Circuits, Systems, and Signal Processing (2022) 41:915–932
https://doi.org/10.1007/s00034-021-01809-3

Identification of the ARX Model with Random Impulse Noise Based on Forgetting Factor Multi-error Information Entropy

Shaoxue Jing[1]

Received: 22 November 2020 / Revised: 22 July 2021 / Accepted: 23 July 2021 /
Published online: 12 August 2021

Abstract

Entropy has been widely applied in system identification in the last decade. In this paper, a novel stochastic gradient algorithm based on minimum Shannon entropy is proposed. Though needing less computation than the mean square error algorithm, the traditional stochastic gradient algorithm converges relatively slowly. To make the convergence faster, a multi-error method and a forgetting factor are integrated into the algorithm. The scalar error is replaced by a vector error with stacked errors. Further, a simple step size method is proposed and a forgetting factor is adopted to adjust the step size. The proposed algorithm is utilized to estimate the parameters of an ARX model with random impulse noise. Several numerical solutions and case study indicate that the proposed algorithm can obtain more accurate estimates than the traditional gradient algorithm and has a faster convergence speed.

Keywords ARX model · Parameter estimation · Minimum error entropy · Information gradient · Multi-error · Forgetting factor

Symbols and abbreviations

$u(k)$	System input at time k
$y(k)$	System output at time k
$y^*(k)$	System output without noise at time k
$n(k)$	System noise at time k
θ	Parameter vector
θ_0	True value of parameter vector
$\hat{\theta}(k)$	Parameter estimate at time k

This work is supported by the Natural Science Research Project of Jiangsu Higher School, China (19KJD510001), and the National Scholarship Foundation of China (201908320048).

✉ Shaoxue Jing
jingsx3@126.com

[1] School of Physics and Electronic Electrical Engineering, Huaiyin Normal University, Huaian 223300, Jiangsu, China

Birkhäuser

$\tilde{\theta}(k)$	$\tilde{\theta}(k) = \hat{\theta}(k) - \theta_0$
n	Dimension of parameter vector
N	Data length
L	Parzen window length
σ	Kernel width of Gaussian kernel
e	Error variable
$e(k)$	Error at time k
pdf	Probability density function
$f(e)$	Pdf of error
$\hat{f}(e(k))$	Estimate of $f(e)$ at time k
$H(e)$	Shannon entropy for e
$\hat{H}(e(k))$	Estimate of $H(e)$ at time k
$E(\cdot)$	Mathematical expectation
Δ_{ki}	$\Delta_{ki} = e(k) - e(i)$
ϵ_{ki}	$\epsilon_{ki} = \varphi(k) - \varphi(i)$
$\mathbf{g}(k)$	Stochastic gradient of Shannon entropy
$\kappa_\sigma(\cdot)$	Gaussian kernel with variance σ^2
$\kappa'_\sigma(\cdot)$	Derivative of $\kappa_\sigma(\cdot)$
SG	Stochastic gradient
SIG	Stochastic information gradient
$\eta(k)$	Step size of SIG
ME	Multi-error
p	Stack length of ME
$\mathbf{E}(p, k)$	Stacked error vector
$\mathbf{\Phi}(p, k)$	Stacked information matrix
$\Delta\mathbf{\Phi}(p; k, i)$	$\Delta\mathbf{\Phi}(p; k, i)$=$\mathbf{\Phi}(p, k) - \mathbf{\Phi}(p, i)$
$\Delta E(p; k, i)$	$\Delta E(p; k, i)$=$E(p, k) - E(p, i)$
FF	Forgetting factor
λ	Forgetting factor
H_1	$H_1 = \frac{\partial \mathbf{g}^T}{\partial \theta}$
I	Identity matrix
γ	Eigenvalue

1 Introduction

ARX model is an AutoRegressive model with eXogenous terms [31]. Because of its simplicity and easy parameterization, the ARX model has been widely used to model a lot of real systems, such as micro-turbines, data improving, fault detection, biomedical signals, COVID-19 case forecasting and communication systems [1,3,7,28,34,44].

Much research has been performed to identify ARX models in the last five decades. A piecewise auto-regressive exogenous structure was adopted to forecast the river floods [18]. A novel automated framework based on generalized spectral decomposition was proposed to estimate the parameters of an ARX model [33]. A new method based on the expectation–maximization (EM) algorithm was utilized for the identification of ARX models subject to missing data [24]. A recursive EM algorithm based on Student's t-distribution was used for robust identification of ARX models [8]. A modified momentum gradient descent algorithm was investigated to identify ARX models [50]. A three-stage algorithm was studied for the identification of fractional differencing ARX with errors in variables [25].

However, most of the noises considered in the aforementioned papers are white noises or Gaussian noises. Random impulse signals can often be found in industrial signals, such as image signals, audio signals and communication signals [2,9,46].

Identification criteria play an important role in system identification. The classical identification criteria include the least square criterion, maximum likelihood criterion and so on. These criteria have the advantages of simple calculation and easy theoretical analysis. However, the performance of the least-squares algorithm is poor for the non-Gaussian case, and the maximum likelihood algorithm needs to know the conditional probability density of the sample. Because of these problems, many researchers have put forward many other criteria, such as $p-$norm error criterion and mixed-norm error criterion [39,53]. In recent years, information criteria have become more widespread in signal processing and system identification [14,23,32]. Compared with mean square error (MSE) criterion, which focuses on second-order statistics, the information-theoretic criterion (e.g., minimum error entropy (MEE) [4], Renyi's entropy [15,41], fixed-point maximum correntropy [21]) is related to various statistical behaviors of the probability density function (pdf) of the error. Algorithms based on information-theoretic criterion may have better performance than MSE-based algorithms [6,15,16].

In the last decade, entropy has found significant applications in system identification. A maximum correntropy criterion (MCC) algorithm was proposed for sparse system identification based on normalized adaptive filter theory [30]. An extended version of correntropy, whose center can be located at any position, and a new optimization criterion based on MCC with a variable center, were proposed [5]. A blocked proportionate normalized maximum correntropy algorithm and a separable maximum correntropy adaptive algorithm were presented to identify dynamic systems [29,45].

