Ultra Orthogonal Forward Regression Algorithms for the Identification of Non-Linear Dynamic Systems

Revised 20.07.2015

Yuzhu Guo, L.Z. Guo, S. A. Billings, and Hua-Liang Wei

Department of Automatic Control and Systems Engineering
The University of Sheffield, Mappin Street, Sheffield, S1 3JD, UK.
INSIGNEO institute for in silico Medicine
The University of Sheffield, Mappin Street, Sheffield, S1 3JD, UK.

Abstract

A new Ultra Least Squares (ULS) criterion is introduced for system identification. Unlike the standard least squares criterion which is based on the Euclidean norm of the residuals, the new ULS criterion is derived from the Sobolev space norm. The new criterion measures not only the discrepancy between the observed signals and the model prediction but also the discrepancy between the associated weak derivatives of the observed and the model signals. The new ULS criterion possesses a clear physical interpretation and is easy to implement. Based on this, a new Ultra Orthogonal Forward Regression (UOFR) algorithm is introduced for nonlinear system identification, which includes converting a least squares regression problem into the associated ultra least squares problem and solving the ultra least squares problem using the orthogonal forward regression method. Numerical simulations show that the new UOFR algorithm can significantly improve the performance of the classic OFR algorithm.

Key words: orthogonal forward regression, system identification, ultra least squares, ultra orthogonal forward regression, ultra orthogonal least squares.

1. Introduction

System identification plays a more and more important role in revealing the unknown mechanisms and rules underlying complex phenomena (Schmidt & Lipson, 2009). System identification includes the detection of the model structure and estimation of the associated parameters. A system identification problem can often be thought of as an optimization problem where the optimal model is searched from a large predefined candidate model dictionary given a criterion. The criterion is used to evaluate the performance of each model by measuring the discrepancy between the observed data and the model predictions. The candidate model dictionary is often chosen to be large enough to include the unknown correct model. Hence an exhaustive search algorithm is often infeasible in these kinds of applications because of the large solution space. Even an evolutionary algorithm which can greatly reduce the search process can still be very computationally intensive. Hence an algorithm which can efficiently find the optimal solution is desired. However, a fast algorithm often dictates an optimal substructure; otherwise the search may converge to a suboptimal solution. Many efforts have been made to improve the search process under a certain specific loss function or performance index, for example, the simulated annealing algorithm, particle swarm optimisation, and so on. In this paper, a different and new methodology will be introduced. Instead of improving the search method, a new and effective criterion will be introduced to describe the objective of the regression more accurately. Under the new criterion, the solution space has a better structure and a fast algorithm is more likely to find the optimal solution.

System identification aims to identify a model from observed data based on a criterion. A good criterion results in not only better parameter estimation but also a good search path along which the search process converges quickly to the optimal solution. Over the years, different criteria have been used in system identification such as the L^2 norm in least squares regression, the L^1 norm in least absolute value regression (Bloomfield & Steiger, 1980; Narula & Wellington, 1982), and zero-norm minimisation (Kaizhu, King, & Lyu, 2008), etc. Among these criteria, the least squares criterion is the most used because of its excellent properties, for example, least squares estimation can be configured to give estimates which are unbiased and efficient when the noise satisfies some basic assumptions. Least squares problems have analytic solutions and can easily be solved using the QR decomposition technique, and least squares regression produces unique and numerically robust solutions. Consequently a large number of system identification algorithms based on the least squares criterion have been developed (Billings, 2013; Li, Peng, & Bai, 2006; Ljung, 1987; Söderström, 1989).

However, the standard least squares method only reveals part of the information in the observed data. The least squares criterion, which considers the datum points individually, discards the connections among the datum points, especially for the identification of dynamic systems where the data set are time series which are samples of continuous functions of time. These individual datum points are time dependent and connected with each other through the derivatives of the time continuous functions, for example, an ordinary differential equation. Many important characteristics of a system can be determined by these interconnections. An absence of this information may lead to over-fitted models in least squares regression, which can be seen in the motivational example described in Figure 1 and discussed in the next section.

The standard least squares regression investigates the problem of model fitting on the space $L^2([0,T])$, where $[0,T]$ represents the time span of a signal. The associated Residual Sum of Squares (RSS), which is the square of the L^2 norm of the residual, is used to measure the fitness of the model. When the model structure is known, the standard least squares algorithm produces the best parameters with which the model will be optimal in the sense of RRS. Considering different model structures, there are plenty of very different models which give the same fitness for a set of observed data in the sense of the RSS criterion. In this paper, an alternative criterion, called ultra least squares (ULS) criterion will be introduced to characterise the model fitness more accurately. Unlike the least squares criterion consider the model fitting on the space L^2, the ULS criterion considers the model fitting in a smaller space, more specifically, the Sobolev space $H^m([0,T])$ (Maz'ia, 1985). The norm defined on this space will be modified and used as the ULS criterion for system identification, where not only residuals but also the associated weak derivatives will be used to measure the model fitness.

Using the derivatives of the data in system identification has been studied, especially in the identification of continuous time models (Brewer, Barenco, Callard, Hubank, & Stark, 2008; Preisig & Rippin, 1993; Schmidt & Lipson, 2009). However, as far as the authors are aware this study is the first time the weak derivatives have been combined with the least squares criterion to build a completely new metric for the prediction errors and which uses the new metric to improve the model structure detection in non-linear system identification.

In this paper, the ULS criterion will be combined with the well known Orthogonal Forward Regression (OFR) algorithm (Billings, 2013) to construct a new Ultra Orthogonal Forward Regression (UOFR) algorithm for nonlinear system identification. The proposed UOFR algorithm is shown to be very powerful for model structure detection in many modelling tasks and is more likely to produce an optimal model.

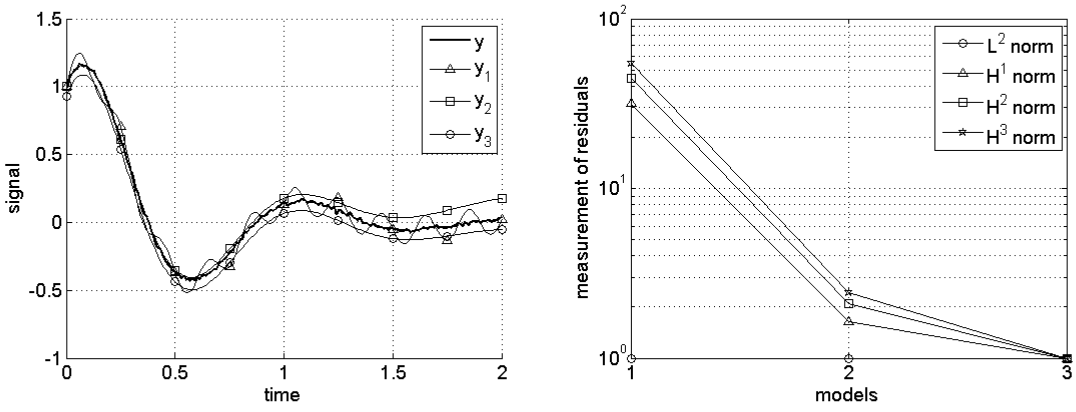
The remainder of the paper is organised as follows: Section 2 briefly reviews some main results on the Lebesgue space L^2 and the Sobolev space H^m. The ULS criterion will be presented by modifying the H^m norm in Section 3. The associated solution to the ultra least squares problem is then defined, and the new UOFR algorithm is described in Section 4. Three benchmark examples are discussed in section 5 to illustrate the efficiency of the new UOFR algorithm. Conclusions are finally drawn in Section 6.

