Monograph:
Guo, L.Z., Billings, S.A. and Zhu, D.Q. (2006) An extended orthogonal forward regression algorithm for system identification using entropy. Research Report. ACSE Research Report no. 931 . Automatic Control and Systems Engineering, University of Sheffield

An Extended Orthogonal Forward Regression Algorithm for System Identification Using Entropy

Guo, L. Z., Billings, S. A. and Zhu, D. Q.

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

Research Report No. 931
July 2006

An extended orthogonal forward regression algorithm for system identification using entropy

Guo, L. Z., Billings, S. A. and Zhu, D. Q.

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

Abstract

In this paper, a fast identification algorithm for nonlinear dynamic stochastic system identification is presented. The algorithm extends the classical Orthogonal Forward Regression (OFR) algorithm so that instead of using the Error Reduction Ratio (ERR) for term selection, a new optimality criterion —Shannon's Entropy Power Reduction Ratio (EPRR) is introduced to deal with both Gaussian and non-Gaussian signals. It is shown that the new algorithm is both fast and reliable and examples are provided to illustrate the effectiveness of the new approach.

1 Introduction

In system identification and modelling, the Orthogonal Forward Regression (OFR) least-squares algorithm (Billings, Chen, and Korenberg 1989, Chen, Billings, and Luo 1989, and Billings, Korenberg, and Chen 1988) has proved to be an effective algorithm for determining significant model terms or the model structure and the associated parameter estimates. The OFR algorithm involves a stepwise orthogonalisation of the regressors and a forward selection of the relevant terms based on the Error Reduction Ratio (ERR) criterion (Billings, Chen, and Kronenberg 1989). In recent years, many variants of the OFR algorithm have been introduced to improve the performance of the algorithm including D-optimality OFR (Hong and Harris 2001), variable pre-selection OFR (Wei, Billing, and Liu 2004), piecewise linearization (Mao and Billings 1999), minimal model structure detection (Mao and Billings 1997) etc. For the past two decades, the OFR algorithm and its variants have been successfully applied in a variety of fields in system identification and modelling (Aguirre and Billings 1995a,b; Billings, Chen, and Backhouse 1989; Billings, Fadzil, Sulley, and Johnson 1988; Coca, Zheng, Mayhew, and Billings 2000; Coca and Billings 2001; Liu, Kung, and Chao 2001; Balikhin, Zhu, and Billings 2005).

A central part of the conventional OFR algorithm is the Error Reduction Ratio (ERR). The ERR of a term represents the percentage reduction in the total mean square error by including this specific term in the final model. By selecting a term with a maximal ERR value during the implementation of the algorithm, a minimum of the mean square error can be achieved. Therefore, the OFR algorithm is basically minimising a mean square error, the variance of the error in the case of zero-mean variables. This criterion is in line with the traditional approach in stochastic system identification and optimisation problems. The main reason behind this choice of criteria is the assumption that most of the random variables in real-life may be sufficiently described by their second-order statistics, that is their mean and variance. It is well known that Gaussian random variables can be completely defined by the mean value and variance. Therefore, for linear Gaussian systems a criterion based on mean square error would be sufficient to extract all the necessary information from such systems (Papoulis 1991). However, for nonlinear non-Gaussian systems criteria that not only consider the mean value and variance, but also take into account the higher order statistical behaviour of the systems, are much desired. Some recently published papers have addressed this issue (Erdogmus and Principle 2002a, b, Feng, Loparo, and Fang 1997, Ta and DeBrunner 2004, Stoorvogel and van Schuppen 1998). In this paper, differential entropy/ Shannon's entropy power will be introduced as a new criterion for the nonlinear stochastic system identification problem.

Entropy of a given random variable can be considered as a measure of the average information contained in the probability density function (pdf) of that specific random variable. When the entropy of a random variable is minimised, all moments of the error pdf including the second moments are constrained. It follows therefore that entropy as an optimality criterion extends the concept of mean square error. In particular, it can be shown that the differential entropy is proportional to its variance for Gaussian random variables, and thus minimising entropy is equivalent to the minimisation of the variance/mean square error for Gaussian variables. Using entropy as an optimality criterion has desirable advantages in the dynamic system identification problem because minimising mean square error simply constrains the variance of the identification error between the observed response and the model response, which does not guarantee the capture of all the details of the underlying dynamics. In this paper, an extended OFR algorithm is proposed by accommodating differential entropy/ Shannon's entropy power as an objective function instead of conventional mean square error. By making use of Shannon's entropy power inequality, a new quantity —Shannon's Entropy Power Reduction Ratio (EPRR) —is introduced as an extension of the ERR. It is shown that the model terms can be selected in the same way as the classical OFR algorithm, that is sequentially and independently. This is the main difference between the proposed method and existing methods such as in Erdogmus and Priciple (2002a).

The paper is organised as follows. Section 2 presents a brief introduction to differential entropy and Shannon's entropy power including their estimates using a Parzen window and kernel function approach (Shwartz, Zibulevsky, and Schechner 2005). In section 3, the classical OFR algorithm is reviewed first, the relationship between ERR and EPRR is discussed in detail, and the extended OFR algorithm is then presented. Section 4 illustrates the proposed approach using numerical simulations and real data, and finally conclusions are given in section 5.

2 Differential entropy and Shannon's entropy power inequality

2.1 Differential entropy and its estimation

Differential entropy of a random variable X with a probability density function $f_X(\cdot)$ is defined as

$$H(X) = -\int f_X(x) \log f_X(x) dx \tag{1}$$

Generally, differential entropy can be interpreted as a measure of randomness of X. In the case that X is a Gaussian variable with $f_X(x)$ as follows

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma_X} \exp\left(-\frac{(x-\bar{x})^2}{2\sigma_X^2}\right) \tag{2}$$

where $\bar{x}$ is the mean value and σ_X^2 is the variance. Then it is easy to show that the differential entropy $H(X)$ becomes

$$H(X) = \frac{1}{2}\log(2\pi e\sigma_X^2) \tag{3}$$

This clearly shows that minimising differential entropy is equivalent to the minimisation of the variance for Gaussian signals.

