Two iterative reweighted algorithms for systems contaminated by outliers

Jing Chen, Manfeng Hu, Yanjun Liu, Quanmin Zhu

***Abstract*—This study proposes two iterative reweighted algorithms for systems whose data are contaminated by outliers. For the negative effect caused by the outliers, traditional least squares and gradient descent algorithms cannot obtain unbiased estimates, while the variational Bayesian and expectation-maximization algorithms have the assumption that the prior knowledge of the outlier is available. To deal with these dilemmas, two iterative reweighted algorithms are developed. By assigning suitable weights for each data set, unbiased parameter estimates can be obtained. In addition, the weights of the corrupted data sets become smaller and smaller with the increased number of iterations, and then the contaminated data can be picked out from the data sets. The proposed algorithms do not require the prior knowledge of the outliers. Convergence analysis and numerical experiments show effectiveness of the iterative reweighted algorithms.**

***Index Terms*—System identification, iterative reweighted algorithm, outlier, weight matrix, least squares algorithm**

I. Introduction

With the development of big data and machine-learning technologies, system identification becomes more and more important in modern society [1]–[4]. For example, in engineering practices [5], chemical industries [6], and medical areas [7]. System identification is constituted of two parts: one is structure identification [8], [9], and the other is parameter estimation [10]–[12]. For a black box model, one should determine its structure first, and then identify the parameters. In this paper, we try to estimate the parameters with the assumption that the structure of the considered model is known a priori.

There exist plethora of parameter estimation algorithms in the past few decades, e.g., the least squares (LS) algorithm, the gradient descent (GD) algorithm, the expectation maximization (EM) algorithm, and the variational Bayesian (VB) algorithm. Among them, the LS and GD algorithms are most widely used [13]–[17]. The LS algorithm has faster convergence rates with the cost of heavy computational efforts, while the GD algorithm has less computational costs but slower convergence rates. Therefore, the LS algorithm is a better choice for systems with simple structures and low-order, and the GD algorithm performs better for complex nonlinear models and high-order models [18], [19]. For systems with good data, these two algorithms can achieve satisfactory outcomes. However, in application, some data are usually noisy or contaminated by outliers, e.g., transmission errors, process disturbances, and instrument degradation [20]–[22]. The LS and GD methods are typically sensitive to outliers, resulting models usually cause biased estimates and model-order mismatch. Therefore, developing some novel identification algorithms which can identify systems with outliers is essential to parameter estimation.

To diminish the effect caused by the outliers, the most widely used method is to assign different kinds of noises to describe the dynamics of the outliers. Since Gaussian distribution noise cannot be sufficient to describe the outliers, other probabilistic distributions, such as t-distribution and Laplace distribution are usually considered [23], [24]. For example, in [25], a VB approach combining a t-distribution noise is utilized to identify the models with outliers, while the statistical parameters of the noises and the model parameters are interactively updated. In [10], Liu et al integrates the EM algorithm with a t-distribution noise to remove the effect of outliers. In [26], an iterative maximum likelihood estimator combining an outlier-robust bipercentile estimator is proposed for systems with outliers, where the K-distribution is applied to model the outliers. These methods have an assumption that the distribution of the outlier should be known a priori; otherwise, they will be inefficient. In addition, one should study the statistical characteristics of the noise to diminish the negative effect caused by outliers [27], [28].

Inspired by the method in [29], [30], this paper identifies the systems with outliers in another way: design two iterative reweighted algorithms which can automatically pick out all the outliers from the data sets, and then can obtain unbiased parameter estimates based on the good data. The proposed algorithms are resistant to outliers and result in improved accuracy and reliability of process modeling and prediction. In summary, compared with the work in [10], [23]–[26], the advantages of the proposed algorithms are summarized as follows:

1. do not require any prior knowledge about the outliers, e.g., the prior distributions of the outliers;

2. do not need to study the statistical characteristics of the outliers, for example, the distributions of the outliers do not need to be updated in each iteration;

3. can automatically pick out the contaminated data based on the weight estimates.

The rest of this study is organized as follows. Section II introduces the systems with outliers and the traditional LS and GD algorithms. Section III develops two iterative reweighted algorithms. Section IV presents some properties of the iterative reweighted algorithms. Numerical experiments are provided in Section V. Section VI concludes this paper and points out future directions.