To decrease the entropy estimators' complexity, a stochastic information gradient (SIG) algorithm was proposed and its performance was investigated [14]. To improve the estimates, a joint stochastic gradient algorithm based on MSE and MEE was proposed and applied to identify an RBF network [4]. Though having less complexity, the SIG converges very slowly. To speed up the SIG, like the multi-innovation used in [12], a multi-error strategy is adopted and a feasible equation for calculation of the step size is introduced.

Since its introduction in 2003, the SIG algorithm has been widely used in system identification and machine learning. For example, a kernel-based gradient descent algorithms based on MEE was proposed to find nonlinear structures in the data, and its convergence rate was deduced [22]. A kernel adaptive filter for quaternion data was developed, and a new algorithm based on the SIG approach was applied to this filter [37]. To avoid unstable training or poor performance in deep learning, a strategy of directly estimating the gradients of information measures with respect to model parameters was explored, and a general gradient estimation method for information measures was proposed [52]. To avoid potentially sub-optimal solutions with respect to class separability, a dimensionality reduction network training procedure based on the stochastic estimate of the mutual information gradient was presented [38].

For several decades, data-driven techniques have been used in modeling and fault detection. For example, a surrogate model was developed based on a data-driven approach, and it can facilitate the design and optimization processes of permanent magnet systems [36]. A Matlab toolbox for data-based fault detection was developed in a unified data-driven framework. [27]. A new recursive total principle component regression-based design and implementation approach was proposed for efficient data-driven fault detection and applied to vehicular cyber-physical systems [26].

In this paper, the problem of parameter identification of the ARX model disturbed by random impulse noise is studied. The premise is that the structure of the model is known, the type of noise is known, the identification data are normal measurement data, and there are no outliers except impulse noise. The possible outliers, modeling errors, and other uncertainties in practice are not considered. Interested readers can refer to the recent literature [13,47–49]. The main contributions of this work are as follows:

(1) For the SIG algorithm, a simple step size method is proposed.
(2) To make the algorithm faster, a multi-error method that uses stack error instead of instantaneous scalar error is applied.
(3) Since the stack length can only be taken as an integer, a forgetting factor is used to further accelerate the algorithm.
(4) The proposed algorithm is utilized to estimate the parameters of an ARX model with random impulse noise. Several numerical simulations and a case study show the effectiveness of the algorithm.

The rest of this work is organized as follows. In the next section, we describe the ARX model to be estimated. Based on an SIG algorithm in Sect. 3, a multi-error SIG with a forgetting factor is presented in Sect. 4. The convergence and the computational cost are analyzed in Sect. 5. Then, parameter estimation of an ARX model with a random impulse noise and a gas furnace dataset from the literature [42] is used to validate the proposed algorithm in Sect. 6. Finally, conclusions are presented in Sect. 7.

2 Problem Description

Consider an ARX model depicted in Fig. 1, where $u(k)$ is the input and $y(k)$ is the output. $A(z^{-1})$ and $B(z^{-1})$ are two polynomials with respect to z^{-1}, and their degrees are n_a and n_b, respectively. The model is polluted by a random impulse noise $v(k)$.

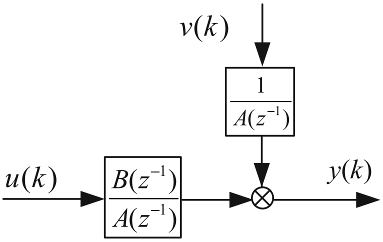


Fig. 1 Block diagram of an ARX model

It is easy to find that

$$y(k) = \frac{B(z^{-1})}{A(z^{-1})}u(k) + \frac{1}{A(z^{-1})}v(k). \tag{1}$$

Multiplying both sides of Eq. 1 by $A(z^{-1})$ gives

$$A(z^{-1})y(k) = B(z^{-1})u(k) + v(k). \tag{2}$$

Suppose $A(z^{-1}) = 1 + a_1z^{-1} + a_2z^{-2} + \cdots + a_{n_a}z^{-n_a}$ and $B(z^{-1}) = b_1z^{-1} + b_2z^{-2} + \cdots + b_{n_b}z^{-n_b}$, then we can directly parameterize the model as follows,

$$\begin{aligned} y(k) &= \left(1 - A(z^{-1})\right) y(k) + B(z^{-1})u(k) + v(k), \\ &= \varphi^T(k)\theta + v(k), \end{aligned} \tag{3}$$

with

$$\begin{cases} \theta = \left[a_1, \cdots, a_{n_a}, b_1, \cdots, b_{n_b}\right]^T \in \mathbb{R}^{n\times 1}, \\ \varphi(k) = [-y(k-1), \cdots, -y(k-n_a), u(k-1), \cdots, u(k-n_b)]^T \in \mathbb{R}^{n\times 1}, \\ n = n_a + n_b. \end{cases} \tag{4}$$

Then, the identification of the ARX model shown in Fig. 1 can be transformed into the estimation of the parameters θ based on the observations $\{u(k), y(k)\}_{k=1}^{N}$, where N is the data length.

However, traditional identification algorithms, such as the least-square algorithm and the mean square error algorithm, only consider the second moment of the error, and in some cases (such as the presence of random impulse noise) identification results deteriorate further. The information criterion algorithm based on a probability density function (pdf) considers the statistical information of each order of the error and is expected to achieve better estimates.

Next, we introduce the SIG algorithm and then describe our algorithm based on information gradient, which is integrated with the multi-error and a forgetting factor.

3 SIG of Shannon's Error Entropy

Consider the parameterized system in Eq. 3, and denote a random error $e(k)$ as

$$e(k) = y^*(k) - \varphi^T(k)\theta, \tag{5}$$

where $y^*(k)$ is the system output without noise.

Shannon's entropy for e with pdf $f(e)$ is [43]

$$H(e) = -\int_{-\infty}^{\infty} f(e)\log f(e)de = E\left[-\log f(e)\right]. \tag{6}$$

In practice, the pdf of e, i.e., $f(e)$, is unknown. Thus, Eq. 6 cannot be used to calculate the entropy of e directly. One way is to utilize a Parzen window to approximate the unknown pdf underlying the N observations by [41]

$$\hat{f}(e) = \frac{1}{N}\sum_{i=1}^{N}\kappa_\sigma\left(e - e(i)\right), \tag{7}$$

where $\kappa_\sigma(\cdot)$ is the kernel function with size σ [40].