2. Problems of least squares regression and model fitting in Sobolev space

In this section, a motivational example is first given to show the problems that can arise when using a standard least square criterion. The reasons which cause these problems will then be discussed in detail and an alternative criterion will be proposed.

Consider the time series fitting problem shown in Figure 1. In this example, three models were identified from an observed signal y which is represented by a thick solid line in Figure 1(a). The reproduced signals by the three models are represented by the curves y_1, y_2, and y_3 in the Figure 1(a) respectively. Figure 1(b) shows the different measurements of the model fitness of the three models: the L^2 norm and the H^m $(m = 1,2,3)$ norms of the residuals.

From Figure 1(b), it can be observed that the three models give the same fitness in the sense of the least squares criterion, which is presented by the line with the circle marks along the abscissa in 1(b), although the reproduced signal y_1 looks significantly different from y_2 and y_3 in 1(a).



(a) (b)

Figure 1 A motivational example for model fitting of a noisy signal

(a) Observed data and reproduced signals for three different models (b) measurement of the fitness of the models using different criteria

Figure 1 (b) shows also the measurements of the errors in the sense of H^m norms when $m = 1,2,3$. It can be observed that the performances of the three models under the H^m norms are significantly different. Model 3 fitted the signal y better than models 1 and 2 did. The system identification problems consists of finding the function on $H^m([0,T])$ which best fits the observed data $\{y_n\}$, *n=1,* *2, …, N*, where both the data points and the interconnections among the datum points (described by the weak derivatives) are considered.

This example shows that the least squares criterion which defined on the L^2 space neglects some very important information in the observations. This information is crucial for identifying a correct model. Alternatively, the model fitness can more accurately be characterised on a smaller space, the Sobolev space H^m , which consists of all the functions which are L^2 integrable and the where up to mth weak derivatives exist and are also L^2 integrable. The new introduced ULS criterion is a realisation of the H^m norm based on the observations.

The generic least squares regression problem includes determining the structure of a linear-in-the-parameters model and estimating the associated coefficients

$$y = \sum_{i=1}^{K} \theta_i x_i + e \tag{1}$$

via minimising the loss function

$$J_{LS} = \left\| y - \sum_{i=1}^{K} \theta_i x_i \right\|_2^2 \ . \tag{2}$$

The aim is to produce a parsimonious model, where y represents the dependant variable and the x_i's are the explanatory variables. In the system identification of dynamic systems, the y and x_i's are time dependent signals with finite energy, that is, these signals are L^2 integrable functions in the Lebesgue space $L^2([0,T])$, where $[0,T]$ is the time span of the signals. Least squares regression involves finding a model to minimise the square of the L^2 norm (2).

Although the least squares loss function is defined based on the L^2 norm, the signals in a dynamic system are more regular than a general L^2 function because of the fact that most physical systems behave essentially as a low-pass filter. These signals are actually functions defined on the subspace of $L^2([0,T])$, specifically, the Sobolev space $H^m([0,T]) = W^{m,2}([0,T])$ (Maz'ia, 1985)

$$H^m([0,T]) = \left\{ x(t) \in L^2([0,T]) \middle| D^l x \in L^2([0,T]), l = 1,2,\cdots,m \right\}, \tag{3}$$

which is the space of functions defined on $[0,T]$, the weak derivatives up to mth order are also L^2 integrable and t represents continuous time. The weak derivatives $D^l x(t)$ satisfy

$$\int_{[0,T]} x(t) D^l \varphi(t) dt = (-1)^l \int_{[0,T]} \varphi(t) D^l x(t) dt \tag{4}$$

for any test function $\varphi(t) \in C_0^\infty([0,T])$, which is smooth and possesses compact support on $[0,T]$.

The metric in $L^2([0,T])$ space is defined by the Lebesgure integral. Hence the distance $\|x-\hat{x}\|_2$, which measures the differences on the interval $[0,T]$ between functions $x(t)$ and $\hat{x}(t)$ as a whole, cannot characterise how the differences are distributed at each time instance. As a result, the L^2 norm only emphasises the similarity of two functions as a whole but disregards the closeness or detail in shape. System identification can be interpreted as discovering unknown rules from a set of observations. Every piece of information can be crucial for a method to discover the correct rules, especially when the system is not persistently excited, where many important system characteristics are not fully excited and are inconspicuously contained in a small number of data. The absence of this information can lead to a wrong model structure. However, this unapparent information can easily be overshadowed by a large amount of trivial data in a global criterion such as the L^2 norm. Therefore, an unclear objective function may confuse system identification algorithms and increase the algorithms sensitivity to noise. Nuances in the data may therefore cause the algorithms to produce incorrect models. Hence a stricter criterion which can accurately characterise the objectives of system identification and reveal all the useful information in data should be investigated.

A stricter metric for the Sobolev space $H^m([0,T])$ is the norm defined as

$$\|x\|_{H^m} = \sqrt{\sum_{l=0}^{m} \|D^l x\|_2^2}\,, \tag{5}$$

where D^l represents the l th differentiation operator.

Based on the above norm, a new criterion can then be defined as

$$J_H = \left\| y - \sum_{i=1}^{\kappa} \theta_i x_i \right\|_2^2 + \sum_{l=1}^{m} \left\| D^l \left(y - \sum_{i=1}^{\kappa} \theta_i x_i \right) \right\|_2^2 . \tag{6}$$

Due to the fact that the differentiations are linear operators, the above criterion can be written as

$$J_H = \left\| y - \sum_{i=1}^{\kappa} \theta_i x_i \right\|_2^2 + \sum_{l=1}^{m} \left\| D^l y - \sum_{i=1}^{\kappa} \theta_i D^l x_i \right\|_2^2 . \tag{7}$$

The J_H criterion consists of two parts: the first part is the standard least squares criterion which emphases the agreements over the data set; while the second part represents the agreement of the weak derivatives which essentially emphases the agreement in shape. Any change in the distribution of the differences will be reflected in the second part of the criterion. Hence, the new J_H criterion, which can reveal more information by introducing the second term, is an alternative criterion to the pure least squares criterion. In the next section, an ULS criterion will be derived by adapting the J_H criterion to the nonlinear system identification problem.

Next, two theorems about the relationships between the L^2 and H^m spaces and the associated norms will be given to show that the Sobolev space H^m is an appropriate space, and the associated norm is an appropriate criterion for a least squares problem.