For a general random variable with finite variance $\sigma_X^2 < \infty$, Guo, Shamai, and Verdu (2005) have shown that the differential entropy of X, regardless of the distribution of X, can be represented as

$$H(X) = \frac{1}{2}\log(2\pi e\sigma_X^2) - \frac{1}{2}\int_0^\infty \frac{\sigma_X^2}{1+\gamma\sigma_X^2} - E\{(XE\{X|\sqrt{\gamma}X+N\})^2\}d\gamma \tag{4}$$

where $E\{\cdot\}$ denotes the expectation, $E\{X|\sqrt{\gamma}X+N\}$ is the conditional expectation, $N \sim N(0,1)$ is the standard Gaussian which is independent of X. γ is understood as the (gain of the) signal-to-noise ratio of the Gaussian channel whose input is X.

From (4), it can bee observed that the nongaussianness of X is given by one half of the integral of the difference of the minimum mean square errors achievable by a Gaussian input with variance σ_X^2 and by X, respectively.

In practice, differential entropy can be estimated using a Parzen window and kernel function approach (Shwartz, Zibulevsky, and Schechner 2005). Let $x_1, x_2, \cdots, x_N$ be a sample of a random

variable X, then the Parzen window approximation of its probability density function $f_X(\cdot)$ is

$$\hat{f}_X(x) = \frac{1}{N}\sum_{i=1}^{N} K(x - x_i) \tag{5}$$

where $K(\cdot)$ is the kernel function. A Gaussian kernel function is defined as

$$K(x) = \frac{1}{\sqrt{2\pi}h^2}\exp(-\frac{x^2}{2h^2}) \tag{6}$$

where h is the bandwidth of the kernel function. In practice, h can be selected according to Silverman's rule

$$h = 1.06\sigma N^{-1/5} \tag{7}$$

in which σ is the standard deviation of the data.

According to the Parzen window estimation $\hat{f}_X(x)$ in (5) of the probability density function and the definition of differential entropy, an estimation of differential entropy $H(X)$ in (5) can be derived as follows

$$\begin{aligned} \hat{H}(X) &= -\int \frac{1}{N}\sum_{i=1}^{N} K(x - x_i) \cdot \log(\frac{1}{N}\sum_{i=1}^{N} K(x - x_i))dx \\ &= -\frac{1}{N}\sum_{j=1}^{N} \log(\frac{1}{N}\sum_{i=1}^{N} K(x_j - x_i)) \end{aligned} \tag{8}$$

and if the kernel is Gaussian, this yields

$$\hat{H}(X) = -\frac{1}{N}\sum_{j=1}^{N} \log(\frac{1}{\sqrt{2\pi}Nh}\sum_{i=1}^{N} \exp(\frac{(x_j - x_i)^2}{2h^2})) \tag{9}$$

2.2 Shannon's entropy power inequality

Shannon's power entropy inequality gives a bound on the entropy of the sum of independent random variables as follows

$$\exp(2H(X_1 + X_2 + \cdots + X_n)) \geq \exp(2H(X_1)) + \exp(2H(X_2)) + \cdots + \exp(2H(X_n)) \tag{10}$$

where $X_1, X_2, \cdots, X_n$ are n independent random variables. The entropy power is maximum and equal to the variance when the random variable is Gaussian, and it follows that the essence of Shannon's entropy power inequality (10) is that the sum of independent variables tends to be "more Gaussian" than individual components. Note that if all of the random variables are Gaussian, (3) yields

$$2\pi e\sigma^2_{\sum_i^n X_i} = \exp(\log(2\pi e\sigma^2_{\sum_i^n X_i})) \geq \sum_i^n \exp(\log(2\pi e\sigma^2_{X_i})) \tag{11}$$

Note that because the entropy power of X is $\exp(2H(X))$, from (3), $\exp(2H(x)) = 2\pi e\sigma_X^2$ for X being Gaussian. This indicates that minimising entropy power is also equivalent to the minimisation of the variance for Gaussian signals.

For given samples $x_1, x_2, \cdots, x_N$ of a random variable X, Shannon's entropy power of X can be approximately calculated according to the estimate $\hat{H}(X)$ in (8) of differential entropy as follows

$$\exp(2H(X)) \approx \exp(2\hat{H}(X)) = \exp(-2\frac{1}{N}\sum_{j=1}^{N}\log(\frac{1}{N}\sum_{i=1}^{N}K(x_j - x_i))) \tag{12}$$

3 An extended orthogonal forward regression least squares algorithm

3.1 The classical OFR least-squares algorithm

Let $p_0, p_1, \cdots, p_n$ be random variables and y the output response of a system. Without loss of generality, all involved variables are assumed to be zero-mean. Assume that there is a subset I of $\{0, 1, \cdots, n\}$ such that a linear relationship

$$y = \sum_{i\in I}\theta_i p_i + \xi \tag{13}$$

exists, in which ξ is an independent noise variable with zero mean and finite variance. Given a set of observations, the system modelling problem of interest is to determine the subset I and the values of θ_i. The OFR algorithm for this problem involves three steps

- Orthogonalise the regressors to remove the correlations between these variables;
- Select significant terms using the ERR as a criterion;
- Estimate the corresponding parameters for the selected terms.

Formally, the classical OFR least-squares algorithm can be stated as follows (Billings, Korenberg, and Chen 1988).

Let $p_0(t), p_1(t), \cdots, p_n(t)$ and $y(t)$, $t = 1, 2, \cdots, N$ be the series of observations. Denote $Y = (y(1), y(2), \cdots, y(N))^T$ and $P_i = (p_i(1), p_i(2), \cdots, p_i(N))^T$, $i = 0, 1, \cdots, n$, then the following linear regression model can be formed

$$Y = P\theta + \Xi \tag{14}$$

where $P = (P_0, P_1, \cdots, P_N)$ is the regression matrix, $\theta = (\theta_1, \theta_2, \cdots, \theta_n)^T$ represents the unknown parameters to be estimated, and $\Xi = (\xi(1), \xi(2), \cdots, \xi(N))^T$ is some modelling error vector. The three steps in the OFR algorithm are

1) ORTHOGONALISATION The orthogonal decomposition $P = WA$, where A is an $(n + 1) \times (n + 1)$ upper triangular matrix with unity diagonal elements, of the regression matrix P provides an alternative representation of eqn. (14)

$$Y = P\theta + \Xi = WA\theta + \Xi = Wg + \Xi \tag{15}$$

where W is an $N \times (n + 1)$ matrix with orthogonal columns W_i such that $W^T W = D$ in which D is an $(n+1)\times(n+1)$ diagonal matrix with elements $d_i =< W_i, W_i >$, $i = 0, 1, \cdots, n$. Note that $< \cdot, \cdot >$ denotes the inner product so that $d_i =< W_i, W_i >= \sum_{t=1}^{N} w_i(t) w_i(t)$, $i = 0, 1, \cdots, n$.

