control community. Although the proposed method shares similar conclusion with the stability selection, it deals with a different model construction problem. One of the main contributions of this paper in theory is that we proved that the average number of falsely selection terms for nonlinear dynamic systems can be bounded. Second, for the parameter setting, the threshold λ is given empirically with $\lambda \in (0.6, 0.9)$ in [8] while we give a reference value $\lambda = \frac{1+\frac{q^2}{M}}{2}$ via rigorous theoretical analysis, and we analyse its impact on final model performance with three scenarios in terms of models

with no, few and significant redundant terms. Finally, stability selection only focuses on variable selection but SOR carries out both variable selection and parameter estimation such that the produced model could have a satisfactory generalization performance for system identification.

To help understand SOR well, its advantages, disadvantages and when to use SOR are discussed as follows:

- **Advantages**: Compared to the OFR, the main advantage of SOR is that it could produce a more sparse model and it is less dependent on the stopping criterion as the model selection is based on term selecting probability.

- **Disadvantages**: The main disadvantage of SOR is that it requires more computations than OFR as it used sub-sampling techniques. If the sub-sampling repetition number r is chosen as 100, the computation cost of SOR is roughly 25 times of OFR. Further, SOR has one selection probability threshold λ to be chosen. In general, we can follow the theory analysis and choose λ within [0.6, 0.9], which can reduce the efforts and pains on determining the threshold.

- **When to use**: When dealing with a system identification problem, if we find that OFR or other selection methods are very sensitive to the model stopping criterion or different input-output data, SOR could be a better choice to build a consistent model with fewer or no redundant model terms.

5. Numerical examples

Consider the sparse nonlinear system [10] :

$$\begin{aligned} z(k) =& 0.2z^3(k-1) + 0.7z(k-1)u(k-1) - 0.7z(k-2)u^2(k-2) \\ & + 0.6u^2(k-2) - 0.5z(k-2) \\ y(k) =& z(k) + e(k). \end{aligned} \tag{17}$$

where $u(k)$ and $y(k)$ are the system input and output at interval k, respectively. The system is excited with a uniformly distributed white noise u(k) $\in$ [-1,1]. The system output $z(k)$ is disturbed by a Gaussian noise sequence $e(k)$ with the signal-to-noise rate (SNR) $15dB$. Here the SNR is defined as $SNR = 10log(\frac{\mathbf{Z}^T\mathbf{Z}}{\mathbf{\Xi}^T\mathbf{\Xi}})$, where $\mathbf{\Xi} = [e(1), ..., e(N)]$ and $\mathbf{Z} = [z(1), ..., z(N)]$ with N being the number of input-output data points. A data sequence of 800 samples were generated for system identification and therefore $N = 800$. The delayed input and output $\{y(t-1), y(t-2), y(t-3), y(t-4), u(t-1), u(t-2), u(t-3)\}$ from the nonlinear system are chosen as model input [11]. The polynomial NARX model with order up to 3 is used and it has 120 terms in total.

First of all, the conventional OFR is used to build a NARX model. It has to be mentioning that the following two factors can have an impact on the resultant model performance.

- **Stopping criterion**: The model selection criterion (9) is used and the threshold value ρ has to be chosen first. The threshold value ρ is related to the system noise $\mathbf{\Xi}$. If we use equations (4), (5), (9), the ideal threshold $\rho = 0.0307$ in the case of noise with 15dB SNR. However, noise is often unknown in practice and therefore the optimal value may not be given. Here, we suppose we can use near-optimal value, $\rho = 0.03$ as the threshold, which represents SNR=15.2dB. According to the definition of SNR, the larger SNR value means smaller noise. Using $\rho = 0.03$, the selection process only can stop when the model produces a equivalent SNR 15.2dB noise. The resultant model may have to include some redundant terms, in where we can test how many redundant terms will be produced under slightly over-fitted setting.

- **Random input data**: When repeating the same modelling process, different input-output data due to the randomly generated input data $u(k)$ can result in different models even if the settings, such as stopping threshold, are fixed. To study the impact of random input data, a Monte Carlo simulation with 100 repetitions is carried out.