Theorem 1. For any approximation $\hat{y} \in L^2([0,T])$ to a function y in $L^2([0,T])$ satisfying $\|y-\hat{y}\|_2 \le \varepsilon$, there exists an $\tilde{y} \in H^m([0,T])$ satisfying $\|y-\tilde{y}\|_2 \le \gamma\varepsilon$ for any $\gamma > 1$.

This theorem can easily be proved from the result that H^m is dense in L^2 so that there exists a function $\tilde{y}$ in H^m subject to $\|\tilde{y}-\hat{y}\|_2 < (\gamma - 1)\varepsilon$. Based on Minkowski's inequality $\|y-\tilde{y}\|_2$ = $\|y-\hat{y}+\hat{y}-\tilde{y}\|_2$ $\le \|y-\hat{y}\|_2 + \|\hat{y}-\tilde{y}\|_2$.

Theorem 1 shows that there exists an approximation in H^m which is not significantly different from an approximation $\hat{y}$ in L^2 but is more regular. Therefore, the Sobolev space $H^m([0,T])$ is smaller than the $L^2([0,T])$ space but large enough for a least squares approximation.

Theorem 2. For any small positive real number ε, $\|y-\tilde{y}\|_{H^m} < \varepsilon$ means $\|y-\tilde{y}\|_2 < \varepsilon$.

Proof:

The result is straightforward because

$$\|y-\tilde{y}\|_{H^m} = \sqrt{\|y-\tilde{y}\|_2^2 + \sum_{l=1}^{m} \|D^l y - D^l \tilde{y}\|_2^2} \ge \sqrt{\|y-\tilde{y}\|_2^2} = \|x\|_2 . \qquad (8)$$

Hence, for any small positive number $\varepsilon > 0$, $\|y-\tilde{y}\|_2 \le \|y-\tilde{y}\|_{H^m} < \varepsilon$. This proves the theorem. □

Theorem 2 indicates that the J_H criterion is a stricter than the least squares criterion. If a model fits the data well in the sense of J_H, the model is also good in the sense of least squares. The reverse is not true.

3. Ultra least squares problems and the ultra least squares criterion

Definition 1. Under the new H^m norm, the least squares problem (1) is equivalent to a new least squares problem

$$\begin{bmatrix} y \\ D^1 y \\ \vdots \\ D^m y \end{bmatrix} = \sum_{i=1}^{\kappa} \theta_i \begin{bmatrix} x_i \\ D^1 x_i \\ \vdots \\ D^m x_i \end{bmatrix} . \tag{9}$$

The new least squares problem (9) will be defined as the **ultra least squares problem** corresponding to the original least squares problem. The solution of the ultra least squares problem will be referred to as the **ultra least squares solution** of the original least squares problem.

The ultra least squares solution can be obtained by solving the ultra least squares problem. However, some more work is still needed before this can be used for data-driven system identification problems. Firstly, the weak derivatives are usually not known in many system identification problems. Secondly, the contribution of each component $\left\| D^l y - \sum_{i=1}^{\kappa} \theta_i D^l x_i \right\|_2^2$ to the J_H criterion may not be balanced. The differentiation terms $\left\| D^l y - \sum_{i=1}^{\kappa} \theta_i D^l x_i \right\|_2^2$ may magnify the effects of noise and incorrectly dominate the J_H criterion. Consequently, J_H criterion will not be robust to noise. Thirdly, the H^m norm looks quite mathematical when it would be preferable that the identification process is physically easy to understand and computationally cheap.

In order to evaluate the contribution of the unknown weak derivatives in the J_H criterion, the distributions corresponding to the signals y and x_i will be introduced. For the signal $y(t)$, the associated distribution T_y can be defined as a functional $T_y : C_0^\infty([0,T]) \to R$

$$\langle T_y, \varphi \rangle = \int_{[0,T]} y(t) \varphi(t) dt \tag{10}$$

for all $\varphi \in C_0^\infty([0,T])$. The distribution T_y now has weak derivatives which are defined as

$$\langle D^l T_y, \varphi \rangle = (-1)^l \int_{[0,T]} y(t) \varphi^{(l)}(t) dt . \tag{11}$$

Similarly, the distributions corresponding to x_i can be defined as

$$\langle T_{x_i}, \varphi \rangle = \int_{[0,T]} x_i(t) \varphi(t) dt , \ \ \varphi \in C_0^\infty([0,T]) \tag{12}$$

The regression is now solved in the sense of distribution. The system identification problem involves fitting the distribution T_y by the combination of a set of distributions T_{x_i}'s. The ultra least squares problem (9) then becomes

$$\begin{bmatrix} y \\ \left\langle D^1 T_y, \varphi \right\rangle \\ \vdots \\ \left\langle D^m T_y, \varphi \right\rangle \end{bmatrix} = \sum_{i=1}^{\kappa} \theta_i \begin{bmatrix} x_i \\ \left\langle D^l T_x, \varphi \right\rangle \\ \vdots \\ \left\langle D^l T_x, \varphi \right\rangle \end{bmatrix}, \ \varphi \in C_0^\infty\left([0,T]\right) \tag{13}$$

Data can be collected by evaluating the values of these distributions for different test functions $\varphi(t)$ in $C_0^\infty([0,T])$. The regression matrix can then be constructed and the parameters can be estimated based on the regression matrix.

However, there are not a finite number of functions which form a basis of $C_0^\infty([0,T])$. Hence, it is infeasible to evaluate the values of the distributions over the whole $C_0^\infty([0,T])$ space. A trade-off is needed between incorporating all the information of the distributions in the ULS problem and computational efficiency.

The weak derivatives of a function based on a locally defined test function $\varphi(t)$ reveal the local information of the function. Shifting the test function along the time axis yields different test functions and the associated weak derivatives contain local information of the signal at the new positions. Instead of all the test functions in $C_0^\infty([0,T])$, a locally defined test function $\varphi(t)$ and time shifts $\varphi(t-\tau)$ are used in the following discussion.

Given a test function $\varphi(t)$ with a finite support on $[0,T_0]$, $T_0 < T$, the ultra least squares problem can then be approximately described as

$$\begin{bmatrix} y \\ \left\langle D^1 T_y, \varphi(t-\tau) \right\rangle \\ \vdots \\ \left\langle D^m T_y, \varphi(t-\tau) \right\rangle \end{bmatrix} = \sum_{i=1}^{\kappa} \theta_i \begin{bmatrix} x_i \\ \left\langle D^1 T_{x_i}, \varphi(t-\tau) \right\rangle \\ \vdots \\ \left\langle D^m T_{x_i}, \varphi(t-\tau) \right\rangle \end{bmatrix}, \ \tau \in [0, T-T_0] \tag{14}$$

$$\left\langle D^l T_y, \varphi(t-\tau) \right\rangle = (-1)^l \int_{[0,T]} y(t)\, \varphi^{(l)}(t-\tau)\, dt \tag{15}$$

The distribution $\left\langle D^m T_y, \varphi(t-\tau) \right\rangle$ can then be thought as a function of τ. Denote the function as

$$y^l(\tau) \triangleq \left\langle D^l T_y, \varphi(t-\tau) \right\rangle = (-1)^l \int_0^T y(t)\varphi^{(l)}(t-\tau)\,dt\,. \tag{16}$$

Here $y^l(\tau)$ is the convolution of the signal $y(t)$ with the lth derivative of a function which is defined as $g(t) \triangleq \varphi(-t)$. The weight function $g^{(l)}(t)$ can then be thought as the impulse response of a linear filter and $y^l(\tau)$ is the output of the filter to the input $y(t)$.