At time k, a Parzen window estimate of e with window length L is

$$\hat{f}\left(e(k)\right) = \frac{1}{L}\sum_{i=k-L}^{k-1}\kappa_\sigma\left(\varDelta_{ki}\right), \tag{8}$$

where $\varDelta_{ki}$=$e(k) - e(i)$, and $e(k)$, $e(i)$ denote the error at time k, i, respectively.

Thus, the stochastic entropy estimate at time k becomes

$$\hat{H}(e(k)) = E\left[-\log\left(\frac{1}{L}\sum_{i=k-L}^{k-1}\kappa_\sigma\left(\varDelta_{ki}\right)\right)\right]. \tag{9}$$

Dropping the expectation in Eq. 9 [14], we obtain

$$\hat{H}(e(k)) \approx -\log\left(\frac{1}{L}\sum_{i=k-L}^{k-1}\kappa_\sigma\left(\varDelta_{ki}\right)\right). \tag{10}$$

The stochastic gradient of Shannon entropy concerning θ at time k, $\mathbf{g}$, is

$$\mathbf{g} = -\frac{\sum\limits_{i=k-L}^{k-1}\kappa'_\sigma\left(\varDelta_{ki}\right)\left(\frac{\partial e(k)}{\partial\theta} - \frac{\partial e(i)}{\partial\theta}\right)}{\sum\limits_{i=k-L}^{k-1}\kappa_\sigma\left(\varDelta_{ki}\right)}, \tag{11}$$

where $\kappa'_\sigma(\cdot)$ is the derivative of the kernel function.

Using the following Gaussian kernel with variance σ^2, i.e.,

$$\kappa_\sigma\left(\varDelta_{ki}\right) = \exp\left(-\frac{\|\varDelta_{ki}\|^2}{2\sigma^2}\right), \tag{12}$$

Equation 11 becomes

$$\mathbf{g} = \frac{\sum\limits_{i=k-L}^{k-1} \kappa_\sigma\left(\varDelta_{ki}\right) \epsilon_{ki} \varDelta_{ki}}{\sigma^2 \sum\limits_{i=k-L}^{k-1} \kappa_\sigma\left(\varDelta_{ki}\right)}, \tag{13}$$

with

$$\epsilon_{ki} = \varphi(k) - \varphi(i). \tag{14}$$

The SIG for estimating the parameter vector θ is obtained as follows:

$$\hat{\theta}(k) = \hat{\theta}(k-1) + \eta(k)\mathbf{g}, \tag{15}$$

where $\eta(k)$ is the step size and is critical for convergence speed. However, equations to calculate the step size in [4] and [14] are too complicated to operate online. Here, we utilize the equation in stochastic gradient [10]:

$$\begin{cases} r(k) = r(k-1) + \|\varphi(k)\|^2, \quad r(0) = 1, \\ \eta(k) = \dfrac{1}{r(k)}. \end{cases} \tag{16}$$

In practice, the θ and the output without noise $y^*(k)$ in Eq. 5 are unknown. A feasible way is to replace them by $\hat{\theta}(k-1)$ and $y(k)$, respectively. Thus, Eq. 5 becomes

$$e(k) = y(k) - \varphi^T(k)\hat{\theta}(k-1). \tag{17}$$

4 Forgetting Factor Multi-error SIG Algorithm

One drawback of the SIG algorithm is its slow convergence. To enable the algorithm converge faster, a multi-error strategy is adopted and Eq. 13 is rewritten as follows:

$$\mathbf{g} = \frac{\sum\limits_{i=k-L}^{k-1} \kappa_\sigma\left(\varDelta_{ki}\right) \varDelta\boldsymbol{\Phi}(p;k,i)\varDelta\mathbf{E}(p;k,i)}{\sigma^2 \sum\limits_{i=k-L}^{k-1} \kappa_\sigma\left(\varDelta_{ki}\right)} \tag{18}$$

with

$$\begin{cases} \varDelta\mathbf{\Phi}(p;k,i)=\mathbf{\Phi}(p,k)-\mathbf{\Phi}(p,i), \\ \varDelta\mathbf{E}(p;k,i)=\mathbf{E}(p,k)-\mathbf{E}(p,i), \end{cases} \tag{19}$$

where p is the stack length and $\mathbf{E}(p,k)$ and $\mathbf{\Phi}(p,k)$ are the stacked error vector and stacked information matrix, respectively,

$$\mathbf{E}(p,k)=\begin{bmatrix} e(k) \\ e(k-1) \\ \vdots \\ e(k-p+1) \end{bmatrix} \in \mathbb{R}^{p\times 1}, \tag{20}$$

and

$$\mathbf{\Phi}(p,k)=[\varphi(k),\varphi(k-1),\cdots,\varphi(k-p+1)] \in \mathbb{R}^{n\times p}. \tag{21}$$

Note that the scalar error $\varDelta_{ki}$ in Eq. 13 is replaced by the vector error $\varDelta\mathbf{E}(k,i)$ in Eq. 18. In other words, multi-error takes the place of a single error. Thus, the algorithm is named a multi-error SIG algorithm (ME-SIG).

The stack length p can only be a positive integer. To make the ME-SIG faster, a forgetting factor (FF) λ is introduced. The first equation of Eq. 16 becomes

$$r(k)=\lambda r(k-1)+\|\varphi(k)\|^2, \quad r(0)=1. \tag{22}$$

Equations 15 and 17–22 construct the FF-ME-SIG algorithm.

5 Performance Analysis

5.1 Convergence Analysis

The approximate linearization approach [17] is used to analyze the convergence of the proposed ME-SIG algorithm in Eq. (15) with Eqs. (18) and (19).

Subtracting θ_0 from both sides of Eq. (15), we obtain

$$\tilde{\theta}(k)=\tilde{\theta}(k-1)+\eta(k)\mathbf{g}, \tag{23}$$

where $\tilde{\theta}(k)=\hat{\theta}(k)-\theta_0$ is the estimation error vector of the parameter.