2) TERM SELECTION The orthogonal least squares solution to g is given by

$$\hat{g}_i = \frac{< Y, W_i >}{< W_i, W_i >} = \frac{W_i^T Y}{W_i^T W_i}, i = 0, 1, \cdots, n \tag{16}$$

The fraction of variance not explained by a regression of Y on Wg is

$$\frac{< \Xi, \Xi >}{< Y, Y >} = \frac{< Y - Wg, Y - Wg >}{< Y, Y >} = \frac{< Y, Y > - < Wg, Wg >}{< Y, Y >} \tag{17}$$

Thus the error reduction ratio (ERR) caused by term i, $i = 0, 1, \cdots, n$ is defined as

$$ERR_i = \frac{< W_i g_i, W_i g_i >}{< Y, Y >} \tag{18}$$

The OFR least-squares algorithm selects the subset I, that is a subset of regressors in a forward-regression manner by maximising the contribution of a regressor to the explained desired response variance, that is its ERR.

3) PARAMETER ESTIMATION Once the parameters g_i, $i \in I$ have been estimated using (16) the parameters θ_i, $i \in I$ in the regression equation (13) can be calculated as

$$\hat{\theta} = A^{-1}\hat{g} \tag{19}$$

From the definition of ERR (18), it can be observed that the OFR is equivalent to maximising the product moment correlation coefficient. In fact, the product moment correlation coefficient ρ_i of term i satisfies

$$\rho_i^2 = \frac{< Y, W_i >^2}{< Y, Y >< W_i, W_i >} = \frac{\frac{<Y,W_i>^2}{<W_i,W_i>^2} < W_i, W_i >}{< Y, Y >} = \frac{< W_i g_i, W_i g_i >}{< Y, Y >} = ERR_i \tag{20}$$

3.2 Error reduction ratio (ERR) vs. Shannon's entropy power reduction ratio (EPRR)

Consider the above step 2), that is the Term Selection. This term selection procedure is actually based on the ERR values of each candidate terms. The rationale can be explained as follows (Billings, Chen, and Kronenberg 1989).

For the regression problem (14), the orthogonalised version is that of equation (15). Taking the inner product to (15) gives

$$< Y, Y >=< Wg, Y > + < \Xi, Y > \tag{21}$$

Substituting $Y = Wg + \Xi$ into the right hand side of (21) yields

$$< Y, Y >=< Wg, Wg > + < \Xi, \Xi >= \sum_{i=0}^{n} < W_i g_i, W_i g_i > + < \Xi, \Xi > \tag{22}$$

that is

$$\sum_{t=1}^{N} y^2(t) = \sum_{i=0}^{n} \sum_{t=1}^{N} g_i^2 w_i^2(t) + \sum_{t=1}^{N} \xi^2(t) \tag{23}$$

assuming that $\xi(t)$ is an independent noise sequence with zero mean and finite variance, and the orthogonality property of columns of the matrix W holds. The maximum mean square error is achieved when no terms are selected to give

$$< \Xi, \Xi >=< Y, Y > \tag{24}$$

Therefore, the reduction in mean square error by including a term $W_i g_i$ (equivalently $P_i\theta_i$) in the model will be equal to

$$< W_i g_i, W_i g_i >= \sum_{t=1}^{N} g_i^2 w_i^2(t) \tag{25}$$

It follows that the reduction ratio as a result of including the term $P_i\theta_i$ is the percentage reduction in the total mean square error

$$ERR_i = \frac{< W_i g_i, W_i g_i >}{< Y, Y >} \tag{26}$$

which is defined as error reduction ratio (ERR) as (18). By selecting a term with a maximal ERR value at a time during the implementation of the algorithm, a minimum of the mean square error can be achieved. This clearly indicates that the above algorithm minimises the mean square error

$$< \Xi, \Xi >=< Y - Wg, Y - Wg >= \sum_{t=1}^{N}(y(t) - \sum_{i=0}^{m} g_i w_i(t))^2 \tag{27}$$

for any non-negative integer $m \leq n$ because of the orthogonality property. Note that under the assumption that ξ has zero mean and finite variance, minimising the mean square error $< \Xi, \Xi >$ is equivalent to minimising the variance of ξ.

Now assume that all involved random variables w_i are jointly Gaussian, then they are mutually independent because the orthogonality of Gaussian variables implies independence. It follows that Shannon's power entropy inequality holds, that is

$$\exp(2H(y)) \geq \exp(2H(w_1 g_1)) + \exp(2H(w_2 g_2)) + \cdots + \exp(2H(w_n g_n)) + \exp(2H(\xi)) \tag{28}$$

This relationship can be explained as (22). The maximum entropy power of error (equivalently differential entropy of error) is achieved when no terms are selected to give

$$\exp(2H(\xi)) = \exp(2H(y)) \tag{29}$$

Equation (29) indicates that Shannon's entropy power of error can never go beyond the Shannon's entropy power of y for Gaussian stochastic systems. As in the ERR approach, to obtain a minimum of Shannon's entropy power or simply differential entropy of error, those terms with maximal Shannon's entropy values should be included in the regression model. As mentioned earlier, minimising differential entropy is equivalent to the minimisation of the variance for Gaussian

variables while the ERR approach is designed to minimise such a variance. Since differential entropy has a more general meaning than that of variance for any random variables, it can be used to measure and form a term selection criterion for the general stochastic system identification problem. Following the above discussion, a new criterion, which is called the EPRR (Entropy Power Reduction Ratio), is proposed in this paper as follows.