According to Leibniz integral rule, differentiation under the integral sign satisfies

$$\frac{d^l}{dt^l}\int_0^t y(\tau)g(t-\tau)\,d\tau = \int_0^t y(\tau)\frac{\partial^l}{\partial t^l}g(t-\tau)\,d\tau = \int_0^t y(\tau)\frac{d^l}{dt^l}g(t-\tau)d\tau \tag{17}$$

That is, the order of the differentiation and the integral can be interchanged.

Now the newly introduced functions $y^l(\tau)$ have a new physical interpretation. The function $y^l(\tau)$ represents a signal which is obtained from the original signal $y(t)$ through two steps: the signal $y(t)$ is initially smoothed by a filter with an impulse response $g(t)$, the smoothed signal then passes through an lth order pure differentiation system. This is equal to smoothing the signal first and then calculating the derivatives of the smoothed signals.

Similarly, the functions $x_i^l(\tau)$ can be defined as

$$x_i^l(\tau) \triangleq \left\langle D^l T_{x_i}, \varphi(t-\tau) \right\rangle = (-1)^l \int_0^T x_i(t)\varphi^{(l)}(t-\tau)\,dt \tag{18}$$

The ULS problem (14) then becomes

$$\begin{bmatrix} y \\ y^1 \\ \vdots \\ y^m \end{bmatrix} = \sum_{i=1}^{\kappa}\theta_i \begin{bmatrix} x_i \\ x_i^1 \\ \vdots \\ x_i^m \end{bmatrix} \tag{19}$$

where y^l and x_i^l are the signals defined by (16) and (18). The ULS problem involves detecting the model structure and estimating the associated parameters based on both the observed signals and the derivatives of the smoothed signals. The system identification problem is therefore a signal processing problem, which is physically easy to understand and computationally cheap.

Another problem which may be caused by the J_H criterion in the application of system identification is that the difference arising from the derivatives can be much larger than the errors arising from the data themselves, that is, $\sum_{l=1}^{m}\left\| D^l y-\sum_{i=1}^{\kappa}\theta_i D^l x_i\right\|_2^2 \gg \left\| y-\sum_{i=1}^{\kappa}\theta_i x_i\right\|_2^2$, especially, when the residuals change quickly. The errors arising from the derivatives can dominate the value of the J_H defined in (7). For example, in the curve fitting problem shown in Figure 1, the H^m (m=1, 2, 3) norms are much greater than the L^2 norm. The measurements of the derivatives have been introduced to help the new ULS criterion to be more sensitive to the differences in the shape of the residuals. However a good criterion should be robust and not sensitive to the noise. Therefore, some further modifications need to be made to the test function and its derivatives. The test function and the associated derivatives will therefore be normalised before they are applied to the signal to give

$$\bar{\varphi}^l = \varphi^{(l)} \Big/ \left\| \varphi^{(l)} \right\|_2 , \; l = 1, 2, \cdots, m. \tag{20}$$

which satisfies

$$\int_{[0,T]} \bar{\varphi}^l (t)\, dt = 1 . \tag{21}$$

These normalised test functions will be used to modulate the signals instead of $\varphi^{(l)}$ in equation (16) and (18). This ensures that each datum from the modulated function $y^l(\tau)$ have the same weight in the criterion as the data in the original signal $y(t)$.

Given a test function $\varphi(t)$, the ULS problem corresponding to the LS problem (1) can then be summarised as follows

$$\begin{bmatrix} y \\ \bar{y}^1 \\ \vdots \\ \bar{y}^m \end{bmatrix} = \sum_{i=1}^{\kappa}\theta_i \begin{bmatrix} x_i \\ \bar{x}_i^1 \\ \vdots \\ \bar{x}_i^m \end{bmatrix} , \tag{22}$$

where

$$\begin{aligned} \bar{y}^l(\tau) &= \int_0^t y(t)\bar{\varphi}^{(l)}(t-\tau)\,dt \\ \bar{x}_i^l(\tau) &= \int_0^t x_i(\tau)\bar{\varphi}^{(l)}(t-\tau)\,dt \end{aligned} . \tag{23}$$

The ULS criterion is then be given by

$$J_{ULS} = \left\| y-\sum_{i=1}^{\kappa}\theta_i x_i\right\|_2^2 + \sum_{l=1}^{m}\left\| \bar{y}^l-\sum_{i=1}^{\kappa}\theta_i \bar{x}_i^l\right\|_2^2 \tag{24}$$

Since the objective of the test function $\varphi(t)$ is to smooth the observed signals, the test function is chosen to have a bell shaped Gaussian like function shape. Actually, the test functions do not need to be infinitely differentiable. A test function that has up to mth order continuous derivatives is enough for the ULS criterion. In this paper, the ($m+1$)th order B-spline functions which have finite support and continuous mth order derivatives will be used as the modulating functions. The definitions of the B-spline basis function and the associate derivatives are given in the Appendix A.

Given discrete observations of the signals, $\{y(n)\}$, $\{x_i(n)\}$, $n=1,2,\cdots,N$, the discrete version of the modulating process (23) can be written as

$$\begin{aligned} \bar{y}^l(k) &= \sum_{n=k}^{k+n_0} y(n)\bar{\varphi}^l(n-k) \\ \bar{x}_i^l(k) &= \sum_{n=k}^{k+n_0} x_i(n)\bar{\varphi}^l(n-k) \end{aligned} \tag{25}$$

where n_0 is the support of the discrete test function and $k=1,2,\cdots,N-n_0$.

The matrix form of the ULS problem can then be written as

$$\mathbf{Y}_{ULS} = \mathbf{\Phi}_{ULS}\boldsymbol{\theta} \tag{26}$$

where

$$\mathbf{Y}_{ULS} = \left[y(1)\cdots y(N)\, \bar{y}^1(1)\cdots \bar{y}^1(N-n_0)\cdots \bar{y}^m(1)\cdots \bar{y}^m(N-n_0)\right]^T \tag{27}$$

$$\mathbf{\Phi}_{ULS} = \begin{bmatrix} x_1(1)\cdots x_1(N)\, \bar{x}_1^l(k)^1(1)\cdots \bar{x}_1^l(k)^1(N-n_0)\cdots \bar{x}_1^l(k)^m(1)\cdots \bar{x}_1^l(k)^m(N-n_0) \\ \ddots \\ x_\kappa(1)\cdots x_\kappa(N)\, \bar{x}_\kappa^l(k)^1(1)\cdots \bar{x}_\kappa^l(k)^1(N-n_0)\cdots \bar{x}_\kappa^l(k)^m(1)\cdots \bar{x}_\kappa^l(k)^m(N-n_0) \end{bmatrix}^T \tag{28}$$

$$\boldsymbol{\theta} = [\theta_1\ \theta_2\ \cdots\ \theta_\kappa]^T\ . \tag{29}$$