To make a fair comparison, SOR uses the same setting with OFR. However, SOR has two additional parameters, the sub-sampling repetition number r and the selecting probability threshold λ. Following the empirical recommendation, $r = 100$ is used. Further, as for λ, instead of only choosing one value from the the recommended range of $\lambda \in [0.6,0.9]$, here we choose $\lambda = 0.6$. We

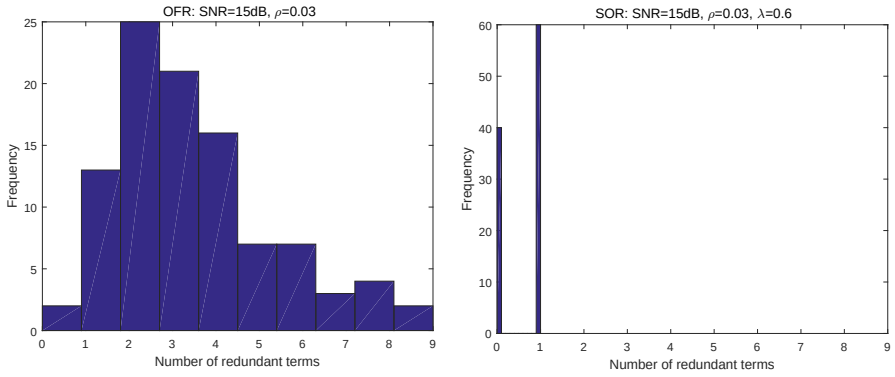

Figure 2: The redundant term distribution produced by OFR (left) and SOR (right) with $\lambda = 0.6$ for example 1 using Monte-Carlo simulation

use number of redundant terms as the performance evaluations. Under the above mentioned settings, the redundant terms distribution produced by OFR is shown in the left of Figure 2 with Monte Carlo simulation. It can seen that the number of redundant terms varies from 0 to 9 where 0 means no redundant terms in the model. More than 90% resultant models have redundant terms and majority have two or three redundant terms and a small proportion has over five redundant terms. Then SOR is used to build the NARX model. Before representing all the Monte Carlo results, one model is chosen from 100 final models is given in table 3, which is used to help understand the new method. It can be seen that all the five true terms have selecting probability above 0.9. One redundant term $u^2(k-2)y(k-4)$ could enter the model due to its high correlation with $u^2(k-2)y(k-2)$, which was also reported in [11, 12]. All other terms have very low selecting probabilities. The reason for this is that these terms enter the model by approximating the noise. When using sub-sampling

Table 3: Term selecting probability using SOR

Term	Probability (SOR)
$\mathbf{y(k-2)}$	1
$\mathbf{y(k-1)u(k-1)}$	1
$\mathbf{u^2(k-2)}$	1
$\mathbf{y^3(k-1)}$	1
$\mathbf{u^2(k-2)y(k-2)}$	0.92
$u^2(k-2)y(k-4)$	0.43
others	≤ 0.1

techniques, the sampling noise varies and their approximating terms also vary accordingly. Therefore, their selecting probabilities are low and these redundant terms can be referenced to as noisy terms. It is worth mentioning that other 99 models may have not the same term selecting probabilities as they use different input-output data. Now the whole results using SOR are shown in the right of Figure 2. It can be seen that majority has no redundant terms and only a small proportion has one redundant term. The significance of the new method is that it can built more spare and robust models than OFR method. Further, SOR produce more consistent results than OFR under different input-output data.

Moreover, least absolute shrinkage and selection operator (LASSO), as a widely used method, is employed here to make further comparison with the new method. There are many LASSO solvers that were published in the past decade. In this paper, the matlab toolbox function $'lasso'$ is used to build the NARX model. LASSO has one regularized parameter to be determined. To avoid unfair comparison, 5000 regularized parameters that are chosen by the matlab function were used to build 5000 models and the best one is picked up for comparison in terms of the least sum squared error with 2-folder and 5-folder cross validation (CV) scheme. Again, a Monte Carlo simulation with 100 realization is performed. The redundant term distribution is given in Figure 3. It can be seen that LASSO produces more redundant terms than SOR.

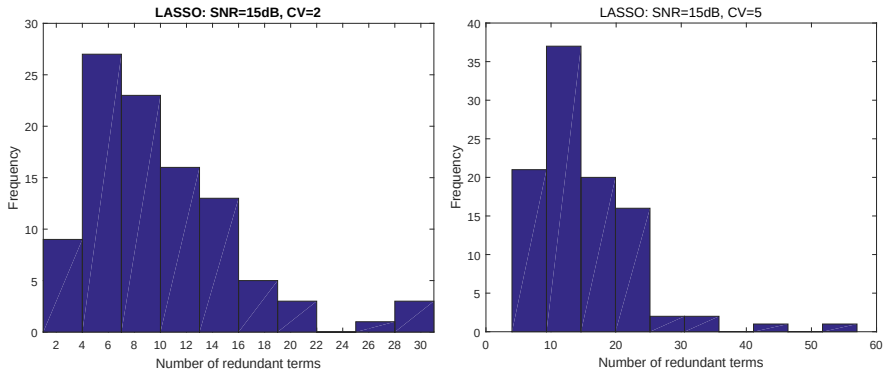

Figure 3: The redundant term distribution produced by lasso method with 2 (left) and 5 (right) folder cross validation for example 1 using Monte Carlo simulation