Approximating the gradient $\mathbf{g}$ in Eq. (18) using the first-order Taylor expansion:

$$\begin{aligned} \mathbf{g} &\approx \mathbf{g}(\theta_0)+H_1\left(\hat{\theta}(k-1)-\theta_0\right) \\ &= H_1\left(\hat{\theta}(k-1)-\theta_0\right) \\ &= H_1\tilde{\theta}(k-1), \end{aligned} \tag{24}$$

where $H_1 = \frac{\partial \mathbf{g}^T(\theta_0)}{\partial \theta}$ is the Hessian matrix and is expressed as

$$\begin{aligned} H_1 =& \frac{\sum\limits_{i=k-L}^{k-1} \kappa_\sigma\left(\Delta_{ki}\right)\left(\Delta \mathbf{E}^T(p;k,i)\right)' \Delta \mathbf{\Phi}^T(k,i)}{\sigma^2 \sum\limits_{i=k-L}^{k-1} \kappa_\sigma\left(\Delta_{ki}\right)} \\ &+ \frac{\sum\limits_{i=k-L}^{k-1} \kappa_\sigma\left(\Delta_{ki}\right) \Delta \mathbf{E}^T(p;k,i)\left(\Delta \mathbf{\Phi}^T(k,i)\right)'}{\sigma^2 \sum\limits_{i=k-L}^{k-1} \kappa_\sigma\left(\Delta_{ki}\right)} \\ &- \frac{\sum\limits_{i=k-L}^{k-1} \kappa_\sigma\left(\Delta_{ki}\right) \Delta \mathbf{E}^T(p;k,i) \Delta \mathbf{\Phi}^T(k,i) \sum\limits_{i=k-L}^{k-1} \kappa'_\sigma\left(\Delta_{ki}\right)}{\sigma^2 \sum\limits_{i=k-L}^{k-1} \kappa^2_\sigma\left(\Delta_{ki}\right)}. \end{aligned} \tag{25}$$

Substituting Eq. (24) into Eq. (23), we obtains

$$\tilde{\theta}(k) = \tilde{\theta}(k-1) + \eta(k) H_1 \tilde{\theta}(k-1), \tag{26}$$

We analyze the convergence of Eq. 26 by borrowing the results from the LMS convergence theory [19,20]. Assuming that the Hessian matrix H_1 is a normal matrix and can be decomposed into the following normal form:

$$H_1 = Q_1 \Lambda_1 Q_1^{-1}, \tag{27}$$

where Q_1 is an $m \times m$ orthogonal matrix, $\Lambda_1 = diag[\gamma_1, \gamma_2, \cdots, \gamma_m]$, γ_i is the eigenvalue of H_1. Then, the recursive Eq. 26 can be expressed as

$$\begin{aligned} \tilde{\theta}(k) &= Q_1 \left[I + \eta(k) \Lambda_1\right] Q_1^{-1} \tilde{\theta}(k-1) \\ &= Q_1 \left[\prod_{i=1}^{k} \left(I + \eta(i) \Lambda_1\right)\right] Q_1^{-1} \tilde{\theta}(0). \end{aligned} \tag{28}$$

Clearly, if the following conditions are satisfied, $\tilde{\theta}(k) \to \mathbf{0}$, i.e., $\hat{\theta}(k) \to \theta_0$:

$$\left|1 + \eta(i)\gamma_j\right| < 1, \quad i = 1, 2, \cdots, k, \quad j = 1, 2, \ldots, m. \tag{29}$$

Thus, a sufficient condition that ensures the convergence of the algorithm is as follows:

$$\begin{cases} \gamma_j < 0, \, j = 1, 2, \cdots, m, \\ 0 < \eta(i) < \dfrac{2}{\max\limits_{j} |\gamma_j|}. \end{cases} \tag{30}$$

Table 1 Computational cost of SIG, ME-SIG and FF-ME-SIG

Algorithm	Computation cost (flops)	$n = 4, L = 3, p = 5$	Time (s)
SIG	$3nL + 17L + n + 2$	93	0.1788
ME-SIG	$npL + pL + 16L + 2$	125	0.1860
FF-ME-SIG	$npL + pL + 16L + 3$	126	0.1862

5.2 Computational Analysis

According to the calculation method of [11], the calculation amount of each iteration of the three algorithms is shown in Table 1, where e^x is calculated by its Taylor expansion, and the first three terms are used. 'Time' denotes the time that the numerical example consumes.

From the calculation of complexity and running time, we can see that the former algorithm has lower values than the others, while the latter two algorithms have little difference. The computational complexity of the two algorithms proposed in this paper is larger than that of the first algorithm, because the latter two algorithms need the calculation of multi-error. In terms of running time, the latter two algorithms take approximately 4% more time than the previous SIG algorithm, which is not a significant difference.

6 Experimental Results

Consider the ARX model depicted in Fig. 1 with

$$\begin{cases} A(z^{-1}) = 1.0 - 1.5z^{-1} + 0.7z^{-2}, \\ B(z^{-1}) = 1.0z^{-1} + 0.5z^{-2}, \end{cases} \tag{31}$$

where input data $u(k)$ are an M-sequence and $v(k)$ is a random impulse noise. 5% of the output data (30 outputs) are randomly selected, and 30 noises with random amplitude between 0 and 1 are added, respectively. The curves of input $u(k)$ and output $y(k)$ are shown in Fig. 2. All simulation experiments use this model.

6.1 Numerical Simulation

(1) Results using SIG algorithm

The parameter estimates using the SIG with window length $L = 3$ are shown in Table 2, where the estimation error δ is defined as $\delta = \frac{\left\|\hat{\theta}(k) - \theta_0\right\|}{\|\theta_0\|} \times 100$. It is easy to determine that estimation error decreases as data length k increases (for a given L). However, the errors are very large (38.9319%) at the end of the estimation.