4. The ultra orthogonal forward regression algorithm

Nonlinear system identification involves both the estimation of the parameters and more importantly the problem of how to detect the structure of the unknown model. Model structure detection for linear systems is relatively easy and usually involves determining the order and time delay in a linear model. However, model detection can be very complicated when the system is nonlinear because of the many potential model terms and complex dynamics. Several model detection methods have been developed, for example, the MP (Matching Pursuit) algorithms (Mallat & Zhang, 1993), OFR (Orthogonal Forward Regression) algorithm(Billings, 2013), and SR (Symbolic Regression) algorithms (Koza, 1992). One of the most popular nonlinear system identification methods is based on the NARMAX (Nonlinear AutoRegressive Moving Average with eXogenous input) model and the associated Orthogonal Forward Regression (OFR) algorithm (Billings, 2013) (also referred to as the OLS (Orthogonal Least Squares) or the FOLSR (Forward Orthogonal Least Squares Regression)).

Consider a nonlinear dynamic system which is represented by an NARX (Nonlinear Auto-Regressive with eXogenous input) model as

$$y(k)=F\left(y(k-1),y(k-2),\cdots,y\left(k-n_y\right),u(k-1),u(k-2),\cdots,u(k-n_u)\right)+e(k) \tag{30}$$

where y, u, and e are the output, input, and the noise sequences, respectively.

Function $F(\cdot)$ is a nonlinear function of the system input and output, which is often approximated by the linear combination of a set of basis terms ϕ_i when the structure is unknown.

$$y(k)=\sum_{i=1}^{K}\theta_i\phi_i\left(y(k-1),\cdots,y\left(k-n_y\right),u(k-1),\cdots,u(k-n_u)\right)+e(k) \tag{31}$$

Where the ϕ_i's are some basis functions of the system input and output; θ_i are the associated parameters.

The system identification problem involves selecting the most significant terms from a pre-defined candidate dictionary to build a model which is sufficient to describe the observed system behaviours. System identification then involves model structure detection and parameter estimation. In a system identification problem, these two processes are closely connected with each other. The parameter estimation depends on a certain model structure. Conversely, when the performance of a model structure is assessed, the associated parameters need to be estimated before this can be achieved. Hence, system identification can involve tedious trial and error processes, where parameters are re-estimated for each assumed model structure, unless a more principled approach is employed to efficiently select model terms.

The orthogonal forward regression decouples the model structure detection and the parameter estimation by orthogonalising the model terms and selecting terms stepwise based on the ERR (Error Reduction Ratio) significance criterion to build a parsimonious model in an efficient model selection and estimation algorithm. The forward regression method also avoids the singularity of the regression matrix caused by redundant terms in a model.

The orthogonalisation can be done using a Gram-Schmidt algorithm as follows

$$\begin{aligned} w_1 &= \phi_1 \\ w_k &= \phi_k - \sum_{i=1}^{k-1} \frac{\langle w_i, \phi_k \rangle}{\langle w_i, w_i \rangle} \end{aligned} \tag{32}$$

The contribution of a term ϕ_k can then be assessed by evaluating the ERR significance criterion which is defined based on the associated orthogonalised term w_k

$$ERR(\phi_k) = \frac{\langle w_k, y \rangle^2}{\langle w_k, w_k \rangle \langle y, y \rangle} = \frac{g_k^2 \langle w_k, w_k \rangle}{\langle y, y \rangle}, \quad g_k = \frac{\langle w_k, y \rangle}{\langle w_k, w_k \rangle}. \tag{33}$$

The OFR algorithm selects terms in a forward manner to build a better model by adding an extra term into the model one at a time. At each step all the remaining candidate terms in the dictionary are orthogonalised with the terms which are already in the model and the term which gives the greatest ERR value is selected as the next term in the model.

Along the orthogonalisation path, the first k term model should be optimal in all the k term models. However, this condition can occasionally be broken, especially when a system is not persistently excited, as shown for example in the papers (Ayala Solares & Wei, 2015; Mao & Billings, 1997; Piroddi & Spinelli, 2003). While non-persistently exciting inputs should always be avoided as a matter of good scientific practice, there are occasions where this is not possible. An iOFR (iterative Orthogonal Forward Regression) algorithm has therefore been proposed to reduce these problems while maintaining the simplicity of the identification procedure. Since rearranging of the order of terms does not affect the sum of the ERR's, the pre-determination of correct terms with a relatively small ERR value can make the remaining terms more likely to win in the following term selections. In this paper a different philosophy is used, where the UOFR algorithm is employed to solve the original least squares regression problem by solving a corresponding new ULS problem. Using this approach, the ULS criterion provides a more accurate description of the optimal solution. The new ULS solutions will then have better properties than a LS solution. Some of the locally optimal solutions under the LS criterion will not be a suboptimal solution under the new criterion. Hence, the UOFR is more likely to find the global optimal solution without significantly increasing the computation.

The UOFR algorithm can now be summarised as follows:

1) Specify an initial full model dictionary of M candidate terms and a cut-off value ρ;

2) Specify a test function $\varphi(t)$ and calculate the associated derivatives $\varphi^{(1)}(t)$, $\varphi^{(2)}(t)$, … , $\varphi^{(m)}(t)$;

3) Normalise the derivatives $\varphi^{(l)}(t)$ according to equation (20) to produce the normalised modulating functions $\bar{\varphi}^{l}$'s;

4) Calculate $\bar{y}^{l}$'s and $\bar{\phi}^{l}$'s by modulating the dependent variables and the regressors using the normalised functions $\bar{\varphi}^{l}$ and construct the associated ULS problem (22);

5) Evaluate the values of the ERR for each of the terms in the dictionary and select the term which gives the largest value of ERR into the model as the first term and remove the term from the dictionary;

6) At the kth step, orthogonalise each of the remaining terms in the dictionary with the $k-1$ selected terms in the model and calculate the ERR of this term. Compare the ERR significance of all the remaining terms and select the term which gives the largest ERR of the remaining terms into the model as the kth term.

7) Evaluate the value of the sum of the ERR's. Terminate the forward regression if the termination criterion $1-\sum_{j=1}^{k} ERR(\phi_j) < \rho$ is satisfied. Otherwise set k=k+1 and repeat step (6) until the condition is satisfied.

8) Estimate the parameters of the model using a least squares method.

Remarks:

a) The new UOFR algorithm, which employs the classic OFR algorithm to solve the proposed ULS problem, inherits the computational efficiency of the OFR algorithm in term selection.

b) The purely forward selection process in the UOFR algorithm can be greedy and produce suboptimal solutions when the test functions are not appropriately selected and the ULS problem does not have an optimal substructure (Cormen, Leiserson, Rivest, & Stein, 2009).

c) Different termination criteria can be used in step 7) to stop the regression process, for example, the APRESS (adaptive prediction sum of squares) criterion (Billings & Wei, 2008).