Notice that Shannon's entropy power reduction ratio is, as a result of including the term $w_i g_i$ or $p_i \theta_i$, the percentage reduction in the total entropy power

$$EPRR_i = \frac{\exp(2H(w_i g_i))}{\exp(2H(y))} \tag{30}$$

When w_i and y are zero-mean and Gaussian, using (3) and $H(w_i g_i) = H(w_i) + \log|g_i|$, EPRR can be expressed as

$$\begin{aligned} EPRR_i &= \frac{\exp(2H(w_i g_i))}{\exp(2H(y))} \\ &= \frac{\exp(\log(2\pi e\sigma_{w_i}^2) + \log(g_i^2))}{\exp(\log(2\pi e\sigma_y^2))} \\ &= \frac{\sigma_{w_i}^2 g_i^2}{\sigma_y^2} \\ &= \frac{< W_i g_i, W_i g_i >}{< Y, Y >} \\ &= ERR_i \end{aligned} \tag{31}$$

From the definitions of ERR and EPRR, and (4), it can be shown that the EPRR for general random variables can be expressed as follows

$$\begin{aligned} EPRR_i = ERR_i \cdot \exp(\int_0^\infty & \frac{< Y, Y > - < W_i, W_i >}{(1+ < Y, Y >)(1+ < W_i, W_i >)} \\ & +(E\{(y - E\{\sqrt{\gamma}y + N_y\})^2\} - E\{(w_i - E\{w_i|\sqrt{\gamma}w_i + N_{w_i}\})^2\})d\gamma) \end{aligned} \tag{32}$$

In practice, the EPRR values can be approximately calculated using (12)

$$EPRR_i \approx g_i^2 \cdot \left(\prod_{t=1}^{N} \frac{\sum_{k=1}^{n} K(y(t) - y(k))}{\sum_{k=1}^{N} K(w_i(t) - w_i(k))} \right)^{\frac{2}{N}} \tag{33}$$

3.3 A summary of the proposed extended OFR algorithm

Let $Y = (y(1), y(2), \cdots, y(N))^T$ and $P_i = (p_i(1), p_i(2), \cdots, p_i(N))^T$, $i = 0, 1, \cdots, n$ be defined as before, where N is the length of the samples. Then the extended OFR algorithm can now be summarised as follows

Step 1 For $i = 0, 1, \cdots, n$, let $W_1^{(i)} = P_i$ and compute the coefficients g

$$g_0^{(i)} = \frac{< Y, W_0^{(i)} >}{< W_0^{(i)}, W_0^{(i)} >} \tag{34}$$

and their corresponding EPRR using (33)

$$EPRR_0^{(i)} \approx (g_0^{(i)})^2 \cdot \left(\prod_{t=1}^{N} \frac{\sum_{k=1}^{n} K(y(t) - y(k))}{\sum_{k=1}^{N} K(w_0^{(i)}(t) - w_0^{(i)}(k))} \right)^{\frac{2}{N}} \tag{35}$$

Find the index with a maximal EPRR value $h_0 = arg[max(EPRR_0^{(i)}, 0 \leq i \leq n)]$ and select the corresponding term $W_0 = W_0^{(h_0)}$.

Step l+1 , $l \geq 1$ For $0 \leq i \leq n, i \neq h_0, h_1, \cdots, h_{l-1}$, calculate the orthogonal projection of P_i on the linear subspace spanned by $W_0, W_1, \cdots, W_{l-1}$ as follows

$$W_l^{(i)} = P_i - \sum_{j=0}^{l-1} \frac{< P_i, W_j >}{< W_j, W_j >} W_j \tag{36}$$

Compute the coefficients g

$$g_l^{(i)} = \frac{< Y, W_l^{(i)} >}{< W_l^{(i)}, W_l^{(i)} >} \tag{37}$$

and their corresponding EPRR using (33) again

$$EPRR_l^{(i)} \approx (g_l^{(i)})^2 \cdot \left(\prod_{t=1}^{N} \frac{\sum_{k=1}^{n} K(y(t) - y(k)}{\sum_{k=1}^{N} K(w_l^{(i)}(t) - w_l^{(i)}(k))} \right)^{\frac{2}{N}} \tag{38}$$

Find the index with a maximal EPRR value $h_l = arg[max(EPRR_l^{(i)}, 0 \leq i \leq n), i \neq h_0, h_1, \cdots, h_{l-1}]$ and select the corresponding term $W_l = W_l^{(h_l)}$.

The procedure is terminated at the n_s^{th} step when

$$1 - \sum_{i=1}^{n_s} EPRR_i < \rho \tag{39}$$

where $0 < \rho < 1$ is a prescribed tolerance. This gives a subset model containing n_s significant terms.

Following the determination of the most significant terms $W_0, W_1, \cdots, W_{n_s}$ according to the above steps, estimates of the parameters $\theta_i, i = 0, 1, \cdots, n_s$ can be obtained. This can be done using the following equations

$$\theta_i = \sum_{j=i}^{n_s} g_j v_j \tag{40}$$

where $g_i, i = 0, 1, \cdots, n_s$ are calculated as follows

$$g_i = \frac{< Y, W_i >}{< W_i, W_i >} = \frac{\sum_{t=1}^{N} y(t) w_i(t)}{\sum_{t=1}^{N} w_i^2(t)}, i = 0, 1, \cdots, n_s \tag{41}$$

and

$$\begin{aligned} v_i &= 1 \\ v_j &= -\sum_{k=i}^{j-1} \alpha_{k,j} v_k, j = i+1, \cdots, n \end{aligned} \tag{42}$$

in which

$$\alpha_{k,j} = \frac{< P_j, W_k >}{< W_k, W_k >} = \frac{\sum_{t=1}^{N} p_j(t) w_k(t)}{\sum_{t=1}^{N} w_k^2(t)}, k = 0, 1, \cdots, j-1 \tag{43}$$

Remark 1 It is worth noting that the algorithm is developed under the assumption that the involved variables are mutually independent or uncorrelated with a jointly Gaussian distribution. There are some recent results about Shannon's entropy power inequality for dependent variables (Johnson 2004, Takano 1996). The proposed algorithm projects the regression matrix into an orthogonal space which implies the involved variables are uncorrelated. Although uncorrelationness does not mean independence, it is, in practice, a good estimation of independence. The examples in this paper show that the extended OFR algorithm works well with this uncorrelationness.