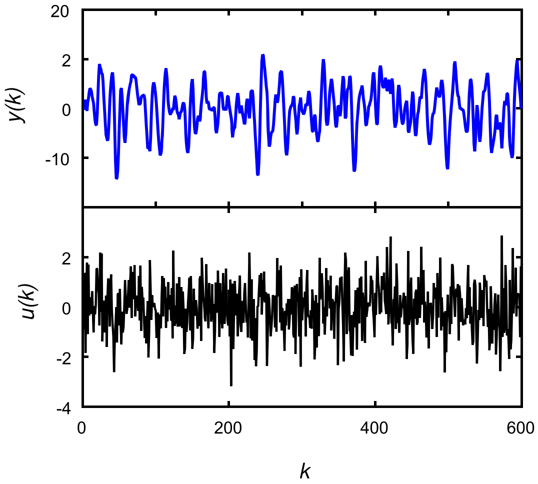


Fig. 2 Curves of input–output data

Table 2 Results using SIG algorithm ($L = 5$)

k	25	50	100	200	400	600	True value
a_1	− 0.8723	− 0.8514	− 0.8443	− 0.8453	− 0.8461	− 0.8436	− 1.5000
a_2	− 0.1863	− 0.0802	− 0.0595	− 0.0454	− 0.0264	− 0.0135	0.7000
b_1	0.1902	0.2164	0.2192	0.2238	0.2288	0.2321	1.0000
b_2	− 0.0186	0.0397	0.0456	0.0526	0.0612	0.0666	0.5000
$\delta(\%)$	69.6451	68.1897	67.5927	66.9201	66.0873	65.6092	

Table 3 Results using ME-SIG algorithm ($p = 5$, $L = 3$)

k	25	50	100	200	400	600	True value
a_1	− 1.3218	− 1.3024	− 1.3212	− 1.3428	− 1.3725	− 1.3874	− 1.5000
a_2	− 0.2220	0.4978	0.5452	0.5706	0.5970	0.6133	0.7000
b_1	0.3139	0.8141	0.8219	0.8260	0.8375	0.8454	1.0000
b_2	− 0.0422	0.2836	0.3057	0.3321	0.3684	0.3854	0.5000
$\delta(\%)$	29.3258	20.1085	17.7285	15.8231	13.3009	11.9745	

(2) Results using ME-SIG algorithm

The parameter estimates using the proposed ME-SIG with stack length $p = 5$ and $L = 3$ are shown in Table 3. The estimation errors with different p are depicted in Fig. 3 ($L = 3$). It can be seen that:

(1) Estimation error of a given p decreases when data length k increases;
(2) With stack length p increasing, the estimation error decreases quickly.

(3) Results using FF-ME-SIG algorithm

The parameter estimates using proposed FF-ME-SIG with forgetting factor $\lambda = 0.99$ are shown in Table 4, where $L = 3$ and $p = 5$. It can be seen that:

Fig. 3 Estimation errors using ME-SIG with different stack length p

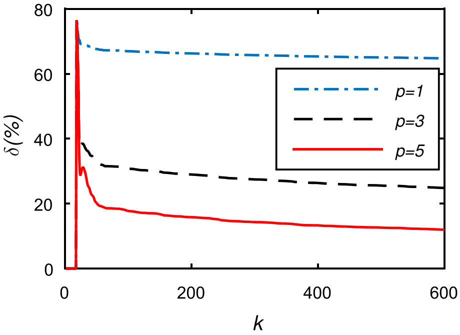
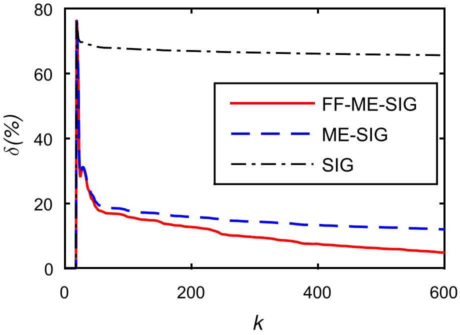


Table 4 Results using FF-ME-SIG algorithm ($p = 5$, $L = 3$, $\lambda = 0.99$)

k	25	50	100	200	400	600	True value
a_1	− 1.3236	− 1.3184	− 1.3425	− 1.3746	− 1.4372	− 1.4641	− 1.5000
a_2	− 0.2209	0.5130	0.5627	0.6075	0.6513	0.6741	0.7000
b_1	0.3152	0.8302	0.8386	0.8463	0.8864	0.9188	1.0000
b_2	− 0.0425	0.3015	0.3283	0.3724	0.4446	0.4723	0.5000
$\delta(\%)$	29.1555	18.4745	15.7657	12.6848	7.4742	4.8337	

Fig. 4 Estimation errors using SIG, ME-SIG, FF-ME-SIG



(1) Estimation error decreases when data length k increases;
(2) Compared with Table 3, the estimation error of the FF-ME-SIG is smaller at the same data length k.

(4) Comparison of the results of SIG, ME-SIG and FF-ME-SIG algorithms

The estimation errors using SIG, ME-SIG and FF-ME-SIG are depicted in Fig. 4. It can be seen that:

(1) All curves decrease when data length k increases;
(2) The estimation error of the SIG algorithm is larger than that of ME-SIG, which means that multi-error can improve the accuracy of the SIG's estimate;
(3) The estimate of the FF-ME-SIG is the most accurate one of the three. In other words, the introduction of the forgetting factor improves the accuracy of the estimation.

Table 5 Results using FF-ME-SIG algorithm under different noise additions ($p = 5, L = 3, \lambda = 0.99$)

Estimates	5%	10%	30%	50%	True value
a_1	− 1.4641	− 1.4322	− 1.4807	− 1.3607	1.5000
a_2	0.6741	0.6419	0.6910	0.5651	0.7000
b_1	0.9188	0.9084	0.9329	0.9424	1.0000
b_2	0.4723	0.5026	0.4043	0.4841	0.5000
$\delta(\%)$	4.8337	6.4037	5.9477	10.1580	