5. Test examples

The classic OFR algorithm can occasionally converge to suboptimal solutions, especially when the system is not persistently excited or the signals are incorrectly (usually over) sampled. Some systems have been proposed as benchmark examples for the study of variations of OFR algorithms and for comparisons of OFR with other algorithms (Baldacchino, Anderson, & Kadirkamanathan; Guo, Guo, Billings, & Wei, 2015; Mao & Billings, 1997; Piroddi & Spinelli, 2003). In this section, these examples will be used to test the new UOFR algorithm. In all the benchmark examples, the UOFR algorithm successfully detects the correct model structure. Based on the new ULS loss function, the ERR significance criterion works better in the forward term selection. All the correct terms are stepwise selected. The redundant terms which confused the OFR algorithm are less significant under the new criterion and are excluded from the correct model.

While these examples have been selected to allow comparisons with often solutions it should be emphasised that these are worst case examples. Normally any data which is not persistently exciting should not be used irrespective of which identification procedure is to be employed. Ideally non-persistently exciting data should not be used rather new experiments should be conducted to obtain good quality data sets. All algorithms for linear and nonlinear system identification may not give correct results when using non-persistently exciting data.

5.1 Example 1

This example is taken from (Mao & Billings, 1997). It has been shown that the classic OFR algorithm can produce a suboptimal model containing redundant terms. Consider the nonlinear system

$$\begin{aligned} y(k) = 0.2y^3(k-1) + 0.7y(k-1)u(k-1) + 0.6u^2(k-2) - 0.5y(k-2) \\ -0.7y(k-2)u^2(k-2) + e(k) \end{aligned} \tag{34}$$

The system is excited with a uniformly distributed white noise $u(k) \sim U(-1,1)$ and the output $y(k)$ is disturbed by a normally distributed white noise $e(k) \sim N(0, 0.1^2)$. A total number of 1000 input and output datum points were used for the system identification.

Up to third order polynomials of the delayed inputs and outputs $\{y(k-1)$, $y(k-2)$, $y(k-3)$, $y(k-4)$, $u(k-1)$, $u(k-2)$, $u(k-3)\}$ were used as the initial potential model terms. A total number of 120 terms were therefore included in the initial term dictionary. Applying the OFR algorithm yields a six-term model which is shown in Table 1. The terms in bold font are the correct model terms. Notice that a redundant term $y(k-4)u^2(k-2)$ which is not in system (34) was selected by the classic OFR algorithm. Under the LS criterion, the correct terms are less significant than the redundant term $y\left(k-4\right)u^{2}\left(k-2\right)$, which is unreasonable. For the reason why the redundant term was selected at the first step, refer to the discussion in our earlier paper (Guo et al., 2015).

Table 1 Results produced by the classic OFR algorithm for example 1

No.	Terms	ERRs	Coefficients	Standard Deviation
1	y(k-4)u^2(k-2)	30.265029	-0.0582872	0.02912
2	**y(k-1)u(k-1)**	13.781921	0.692277	0.01575
3	**u^2(k-2)**	14.405237	0.613678	0.009717
4	**y(k-2)**	28.312587	-0.493238	0.01249
5	**y(k-2)u^2(k-2)**	3.354679	-0.755392	0.03692
6	**y^3(k-1)**	2.439386	0.205762	0.01139
SERR	--	92.559%	--	--

The UOFR was also used to identify the model from the same candidate term dictionary. In the UOFR algorithm, cubic B-spline basis function is used as the modulating function and the first and second order derivatives of the smoothed signals are considered in the ULS criterion. The associated ULS problem in (26) ~ (29) with $m=2$ is then identified. Table 2 shows the output of the UOFR algorithm. This time the correct term $y\left(k-2\right)$ was selected as the most significant term overwhelming the wrong term $y\left(k-4\right)u^{2}\left(k-2\right)$. Under the new ULS criterion, the significance of the redundant term is greatly reduced and does not appear in the model at the UOFR regression process. Figure 2 gives the comparison of the UOFR algorithm and the classic OFR algorithms. The UOFR converged faster than the OFR and obtained the optimal model at the fifth step.

Table 2 Results produced by the UOFR algorithm for example 1

No.	Terms	ERRs	Coefficients	Standard Deviation
1	**y(k-2)**	49.083789	-0.50221	0.007378
2	**u^2(k-2)**	26.235897	0.602759	0.005481
3	**y(k-1)u(k-1)**	12.23879	0.682345	0.009117
4	**y(k-4)u^2(k-2)**	3.083899	-0.709869	0.01838
5	**y^3(k-1)**	2.127315	0.194917	0.006586
SERR	--	92.770%	--	--

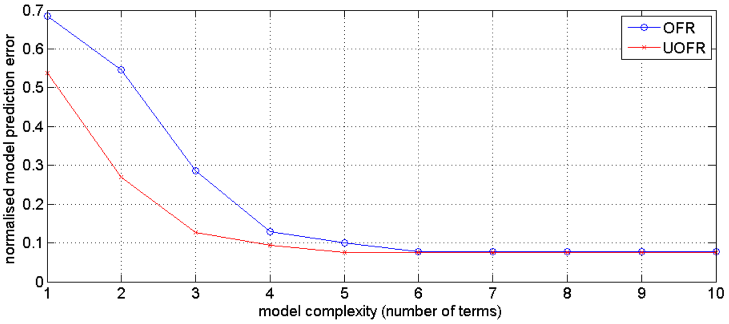


Figure 2 Comparison of the UOFR with the classic OFR for example 1

5.2 Example 2

Piroddi and Spinelli (2003) has shown that a classic OFR algorithm may include redundant autoregressive terms, even when the data set was produced from a purely moving average model. This is because the time series $\{y(k-1)\}$ is always "close" to $\{y(k)\}$ when the sample rate is high. The "closeness" is represented by the cross correlation coefficient, which depends on the inner product in L^2 space. Hence, when a LS criterion is used, the term $y(k-1)$ will always be highly significant. It will be shown that the significance of the tricky term will be greatly reduced when the ULS criterion is used if it is actually not a correct term.

The example is given as follows

$$\begin{cases} w(k) = u(k-1) + 0.5u(k-2) + 0.25u(k-1)u(k-2) - 0.3u^3(k-2) \\ y(k) = w(k) + \dfrac{1}{1-0.8z^{-1}} e(k) \end{cases} \tag{35}$$

where $u(k)$ represents the input signal and $y(k)$ represents the observation of the output $w(k)$ which is polluted by noise $e(k)$. The classic OFR algorithm can correctly detect the model structure when the system is persistently excited. Piroddi and Spinelli has shown that the classic OFR algorithm may incorrectly select autoregressive terms when the input signal is less rich in frequency components and they recommended an input which is generated by an AR process with two real poles between 0.75 and 0.9. Repeating Piroddi and Spinelli's simulation using an input signal which was generated by the following AR process.

$$u(k) = \frac{0.25}{1-1.6z^{-1}+0.64z^{-2}} v(k) \tag{36}$$

where $v(k)$ is Gaussian noise $v(k) \sim N(0,1)$. The AR process has a repeat pole at 0.8 and the coefficient 0.25 is chosen to guarantee the input signal has reasonable amplitude. Here the noise

$e(k)$ is a Gaussian distributed noise with a variance 0.02, that is, $e(k) \sim N(0, 0.02)$. The results produced by the standard OFR algorithm are given in Table 3. Observe that two incorrect autoregressive terms were selected overwhelming the correct terms. A correct term $u(k-1)u(k-2)$ was missed in the identification.