4 Numerical simulations

4.1 Example 1: A linear system with Gaussian and non-Gaussian noise

Consider the following simple AR model

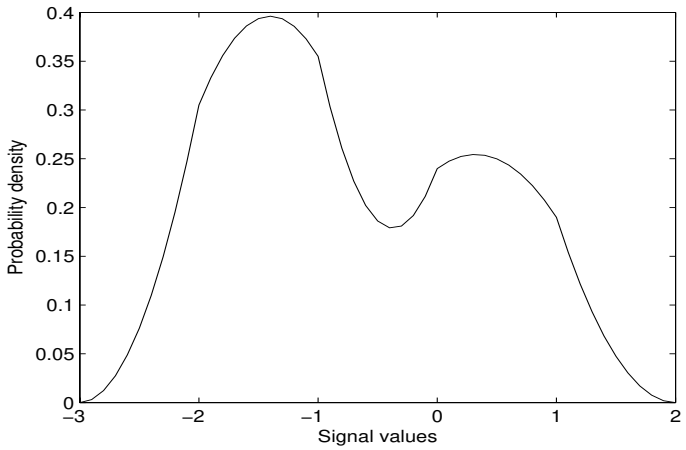

Figure 1: Example 1: Distribution of a non-Gaussian noise signal with two peaks

Table 1: Example1: The terms and parameters of the final model using the OFR algorithm with Gaussian noise

Terms	Estimates	ERR
$y(t-1)$	1.7749e+000	3.7352e-001
$y(t-2)$	-1.9534e+000	2.3680e-001
$y(t-3)$	1.3874e+000	1.8680e-001
$u(t-1)$	9.9466e-001	1.2390e-001
$y(t-4)$	-4.7615e-001	4.5940e-002

$$y(t) = a_1 y(t-1) + a_2 y(t-2) + a_3 y(t-3) + a_4 y(t-4) + a_5 u(t-1) + e(t) \tag{44}$$

where $a_1 = 1.8$, $a_2 = -1.99$, $a_3 = 1.422$, $a_4 = 0.493$, and $a_5 = 1.0$. In order to apply the proposed extended OFR algorithm, system eqn. (44) was simulated where $u(t)$ was a uniformly distributed random signal on the interval $[0, 1]$ and two sets of data were collected for the cases where $e(t)$ had a Gaussian ($\sim N(0, 0.3^2)$) and a non-Gaussian probability distribution. Fig. (1) shows the probability density function of the applied non-Gaussian noise.

The set of terms in an initial candidate model was set to be $\{1, y(t-1), y(t-2), y(t-3), y(t-4), y(t-5), y(t-6), u(t-1)\}$. The identified results using the original and the extended OFR algorithms are shown in Tables 1 to 4.

Table 2: Example 1: The terms and parameters of the final model using the extended OFR algorithm with Gaussian noise

Terms	Estimates	ERR	EPRR
$y(t-1)$	1.7749e+000	3.7352e-001	3.7359e-001
$y(t-2)$	-1.9534e+000	2.3680e-001	2.3814e-001
$y(t-3)$	1.3874e+000	1.8680e-001	1.8806e-001
$u(t-1)$	9.9466e-001	1.2390e-001	1.0005e-001
$y(t-4)$	-4.7615e-001	4.5940e-002	4.6161e-002

From Tables 1 and 2 it can be observed that the selected significant terms (which are the terms in the original system model) and the parameter estimates are exactly the same (which are very close to the real values) for both algorithms with Gaussian noise, and in this case the ERR values and EPRR values for each selected terms are almost identical. This is because the noise is Gaussian, which verifies the theoretical results, that is that minimising entropy is equivalent to the minimisation of variance for Gaussian signals. Furthermore, the estimated higher moments of the errors from both algorithms are coincident (Fig. (3)) : $\mu_3 = 1.4614e-003, \mu_4 = 2.6130e-002, \mu_5 = 1.7881e-003, \mu_6 = 1.3225e-002, \mu_7 = 2.1976e-003$. Note that the errors from both algorithms for Gaussian noise have approximated to a Gaussian distribution. This can be observed from Fig. (2) which is the error probability density function estimate calculated using Parzen windowing and Gaussian kernel functions for both algorithms. Therefore, the extended OFR algorithm is equivalent to the original OFR for such Gaussian systems. However, for non-Gaussian noise the results are different. Tables 3 and 4 show that the ERR values and EPRR values are different although all of the correct terms have been selected in the right order by both algorithms. An investigation shows that the higher moments for both algorithms are slightly different as well (Fig. (4)): for the extended OFR algorithm $\mu_3 = 2.4736e-001, \mu_4 = 3.4510e+000, \mu_5 = 1.6009e+000, \mu_6 = 1.3223e+001, \mu_7 = 9.8424e+000$, and for the original OFR algorithm $\mu_3 = 2.4856e-001, \mu_4 = 3.4393e+000, \mu_5 = 1.6140e+000, \mu_6 = 1.3129e+001, \mu_7 = 9.9515e+000$. It can also be observed from Fig. (5), that there is a slight difference between the estimated probability density functions of errors from the two algorithms. Those observations reveal that instead of the minimisation of conventional mean squares error, minimising entropy/entropy power does introduce more information into the solution.

Table 3: Example1: The terms and parameters of the final model using the OFR algorithm with non-Gaussian noise

Terms	Estimates	ERR
$y(t-1)$	1.9394e+000	5.7413e-001
$y(t-2)$	-2.1653e+000	1.0629e-001
$y(t-3)$	1.5944e+000	1.9971e-001
$y(t-4)$	-5.1538e-001	2.2979e-002
$u(t-1)$	1.0327e+000	2.0446e-002
$y(t-6)$	8.6247e-002	2.3946e-003
$y(t-5)$	-3.2717e-002	2.4375e-005
$u(t-2)$	-1.3459e-002	2.7265e-006

Table 4: Example 1: The terms and parameters of the final model using the extended OFR algorithm with non-Gaussian noise

Terms	Estimates	ERR	EPRR
$y(t-1)$	1.9417e+000	5.7413e-001	5.7422e-001
$y(t-2)$	-2.1801e+000	1.0629e-001	1.5813e-001
$y(t-3)$	1.6272e+000	1.9971e-001	2.6595e-001
$y(t-4)$	-5.5539e-001	2.2979e-002	3.4981e-002
$u(t-1)$	1.0322e+000	2.0446e-002	2.6012e-002
$y(t-6)$	7.2360e-002	2.3946e-003	3.4763e-003