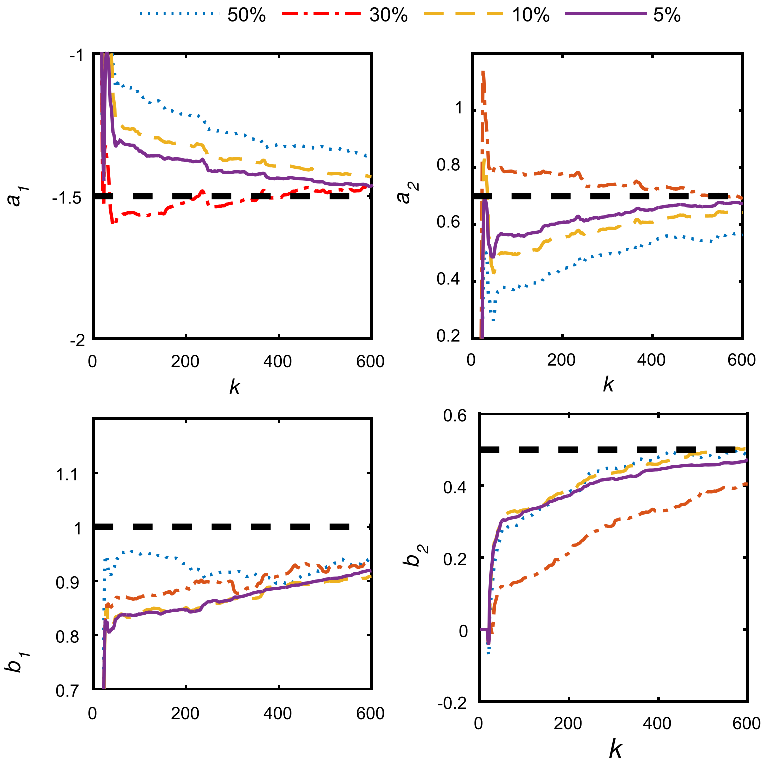
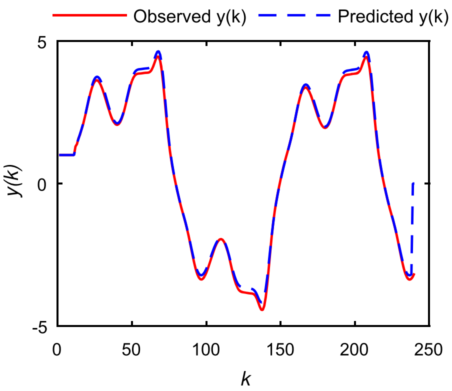


Fig. 5 Estimates of FF-ME-SIG using samples with different noise additions

(5) Results using FF-ME-SIG algorithm under different noise additions

To test the performance of the algorithm under different noise levels, we add 5%, 10%, 30%, and 50% of noises to the samples. The mean of the noise is 0.5, and the amplitude is between 0 and 1. The estimation results when $k = 600$ are shown in Table 5, where $p = 5, L = 3, \lambda = 0.99$. The estimation error curve is shown in Fig. 5. It can be seen from Table 5 and Fig. 5 that with the increase in the added noise, the estimation error tends to increase, but the change is small, which indicates that the proposed algorithm has strong adaptability to noise.

Fig. 6 Predicted and observed output using FF-ME-SIG for Eqs. 32–33



(6) Results using the FF-ME-SIG for the identification of the Narendra difference equations

To support this paper's argument, we add further simulation examples involving synthetic input–output relations such as the Narendra difference equations proposed in [35,51]:

$$\boldsymbol{y}(n+1) = 0.3\boldsymbol{y}(n) + 0.6\boldsymbol{y}(n-1) + f[\boldsymbol{e}(n)], \tag{32}$$

where

$$\begin{cases} f(e) = 0.6\sin(\pi e) + 0.3\sin(3\pi e) + 0.1\sin(5\pi \boldsymbol{e}) \\ \boldsymbol{e}(n) = \sin\left[(1+\boldsymbol{a})\omega_0 n\right]. \end{cases} \tag{33}$$

Using the following structure to the model above equation:

$$y(k) = a_1 y(k-1) + a_2 y(k-2) + a_3 y(k-3) + a_4 y^2(k-1) + a_5 y^2(k-2) \tag{34}$$

Let the data length be 240. The estimate using the proposed algorithm is $[0.3084, 0.3255, 0.3641, -0.0264, 0.0372]$ (when $k = 240$). The predicted $y(k)$ and observed $y(k)$ are depicted in Fig. 6.

(7) Results using FF-ME-SIG algorithm and RLS, SG algorithm

To prove the superiority of the proposed algorithm, the identification results of the stochastic gradient (SG) algorithm and recursive least squares (RLS) algorithm are compared with that of the proposed algorithm. Figure 7 shows the estimation error curves of the three algorithms. When $k = 600$, the estimation error of the SG, RLS and FF-ME-SIG is 55.5395%, 5.2045% and 4.8337%, respectively. It can be seen that the estimation error of SG is very large, and the estimation error of the RLS algorithm is slightly larger than that of the proposed algorithm. However, due to the impulse noise, the estimation error of the RLS algorithm changes dramatically, which indicates that the estimates given by the RLS algorithm change significantly.

Fig. 7 Estimate errors using SG, RLS and FF-ME-SIG

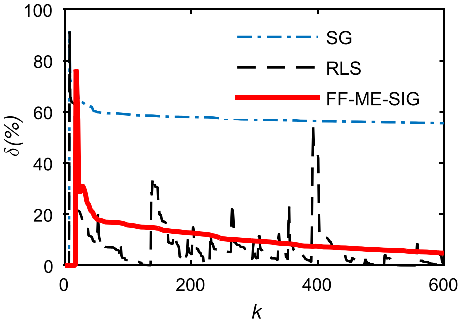


Fig. 8 Curves of the input and output data of a gas furnace

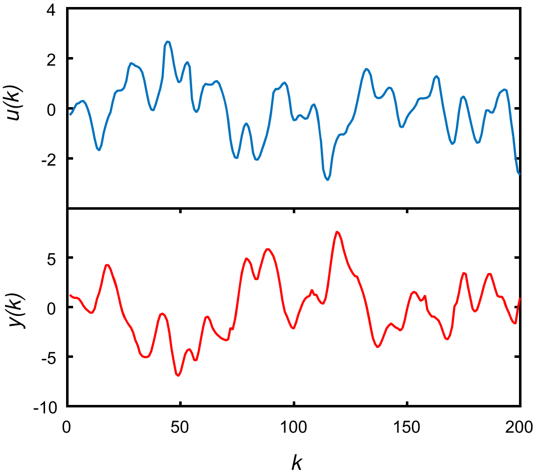


Table 6 Results using SIG, ME-SIG and FF-ME-SIG algorithm

k	SIG	ME-SIG	FF-ME-SIG
a_1	− 0.4295	− 0.6126	− 0.6215
a_2	− 0.3003	− 0.0561	0.0029
b_1	− 0.0868	− 0.0904	0.0614
b_2	− 0.1350	− 0.2137	− 0.1414
b_3	− 0.1891	− 0.4097	− 0.4190
b_4	− 0.1994	− 0.4148	− 0.4577
b_5	− 0.1686	− 0.2513	− 0.2869
RMSE	0.2782	0.1183	0.0594

6.2 Case Study

The data set of a gas furnace from the literature [42] is used to validate the proposed algorithm. These data were continuously collected from a gas furnace and then read every 9 s. The air feed of the furnace was kept constant, but the methane feed rate was varied and the resulting CO_2 concentration in the off gases was measured. There are 296 input–output data in this set. The first 200 data are adopted to estimate the parameters.