Table 3 Results produced by the standard OFR algorithm for example 2

No.	Terms	ERRs	Coefficients	Standard Deviation
1	y(k-1)	88.658562	0.411992	0.00554
2	y(k-2)	3.145405	0.00399801	0.002016
3	**u^3(k-1)**	1.999676	-0.300419	0.0006478
4	u^3(k-2)	4.872847	0.125053	0.001666
5	**u(k-1)**	0.179355	1.14722	0.008823
6	u^2(k-1)	0.96719	0.147869	0.0013
7	**u(k-2)**	0.013696	-0.259794	0.01263
SERR	--	99.837%	--	--

The output signal and all the candidate terms are modulated using the first and second order derivatives of a cubic B-spline basis function and the UOFR is applied. The identified model by the UOFR is given in Table 4. This time all the correct terms were successfully detected. The redundant terms are avoided using the UOFR algorithm.

Table 4 Model identified using the UOFR algorithm for example 2

No.	Terms	ERRs	Coefficients	Standard Deviation
1	**u^3(k-1)**	81.543654	-0.299632	0.0001863
2	**u(k-1)**	10.684663	1.01012	0.004504
3	**u (k-1)u(k-2)**	6.937331	0.250064	0.0005112
4	**u(k-2)**	0.401121	0.490076	0.004161
SERR	--	99.567%	--	--

A comparison of the UOFR and OFR algorithms is shown in Figure 3. The UOFR converges faster than the OFR and produces the optimal model at the fifth step. It can be observed that term $y(t-1)$ was selected by the OFR algorithm because it gives the greatest ERR value at the first step. However, in the UOFR algorithm, terms $y(t-1)$ and $y(t-2)$ which were dominant in the OFR are less significant than the correct terms under the new criterion even though they are close to the dependant variable $y(t)$ in the sense of a LS criterion. The new UOFR can successfully solved the problem caused by $y(t-1)$.

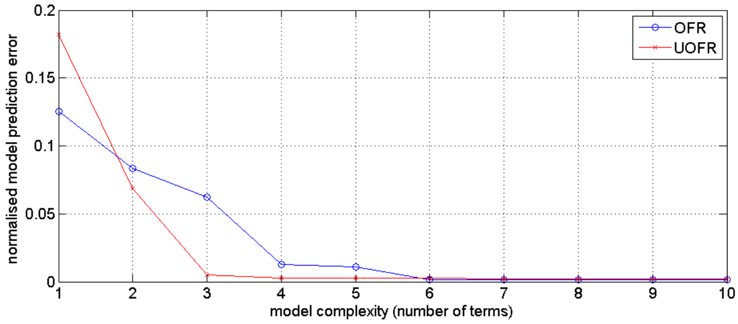


Figure 3 Comparison of the UOFR with the classic OFR for example 2

5.3 Example 3

This example is also taken from (Piroddi & Spinelli, 2003). In this example, the redundant term is selected in the middle of the regression rather than the first step.

Consider the following system.

$$\begin{cases} w(k) = 0.5w(k-1) + 0.8u(k-2) + u^2(k-1) - 0.05w^2(k-2) + 0.5 \\ y(k) = w(k) + \dfrac{1}{1-0.5z^{-1}} e(k) \end{cases} \tag{37}$$

The system was excited by an input defined as

$$u(k) = \frac{0.16}{1-1.6z^{-1}+0.64z^{-2}} v(k) \tag{38}$$

Where $v(k)$ is Gaussian noise $v(k) \sim N(0,1)$. The AR process has a repeat pole 0.8 and the coefficient 0.16 is chosen so that the input signal has a similar amplitude as $v(k)$. Here the noise sequence $e(k)$ is a Gaussian distributed noise with a variance 0.05, that is, $e(k) \sim N(0,0.05)$. The results produced by the standard OFR algorithm are given in Table 5. Observe that several incorrect terms were selected by the OFR algorithm at the middle steps of the forward regression.

Table 5 Results produced by the standard OFR algorithm for example 3

No.	Terms	ERRs	Coefficients	Standard Deviation

1	**y(k-1)**	90.29322	0.108473	0.001819
2	y(k-2)	1.888513	0.002894	0.001484
3	**u^2(k-1)**	1.360357	0.800674	0.002201
4	u(k-1)u(k-2)	3.27943	-0.00893	0.003089
5	u(k-1)	0.672635	-0.00137	0.002658
6	**y^2(k-2)**	1.642735	-0.04993	0.000123
7	**constant**	0.334417	0.494668	0.001746
8	**u(k-2)**	0.487196	-0.79062	0.003265
SERR	--	99.959%	--	--

The system was then identified using the UOFR algorithm. Again the first two derivatives of the cubic B-spline basis function were used to modulate the dependent variable and regressors. The results are given in Table 6. It can be observed that all the correct terms are significant under the new criterion and detected by the UOFR algorithm one by one.

Table 6 Model identified using the UOFR algorithm for example 3

No.	Terms	ERRs	Coefficients	Standard Deviation
1	**u^2(k-1)**	80.28022	0.79721	0.000443
2	**u(k-2)**	10.18902	-0.79724	0.000601
3	**y^2(k-2)**	7.041134	-0.04997	4.34E-05
4	**constant**	2.286183	0.500225	0.000971
5	**y(k-1)**	0.127839	0.103086	0.000648
SERR	--	99.924%	--	--

Figure 4 shows the comparison of the OFR and UOFR regression process.

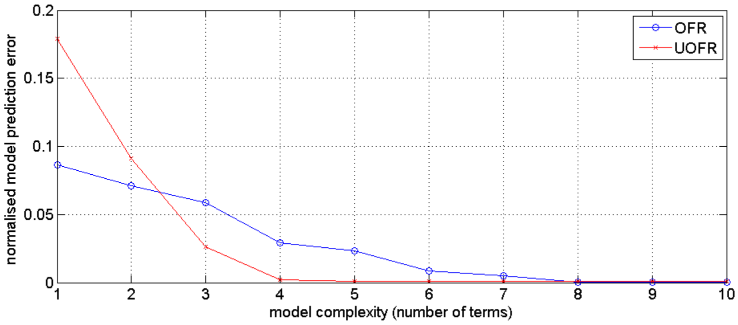


Figure 4 Comparison of the UOFR with the classic OFR for example 3

In all the three examples, the UOFR algorithm identified the correct model structures. Based on these benchmark examples which were deliberately constructed to challenge the model selection, it is fair to draw the conclusion that the performance of the new UOFR is significantly improved by applying the new criterion.