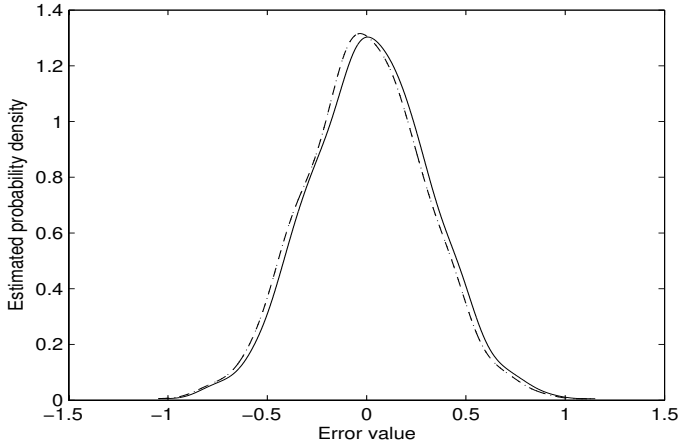

Figure 2: Example 1: Estimated probability distribution of errors of models from original OFR (dotted) and extended OFR algorithm(dashed) for Gaussian noise (the solid is the pdf of the original noise)

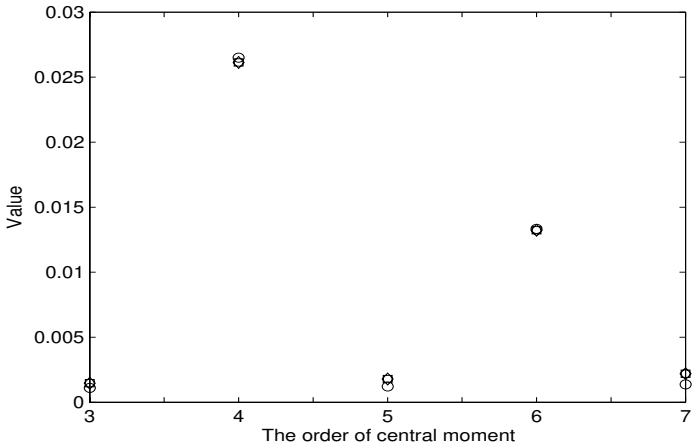

Figure 3: Example 1: Higher central moments of errors of models from conventional OFR (diamond) and extended OFR algorithms (square) for Gaussian noise (the circles denote the higher central moments of the original noise)

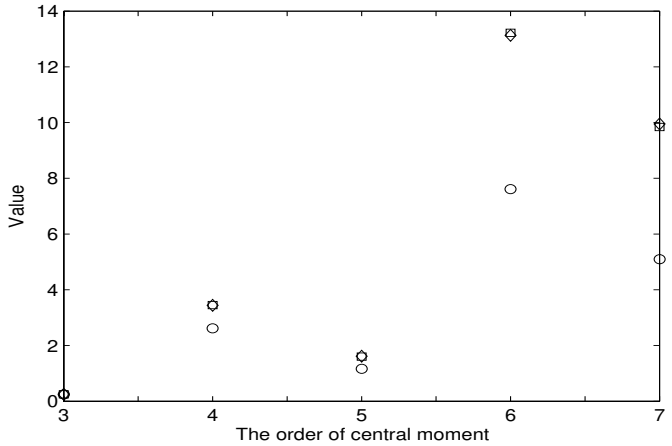

Figure 4: Example 1: Higher central moments of errors of models from conventional OFR (diamond) and extended OFR algorithms (square) for non-Gaussian noise (the circles denote the higher central moments of the original noise)

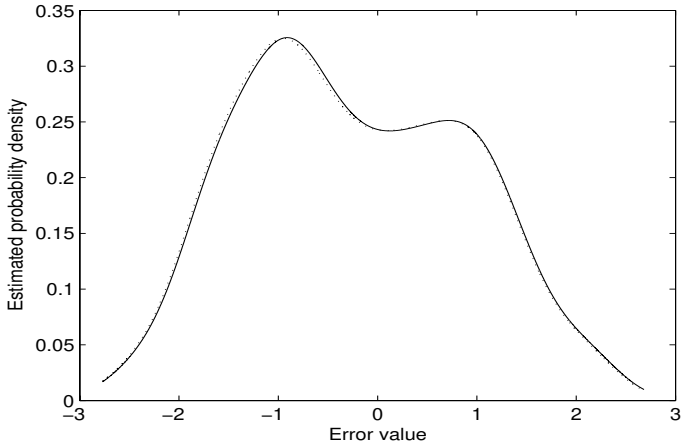

Figure 5: Example 1: Estimated probability distribution of errors of models from original OFR (dotted) and extended OFR algorithms (solid) for non-Gaussian noise

4.2 Prediction of the Dst index in Geophysics

Dst is a geomagnetic index which monitors the world wide magnetic storm level. It is generally constructed by averaging the horizontal component of the geomagnetic field from mid-latitude and equatorial magnetograms from all over the world. Negative Dst values indicate a magnetic storm is in progress, the more negative Dst is the more intense the magnetic storm. The negative deflections in the Dst index are caused by the storm time ring current which flows around the Earth from east to west in the equatorial plane. The ring current results from the differential gradient and curvature drifts of electrons and protons in the near Earth region and its strength is coupled to the solar wind conditions. Only when there is an eastward electric field in the solar wind which corresponds to a southward interplanetary magnetic field (IMF) is there any significant ring current injection resulting in a negative change to the Dst index. In addition to the ring current, other currents such as the current along the magnetopause, a boundary between the terrestrial magnetosphere and the solar wind, also provide some contribution to the evolution of the Dst index. Due to the importance of the Dst index, it is highly desirable to be able to predict its values. However, the relationship between the Dst index and the evolution of the ring, magnetopause and magnetotail currents under the influence of the solar wind is extremely complicated. In this study a low dimensional input-output dynamical system model of the magnetosphere will be identified directly from observations using the new approach, in which the input is the product of the solar wind velocity V and the southward component of the solar wind magnetic field Bs and the output is the Dst index. Note that the input was calculated using the plasma velocity and magnetic field measurements from the Wind satellite. The effects of other factors on the system will be neglected in the present study.