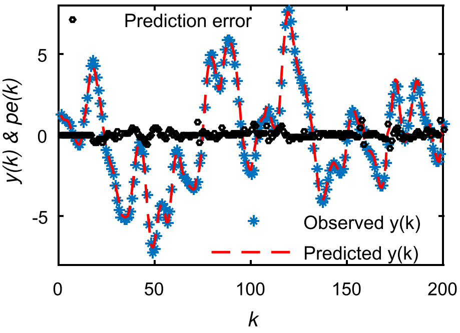


Fig. 9 Curves of the outputs and prediction errors of a gas furnace

The curves of the input and output are shown in Fig. 8. The estimation results and the prediction errors using SIG, ME-SIG and FF-ME-SIG are listed in Table 6. The outputs $y(k)$ and the prediction errors $pe(k)$ using the proposed algorithm are shown in Fig. 9.

(1) It can be seen from Table 6 that among three algorithms, the proposed FF-ME-SIG algorithm has the smallest RMSE, which means that the proposed algorithm can give the most accurate estimate.
(2) As shown in Fig. 9, the outputs of the obtained model using the proposed algorithm can predict the observations well.

7 Conclusions

In this paper, a novel SIG algorithm based on minimum error entropy is presented. The traditional SIG algorithm needs less computation than the MSE algorithm. However, it converges slowly. A multi-error strategy and a forgetting factor are introduced to speed up the SIG. We compared the results of SIG, ME-SIG and FF-ME-SIG by estimating the parameters of an ARX model with random impulse noise and through a case study. It is found that SIG with ME and FF can obtain accurate estimates, and it has a quick convergence rate.

Data Availability Statement All data generated or analyzed during this study are included in the attached file “data used in the paper.xlsx”.

References

1. H.N. Akouemo, R.J. Povinelli, Data improving in time series using ARX and ANN models. IEEE Trans. Power Syst. **32**(5), 3352–3359 (2017)
2. A. Awad, Impulse noise reduction in audio signal through multi-stage technique. Eng. Sci. Technol. Int. J. **22**(2), 629–636 (2018)
3. C. Böck, K. Kostoglou, P. Kovacs, M. Huemer, J. Meier, A linear parameter varying ARX model for describing biomedical signal couplings, in *Computer Aided Systems Theory-EUROCAST 2019. 17th International Conference*. (Las Palmas de Gran Canaria, Spain, 2020), pp. 339–346

4. B. Chen, J. Hu, H. Li, Z. Sun. A joint stochastic gradient algorithm and its application to system identification with RBF networks, in *Proceedings of the 6th World Congress on Intelligent Control and Automation* (Dalian, China, 2006), pp. 1754–1758
5. B. Chen, X. Wang, Y. Li, J.C. Principe, Maximum correntropy criterion with variable center. IEEE Signal Process. Lett. **26**(8), 1212–1216 (2019)
6. B. Chen, Y. Zhu, J. Hu, J.C. Principe, *System Parameter Identification: Information Criteria and Algorithms* (Elsevier, New York, 2013)
7. J. Chen, Y. Liu, F. Ding, Q. Zhu, Gradient-based particle filter algorithm for an ARX model with nonlinear communication output. IEEE Trans. Syst. Man Cybern. Syst. **50**(6), 2198–2207 (2020)
8. X. Chen, S. Zhao, F. Liu, Robust identification of linear ARX models with recursive EM algorithm based on student's t-distribution. J. Frankl. Inst. **358**(1), 1103–1121 (2021)
9. H. Dawood, H. Dawood, P. Guo, Removal of random-valued impulse noise by local statistics. Multimed. Tools Appl. **74**(24), 11485–11498 (2015)
10. F. Ding, System identification. Part F: multi-innovation identification theory and methods. J. Nanjing Univ. Inf. Sci. Technol. **4**(1), 1–28 (2012)
11. F. Ding, *New Theory of System Identification* (Tsinghua University Press, Beijing, 2013)
12. F. Ding, T. Chen, Performance analysis of multi-innovation gradient type identification methods. Automatica **43**(1), 1–14 (2007)
13. X. Dong, S. He, V. Stojanovic, Robust fault detection filter design for a class of discrete-time conic-type nonlinear Markov jump systems with jump fault signals. IET Control Theory Appl. **14**(14), 1912–1919 (2020)
14. D. Erdogmus, K.E. Hild, J.C. Principe, Online entropy manipulation: stochastic information gradient. Signal Process. Lett. **10**(8), 242–245 (2003)
15. D. Erdogmus, J.C. Principe, An error-entropy minimization algorithm for supervised training of nonlinear adaptive systems. IEEE Trans. Signal Process. **50**(7), 1780–1786 (2002)
16. D. Erdogmus, J.C. Principe, Generalized information potential criterion for adaptive system training. IEEE Trans. Neural Netw. **13**(5), 35–44 (2002)
17. D. Erdogmus, J.C. Principe, Convergence properties and data efficiency of the minimum error entropy criterion in Adaline training. IEEE Trans. Signal Process. **51**(7), 1966–1978 (2003)
18. B. Hadid, E. Duviella, S. Lecoeuche, Data-driven modeling for river flood forecasting based on a piecewise linear ARX system identification. J. Process Control **86**, 44–56 (2020)
19. S. Haykin, *Least-Mean-Square Adaptive Filters* (Wiley, New York, 2003)
20. S. Haykin, *Adaptive Filter Theory* (Pearson Education Limited, England, 2014)
21. A.R. Heravi, G.A. Hodtani, Comparison of the convergence rates of the new correntropy-based Levenberg–Marquardt (CLM) method and the fixed-point maximum correntropy (FP-MCC) algorithm. Circuits Syst. Signal Process. **37**(7), 2884–2910 (2018)
22. T. Hu, Q. Wu, D. Zhou, Kernel gradient descent algorithm for information theoretic learning. J. Approx. Theory **263**, 105518 (2021)
23. A. Hyvarinen, E. Oja, Independent component analysis: algorithms and applications. Neural Netw. **13**(4), 411–430 (2000)
24. A.J. Isaksson, Identification of ARX-models subject to missing data. IEEE Trans. Autom. Control **38**(5), 813–819 (2002)
25. D.V. Ivanov, I.L. Sandler, O.A. Katsyuba, V.N. Vlasova, Identification of FARARX models with errors in variables, in *Recent Trends in Intelligent Computing, Communication and Devices. Advances in Intelligent Systems and Computing*, vol. 1006, ed. by V. Jain, S. Patnaik, V.F. Popentiu, I. Sethi (Springer, Singapore, 2020), pp. 481–487
26. Y. Jiang, S. Yin, Recursive total principle component regression based fault detection and its application to vehicular cyber-physical systems. IEEE Trans. Industr. Inf. **14**(4), 1415–1423 (2017)
27. Y. Jiang, S. Yin, Recent advances in key-performance-indicator oriented prognosis and diagnosis with a Matlab toolbox: Db-kit. IEEE Trans. Industr. Inf. **15**(5), 2849–2858 (2018)
28. F. Jurado, A. Cano, Use of ARX algorithms for modelling micro-turbines on the distribution feeder. IEE Proc. Gener. Trans. Distrib. **151**(2), 232–238 (2004)
29. Y. Li, Z. Jiang, W. Shi, X. Han, B. Chen, Blocked maximum correntropy criterion algorithm for cluster-sparse system identifications. IEEE Trans. Circuits Syst. II Express Briefs **66**(11), 1915–1919 (2019)
30. Y. Li, Y. Wang, R. Yang, F. Albu, A soft parameter function penalized normalized maximum correntropy criterion algorithm for sparse system identification. Entropy **19**(1), 1–16 (2017)