6. Conclusions

System identification involves the detection of the model structure and the estimation of the associated parameters under a specific criterion. The drawbacks of the often used least squares criterion have been extensively discussed. Instead of developing a more complex algorithm, a new stricter measurement of the residuals is proposed to improve the system identification performance. The fitness of a model to the weak derivatives of the observed data is combined with the classic least squares criterion to construct a novel ultra least squares criterion. The ULS criterion considers not only the data themselves but also the relations between and amongst the data points. By modifying the H^m norm, the new ULS criterion possesses a clear physical meaning and is easy to implement.

Based on the ULS criterion, a least squares regression problem can be transformed into an associated ultra least squares problem. The ULS criterion characterises the objective model more accurately and the solution space of the ultra least squares problem possesses better properties than that of the original least squares problem. A novel UOFR algorithm was proposed by combining the ULS criterion with the OFR algorithm to efficiently detect the correct model structure. Simulation results shown that the UOFR algorithm significantly improves the performance of the classic OFR algorithm.

In this paper, the ULS criterion has been used for the UOFR algorithm. However, the application of the ULS criterion is not confined to the UOFR algorithm. The ULS criterion can also be used for other optimization methods where the LS criterion has been used.

Acknowledgements

The authors gratefully acknowledge support from the UK Engineering and Physical Sciences Research Council (EP/K03877X/1).

References

Ayala Solares, J. R., & Wei, H.-L. (2015). Nonlinear model structure detection and parameter estimation using a novel bagging method based on distance correlation metric. *Nonlinear Dynamics*, 1-15. doi: 10.1007/s11071-015-2149-3

Baldacchino, T., Anderson, S. R., & Kadirkamanathan, V. Computational system identification for Bayesian NARMAX modelling. *Automatica, 49*(9), 2641-2651. doi: http://dx.doi.org/10.1016/j.automatica.2013.05.023

Billings, S. A. (2013). *Nonlinear system identification : NARMAX methods in the time, frequency, and spatio-temporal domains*. Hoboken, New Jersey: John Wiley & Sons Ltd.

Billings, S. A., & Wei, H. L. (2008). An adaptive orthogonal search algorithm for model subset selection and non-linear system identification. *International Journal of Control, 81*(5), 714-724. doi: 10.1080/00207170701216311

Bloomfield, P., & Steiger, W. (1980). Least Absolute Deviations Curve-Fitting. *SIAM Journal on Scientific and Statistical Computing, 1*(2), 290-301. doi: doi:10.1137/0901019

Brewer, D., Barenco, M., Callard, R., Hubank, M., & Stark, J. (2008). Fitting ordinary differential equations to short time course data. *Phil. Trans. R. Soc. A, 366*(1865), 519-544. doi: 10.1098/rsta.2007.2108

Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C. (2009). *Introduction to Algorithms* (3rd ed.). London, England: MIT Press.

Guo, Y., Guo, L. Z., Billings, S. A., & Wei, H. L. (2015). An iterative orthogonal forward regression algorithm. *Intern. J. Syst. Sci., 46*(5), 776-789. doi: 10.1080/00207721.2014.981237

Kaizhu, H., King, I., & Lyu, M. R. (2008, 15-19 Dec. 2008). *Direct Zero-Norm Optimization for Feature Selection.* Paper presented at the Data Mining, 2008. ICDM '08. Eighth IEEE International Conference on.

Koza, J. R. (1992). *Genetic programming: on the programming of computers by means of natural selection*. Cambridge, MA, USA: MIT Press.

Li, K., Peng, J.-X., & Bai, E.-W. (2006). A two-stage algorithm for identification of nonlinear dynamic systems. *Automatica, 42*(7), 1189-1197. doi: http://dx.doi.org/10.1016/j.automatica.2006.03.004

Ljung, L. (1987). *System Identification: Theory for the User*. Englewood Cliffs, N.J.: Prentice-Hall, Inc.

Mallat, S. G., & Zhang, Z. (1993). Matching pursuits with time-frequency dictionaries. *Signal Processing, IEEE Transactions on, 41*(12), 3397-3415. doi: 10.1109/78.258082

Mao, K. Z., & Billings, S. A. (1997). Algorithms for minimal model structure detection in nonlinear dynamic system identification. *International Journal of Control, 68*(2), 311-330. doi: 10.1080/002071797223631

Maz'ia, V. G. (1985). *Sobolev Space*. Berlin; New York: Springer-Verlag.

Narula, S. C., & Wellington, J. F. (1982). The Minimum Sum of Absolute Errors Regression: A State of the Art Survey. *International Statistical Review / Revue Internationale de Statistique, 50*(3), 317-326. doi: 10.2307/1402501

Piroddi, L., & Spinelli, W. (2003). An identification algorithm for polynomial NARX models based on simulation error minimization. *International Journal of Control, 76*(17), 1767-1781. doi: 10.1080/00207170310001635419

Preisig, H. A., & Rippin, D. W. T. (1993). Theory and application of the modulating function method - I. Review and theory of the method and theory of the spline-type modulating functions. *Computers & Chemical Engineering, 17*(1), 1-16. doi: http://dx.doi.org/10.1016/0098-1354(93)80001-4

Schmidt, M., & Lipson, H. (2009). Distilling Free-Form Natural Laws from Experimental Data. *Science, 324*(5923), 81-85. doi: 10.1126/science.1165893

Söderström, T. (1989). *System Identification*. New York; London: Prentice Hall.

Appendix A

A set of *k*th order B-spline basis function can be recursively defined as follows given a set of knots s_0, s_1, …, s_N, $N \geq k+1$.

$$B_{i,k}(t) = \frac{t - s_i}{s_{i+k-1} - s_i} B_{i,k-1}(t) + \frac{s_{i+k} - t}{s_{i+k} - s_{i+1}} B_{i+1,k-1}(t) \tag{39}$$

with

$$B_{i,1}(t) = \begin{cases} 1, & s_i \leq t < s_{i+1} \\ 0, & otherwise \end{cases} \tag{40}$$

The first index indicates the position of the B-spline basis function and the second index denotes the order of the B-spline functions. Using N knots, (N-k) *k*th order B-spline basis function can be defined. When the knots is uniformly distributed, function $B_{i,k}(t)$ is a time shift of $B_{1,k}(t)$.

The derivative of a k-th order B-spline can be calculated as

$$\frac{dB_{i,k}(t)}{dt} = (k-1)\left(\frac{B_{i,k-1}(t)}{s_{i+k-1} - s_i} - \frac{B_{i+1,k-1}(t)}{s_{i+k} - s_{i+1}} \right) \tag{41}$$

The higher order derivative of a B-spline basis function can be calculated according to the recursive formula:

$$\frac{d^r B_{i,k}(t)}{dt^r} = (k-1)\begin{pmatrix} \frac{1}{s_{i+k-1} - s_i} & \frac{-1}{s_{i+k} - s_{i+1}} \end{pmatrix} \begin{pmatrix} \frac{d^{r-1} B_{i,k-1}(t)}{dt^{r-1}} \\ \frac{d^{r-1} B_{i+1,k-1}(t)}{dt^{r-1}} \end{pmatrix} \tag{42}$$