A set of 1580 pairs of samples of the input and output were collected from satellites with a sampling period of one hour and the data are shown in Fig. (6). A polynomial NARMAX model

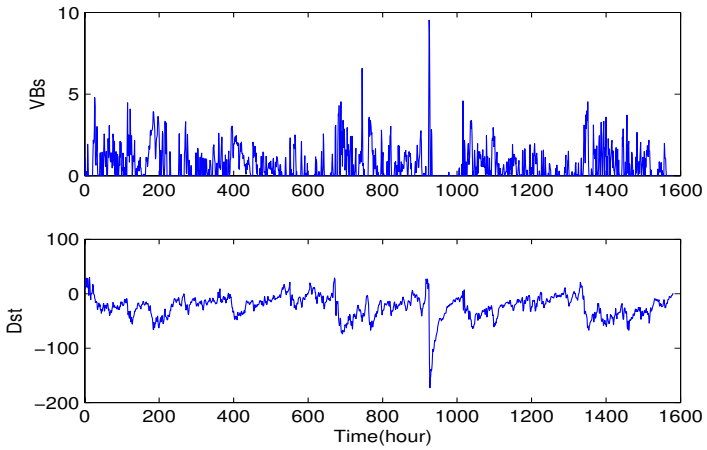

Figure 6: Example 2: The input and output data

Table 5: Example2: The terms and parameters of the final model using the extended OFR algorithm

Terms	Estimates	ERR	EPRR
$y(t-1)$	1.0569e+000	9.9752e-001	9.9838e-001
$y^2(t-2)$	-1.3591e-001	3.7412e-004	1.3399e-002
$y(t-2)u(t-1)$	-4.1862e-001	8.5993e-004	1.3130e-002
$y(t-1)u(t-2)$	1.6460e-001	2.0221e-004	4.9254e-003
$y(t-4)y(t-7)$	7.5858e-002	3.5534e-005	8.2560e-004
$u(t-1)$	3.3662e-001	3.0053e-005	4.3869e-004
$y(t-7)u(t-1)$	-3.6974e-001	4.9174e-005	9.6667e-004

with input lag 3 and output lag 7 and nonlinear degree 2 was used to fit the measured data. The identified final model using the first 500 pairs of data and the extended OFR algorithm is shown in Table 5, which indicates that only 7 out of 66 candidate terms are selected to yields a very simple nonlinear model. A comparison of the test results, that is the output data, the model predicted output and one-step-ahead predicted output, are shown in Fig. (7). The error probability density function estimate calculated using Parzen windowing and Gaussian kernel functions is shown in Fig. (8). The higher moments for the new algorithm are: $\mu_3 = -8.3544e+001, \mu_4 = 1.0190e+005, \mu_5 = -9.9569e+005, \mu_6 = 1.8408e+008, \mu_7 = -4.1276e+009$.

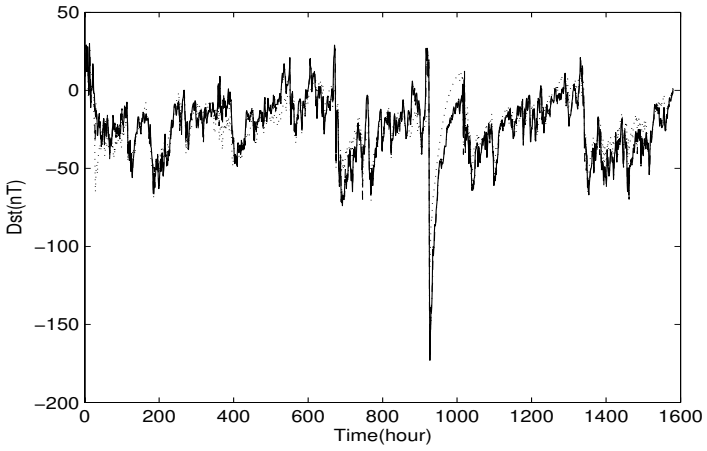

Figure 7: Example 2: Measurement (solid), model predicted output (dotted) and one-step-ahead predicted output(dashed)

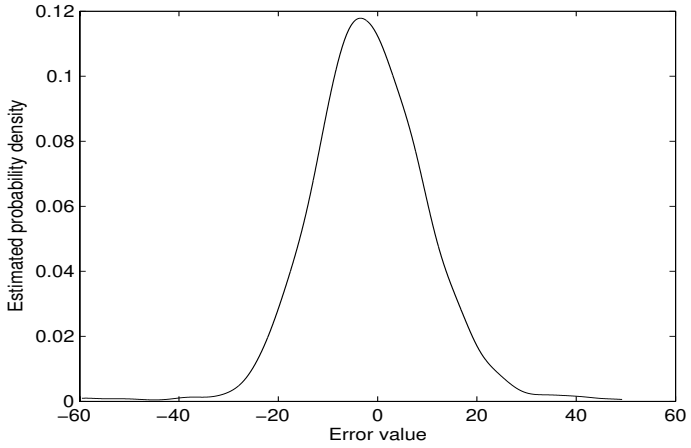

Figure 8: Example 2: Estimated probability distribution of errors of the model from extended OFR algorithms

From the test results some observations can be made:

1). Both the model predicted and one-step-ahead predicted outputs are good. The model predicted output is not as good as the one-step-ahead predicted output but this is to be expected because model predicted output is a much more severe test than one-step-ahead predictions. However, the model predicted output shows good long-term predictions and gives more confidence in the identified model comparing with the one-step-ahead prediction. Such a model with long-term predictive ability provides a basis for any further analysis and control of the underlying dynamics.

2). Fig. (8) shows that the modelling errors have approximated to a Gaussian distribution, which is different from the estimated pdf from the conventional OFR algorithm.

3). From the final model it can be observed that the main influence of the input (VBs) on the system current output (Dst) is approximately two hours behind (only $u(t-1), u(t-2)$ appear in the final model) whilst the influence of the past Dst index on the Dst index is approximately seven hours behind ($y(t-7)$ appears in the identified model).

4). The discrepancy between the model predicted values and the measured values of Dst may result from the errors between the real values of the entropy/Shannon entropy power and the estimated values using Parzen windowing and Gaussian kernel functions, high dependence between some candidate model terms, and/or the other factors which actually affect the system dynamics but were not included in the current model.