31. L. Ljung, *System Identification: Theory for the User* (Tsinghua University Press, Beijing, 2002)
32. W. Magdy, T. Elsayed, Unsupervised adaptive microblog filtering for broad dynamic topics. Inf. Process. Manage. **52**(4), 513–528 (2016)
33. D. Maurya, A. Tangirala, S. Narasimhan. ARX model identification using generalized spectral decomposition. eprint arXiv:2008.04779 (2020)
34. T. Najeh, C.B. Njima, T. Garna, J. Ragot, Input fault detection and estimation using pi observer based on the ARX-Laguerre model. Int. J. Adv. Manuf. Technol. **90**(5), 1317–1336 (2017)
35. K.S. Narendra, K. Parthasarathy, Identification and control of dynamical systems using neural networks. IEEE Trans. Neural Netw. **1**(1), 4–27 (1990)
36. V.T. Nguyen, M. Bermingham, M.S. Dargusch, Data-driven modelling of the interaction force between permanent magnets. J. Magn. Magn. Mater. **532**, 167869 (2021)
37. T. Ogunfunmi, C. Safarian, The quaternion stochastic information gradient algorithm for nonlinear adaptive systems. IEEE Trans. Signal Process. **67**(23), 5909–5921 (2019)
38. O. Özdenizci, D. Erdogmus, Stochastic mutual information gradient estimation for dimensionality reduction networks. Inf. Sci. **570**, 298–305 (2021)
39. E.V. Papoulis, T. Stathaki, A normalized robust mixed-norm adaptive algorithm for system identification. IEEE Signal Process. Lett. **11**(1), 56–59 (2004)
40. E. Parzen, On estimation of a probability density function and mode. Ann. Math. Stat. **33**(3), 1065–1076 (1962)
41. J.C. Principe, *Information Theoretic Learning: Renyis Entropy and Kernel Perspectives* (Springer, New York, 2010)
42. T. Söderström, P. Stoica, Instrumental variable methods for system identification. Lect. Notes Control Inf. Ences **21**(1), 1–9 (1983)
43. C.E. Shannon, A mathematical theory of communication. Bell Syst. Tech. J. **27**(3), 379–423 (1948)
44. R.R. Sharma, M. Kumar, S. Maheshwari, K.P. Ray, Evdhm-Arima-based time series forecasting model and its application for Covid-19 cases. IEEE Trans. Instrum. Meas. **70**, 1–10 (2021)
45. W. Shi, Y. Li, B. Chen, A separable maximum correntropy adaptive algorithm. IEEE Trans. Circuits Syst. II Express Briefs **67**(11), 2797–2801 (2020)
46. W. Shieh, I.B. Djordjevic, *OFDM for Optical Communications* (Elsevier, London, 2010)
47. V. Stojanovic, S. He, B. Zhang, State and parameter joint estimation of linear stochastic systems in presence of faults and non-gaussian noises. Int. J. Robust Nonlinear Control **30**(16), 6683–6700 (2020)
48. V. Stojanovic, D. Prsic, Robust identification for fault detection in the presence of non-gaussian noises: application to hydraulic servo drives. Nonlinear Dyn. **100**, 2299–2313 (2020)
49. H. Tao, P. Wang, Y. Chen, V. Stojanovic, H. Yang, An unsupervised fault diagnosis method for rolling bearing using STFT and generative neural networks. J. Franklin Inst. **357**(11), 7286–7307 (2020)
50. Q. Tu, Y. Rong, J. Chen, Parameter identification of ARX models based on modified momentum gradient descent algorithm. Complexity **2020**(3), 1–11 (2020)
51. C. Turchetti, G. Biagetti, F. Gianfelici, P. Crippa, Nonlinear system identification: an effective framework based on the Karhunen–Loeve transform. IEEE Trans. Signal Process. **57**(2), 536–550 (2009)
52. L. Wen, H. Bai, L. He, Y. Zhou, M. Zhou, Z. Xu, Gradient estimation of information measures in deep learning. Knowl. Based Syst. **224**, 107046 (2021)
53. H. Zayyani, Continuous mixed p-norm adaptive algorithm for system identification. IEEE Signal Process. Lett. **21**(9), 1108–1110 (2014)