A special case of NARX model which identifies nonlinear system with polynomial NAR model. It is taken from [9], the specific format as follows

$$\begin{aligned} z(t) =&(0.8 - 0.5e^{-y^2(t-1)})y(t-1) + 0.1\sin(\pi y(t-1)) \\ &- (0.3 + 0.9e^{-y^2(t-1)})y(t-2) \\ y(t) =&z(t) + e(t). \end{aligned} \tag{18}$$

where y(t) is the system output at interval t and y(t) is disturbed by a Gaussian noise sequence e(t) with the signal-to-noise rate (SNR) 15dB. 800 samples are generated for identification. The delayed output $\{y(t-1), y(t-2), y(t-3), y(t-4)\}$ are used for model input. When the orders of polynomial are up to 3, then there are total 34 polynomial terms. All the experimental conditions are the same with previous examples. At the same time, Monte Carlo simulation is repeated 100 times. The average number of selected terms and average training error of different methods are listed in table 4. It can be found that LASSO has the largest average training error. Although OFR has a smaller error than that of LASSO, which is larger than that of SOR. In addition, SOR builds a parsimonious model with fewer terms compared to LASSO and OFR. The listing results have shown the effectiveness of SOR to build a parsimonious model with a satisfied model performance.

Table 4: The simulation results

Algorithm	No. of Terms	Error
OFR ($\rho = 0.03$)	9.32	0.0064
LASSO ($CV = 5$)	16.80	0.0168
SOR ($\lambda = 0.6$)	3.55	0.0015

6. Conclusion

In the present work, stability orthogonal regression (SOR) has been proposed for system identification. First, multiple intermediate models have been produced using sub-sampling technique and then the final model is a combination of intermediate model terms but does not include low frequently selected terms. SOR has employed the well known orthogonal forward regression (OFR) to build the intermediate models. The main advantage of the new method is that it is capable of building a more parsimonious model with fewer redundant model terms and further it has the potential to produce an optimal model. Theoretical analysis has analyzed the impact of choice of term selecting threshold. Results from numerical examples have confirmed the effectiveness of the proposed method with comparison to two popular methods.

References

[1] S. A. Billings, Nonlinear System Identification: NARMAX Methods in the Time, Frequency, and Spatio-Temporal Domains, John Wiley and Sons Ltd, 2013.

[2] S. Chen, S. A. Billings, W. Luo, Orthogonal least squares methods and their application to non-linear system identification, International Journal of Control 50 (1989) 1873–1896.

[3] L. Zhang, K. Li, Forward and backward least angle regression for nonlinear system identification, Automatica 53 (2015) 94–102.

[4] A. Falsone, L. Piroddi, M. Prandini, A randomized algorithm for nonlinear model structure selection, Automatica 60 (2015) 227–238.

[5] T. Baldacchino, S. R. Anderson, V. Kadirkamanathan, Computational system identification for Bayesian NARMAX modelling, Automatica 49 (9) (2013) 2641–2651.

[6] F. Bianchi, A. Falsone, M. Prandini, L. Piroddi, A randomised approach for NARX model identification based on a multivariate Bernoulli distribution, International Journal of Systems Science 48(6) (2017) 1203–1216.

[7] Y. Z. Guo, L. Z. Guo, S. A. Billings, H. L. Wei, Identification of Nonlinear Systems with Non-Persistent Excitation using an Iterative Forward Orthogonal Least Squares Regression Algorithm, International Journal of Modelling, Identification and Control 23(1) (2015) 1–7.

[8] N. Meinshausen, P. Bühlmann, Stability selection, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72 (4) (2010) 417–473.

[9] L. Zhang, K. Li, E. W. Bai, G. W. Irwin, Two-stage orthogonal least squares methods for neural network construction, Neural Networks and Learning Systems, IEEE Transactions on 26 (8) (2014) 1608 – 1621.

[10] L. Piroddi, W. Spinelli, An identification algorithm for polynomial narx models based on simulation error minimization, International Journal of Control 76 (17) (2003) 1767–1781.

[11] Y. Z. Guo, L. Z. Guo, S. A. Billings, H. L. Wei, An iterative orthogonal forward regression algorithm, International Journal of Systems Science 46 (5) (2015) 776–789.

[12] K. Z. Mao, S. A. Billings, Algorithms for minimal model structure detection in nonlinear dynamic system identification, International Journal of Control 48 (2013) 1553–1565.