5 Conclusions

An extended OFR algorithm for the identification of both the model terms or structure and the unknown parameters of non-linear stochastic systems with Gaussian and non-Gaussian noise has been introduced. It has been shown that by using entropy/entropy power as an optimality criterion, the identification ability of the conventional OFR algorithm can be enhanced. The introduction of EPRR in the extended OFR algorithm not only retains the advantages of the OFR algorithm, that is terms can be selected in a fast, simple, and independent way, but also takes into account the higher statistical behaviour of the systems and signals. In this sense the proposed algorithm can be considered as an extension of the conventional OFR algorithm. The method has been tested on both simulated and real data and was shown to perform very well.

6 Acknowledgement

The authors gratefully acknowledge financial support from EPSRC (UK).

References

[1] Aguirre, L. A. and Billings, S. A., (1995a) Improved structure selection for nonlinear models based on term clustering, *Int. J. Contr.*, Vol. 62, No.3, pp. 569-587.

[2] Aguirre, L. A. and Billings, S. A., (1995b) Retrieving dynamical invariants from chaotic data using narmax models, *Int. J. Bifurcation and Chaos*, Vol. 5, No. 2, pp. 449-474.

[3] Balikhin, M. A., Zhu, D., and Billings, S. A., (2005) Time domain identification of nonlinear systems: from the measurements to continuous differential equations, *The 2005 Joint Assembly Meeting of AGU, SEG, NABS and SPD/AAS*, New Orleans, Louisiana.

[4] Billings, S. A., Chen, S., and Kronenberg, M. J., (1989) Identification of MIMO nonlinear systems using a forward-regression orthogonal estimator, *Int. J. Contr.*, Vol. 49, pp. 2157-2189.

[5] Billings, S. A., Chen, S., and Backhouse, R. J., (1989), The identification of linear and nonlinear models of a turbocharged automotive diesel engine, *Mechanical Systems and Signal Processing*, Vol. 3, pp. 123-142.

[6] Billings, S. A., Fadzil, M. B., Sulley, J., and Johnson P. M., (1988), Identification of a nonlinear difference equation model of an industrial diesel generator, *Mechanical Systems and Signal Processing*, Vol. 2, pp. 59-76.

[7] Billings, S. A., Korenberg, M. J., and Chen, S., (1988) Identification of non-linear output-affine systems using an orthogonal least-squares algorithm, *Int. J. Syst. Sci.*, Vol. 19, pp. 59-76.

[8] Chen, S., Billings, S. A., and Luo, W., (1989) Orthogonal least squares methods and their application to non-linear system identification, *International Journal of Control*, Vol. 50, No. 5, pp. 1873-1896.

[9] Coca,D. and Billings, S. A., (2001) Identification of coupled map lattice models of complex spatio-temporal pattern, *Phys. Lett.*, A287, pp. 65-73.

[10] Coca, D., Zheng, Y., Mayhew, J., and Billings, S. A. (2000) Nonlinear system identification and analysis of complex dynamical behaviour in reflected light measurements of vasomotion, *International Journal of Bifurcation and Chaos*, Vol. 10, No. 2, pp. 461-476.

[11] Erdogmus, D. and Principe, J. C., (2002a) An error-entropy minimization algorithm for supervised training of nonlinear adaptive systems, *IEEE Trans. on Signal Processing*, Vol. 50, No. 7, pp. 1780-1786.

[12] Erdogmus, D. and Principe, J. C., (2002b) Generalized information potential criterion for adaptive system training, *IEEE Trans. on Neural Networks*, Vol. 13, No. 5, pp. 1035-1044.

[13] Feng, X, Loparo, K, and Fang, Y., (1997) Optimal state estimation for stochastic systems: An information theoretic approach, *IEEE Trans. Automatic Control*, Vol. 42, pp. 771-785.

[14] Guo, D., Shamai, S., and Verdu, S., (2005) Mutual information and minimum mean-square error in Gaussian channels, *IEEE Trans. On Information Theory*, Vol. 51, No.4, pp. 1261-1282.

[15] Hong, X. and Harris, C. J., (2001) Nonlinear model structure design and construction using orthogonal least squares and d-optimality design, *IEEE Transactions on Neural Networks*, vol. 13, no. 5, pp. 12451250.

[16] Johnson, O. (2004) A conditional entropy power inequality for dependent variables, *IEEE Trans. Information Theory*, Vol. 50, No. 8, pp. 1581-1583.

[17] Liu, J. J., Kung, I. C., and Chao, H. C., (2001) Speed estimate of induction motor using a nonlinear identification technique, *Proceedings of the National Science Council*, Republic of China. Part A, vol. 25, no. 2, pp. 107-114.

[18] Mao, K. Z., and Billings, S. A., (1997) Algorithms for minimal model structure detection in nonlinear dynamic system identification, *Int. J. Control*, Vol. 68, pp. 311330.

[19] Mao, K. Z., and Billings, S. A., (1999) Variable selection in nonlinear systems modelling, *Mechanical Systems and Signal Processing*, Vol. 13, pp. 351366.

[20] Papoulis, A., (1991)*Probability, Random Variables, and Stochastic Processes*, New York:McGraw-Hill.

[21] Shwartz, S., Zibulevsky, M., and Schechner, Y. Y., (2005) Fast kernel entropy estimation and optimization, *Signal Processing*, Vol. 85, No. 5, pp. 1045-1058.

[22] Stoorvogel, A. A. and van Schuppen, J. H., (1998) Approximation problems with the divergence criterion for Gaussian variables and processes, *System & Control Letters*, Vol. 35, No. 4, pp. 207-218.

[23] Ta, M. and DeBrunner, V., (2004) Minimum entropy estimation as a near maximum-likelihood method and its application in system identification with non-Gaussian noise, in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing(ICASSP '04)*, Vol. 2, pp. 17-21.

[24] Takano, S. (1996) The inequalities of Fisher information and entropy power for dependent variables, in *Proceedings of 7th Japan-Russia Symposium on Probability Theory and Statistics*, Tokyo, pp.460-470.

[25] Wei, H. L., Billings, S. A., and Liu, J., (2004) Term and variable selection for nonlinear system identification, *Int. J. Control*, Vol. 77, No. 1, pp. 86-110.