II. Systems with outliers and traditional algorithms

Some notations are defined first: $\|\mathbf{X}\| = \sqrt{\lambda_{\max}[\mathbf{X}\mathbf{X}^{\mathrm{T}}]}$ denotes the 2-norm of a matrix $\mathbf{X}$; $\lambda_{\max}[\cdot]$ means the maximum eigenvalue of a matrix; T indicates the matrix transpose.

A. Systems with outliers

Consider the following model

$$y(t) = \sum_{i=1}^{m} g_i \varphi_i(t) + v(t),\ t = t_{no_1}, \cdots, t_{no_q}, \tag{1}$$

J. Chen and M.F. Hu are with School of Science, Jiangnan University, Wuxi 214122, PR China (chenjing1981929@126.com, humanfeng@jiangnan.edu.cn)

Y.J. Liu is with Key Laboratory of Advanced Process Control for Light Industry (Ministry of Education), Jiangnan University, Wuxi 214122, PR China (yanjunliu_1983@126.com)

Q.M. Zhu is with Department of Engineering Design and Mathematics, University of the West of England, Bristol BS16 1QY, UK (quan.zhu@uwe.ac.uk)

This work is supported by the National Natural Science Foundation of China (No. 61973137) and the Natural Science Foundation of Jiangsu Province (No. BK20201339)

and

$$y(t)=\sum_{i=1}^{m}g_i\varphi_i(t)+\varpi(t),\ t=t_{ab_1},\cdots,t_{ab_s}, \tag{2}$$

where $y(t)$ is the output, $\varphi_i(t)$ is the information element constituted of the input and output data before sampling instant t, $g_i, i=1,\cdots,m$ are the unknown parameters, $v(t)$ is a Gaussian white noise and satisfies $v(t)\sim N(0,\delta^2)$, and $\varpi(t)$ is an outlier. In general, model (1) contaminated by a Gaussian white noise is termed as normal model; while model (2) contaminated by outliers is called abnormal model. Assume that the number of the collected data is L, $s+q=L$ and $q>m$. Define the cost function

$$J(g_1,\cdots,g_m)=\sum_{t=1}^{L}\sum_{i=1}^{m}[y(t)-g_i\varphi_i(t)]^2.$$

Decompose the above cost function into two parts

$$\begin{aligned}J(g_1,\cdots,g_m)=&\sum_{l=1}^{q}\sum_{i=1}^{m}[y(t_{no_l})-g_i\varphi_i(t_{no_l})]^2+\\&\sum_{o=1}^{s}\sum_{i=1}^{m}[y(t_{ab_o})-g_i\varphi_i(t_{ab_o})]^2.\end{aligned} \tag{3}$$

Indeed, those abnormal sampling instants are unknown a priori, and their corresponding data sets have negative impact on the parameter estimation. The focus of this paper is to use the proposed algorithms to pick out the contaminated data/abnormal model, and then to obtain unbiased parameter estimates.

B. Review of the traditional algorithm

Define the following two vectors,

$$\begin{aligned}\mathbf{G}&=[g_1,\cdots,g_m]^{\mathrm{T}},\\ \boldsymbol{\varphi}(t)&=[\varphi_1(t),\cdots,\varphi_m(t)]^{\mathrm{T}}.\end{aligned}$$

Then, Equation (3) is rewritten as

$$\begin{aligned}J(\mathbf{G})=&\sum_{l=1}^{q}[y(t_{no_l})-\boldsymbol{\varphi}^{\mathrm{T}}(t_{no_l})\mathbf{G}]^2+\\&\sum_{o=1}^{s}[y(t_{ab_o})-\boldsymbol{\varphi}^{\mathrm{T}}(t_{ab_o})\mathbf{G}]^2.\end{aligned} \tag{4}$$

For the systems with outliers, the LS algorithm can be written as [32],

$$\begin{aligned}\hat{\mathbf{G}}=&[\sum_{l=1}^{q}\boldsymbol{\varphi}(t_{no_l})\boldsymbol{\varphi}^{\mathrm{T}}(t_{no_l})+\sum_{o=1}^{s}\boldsymbol{\varphi}(t_{ab_o})\boldsymbol{\varphi}^{\mathrm{T}}(t_{ab_o})]^{-1}\times\\&[\sum_{l=1}^{q}\boldsymbol{\varphi}(t_{no_l})y(t_{no_l})+\sum_{o=1}^{s}\boldsymbol{\varphi}(t_{ab_o})y(t_{ab_o})].\end{aligned} \tag{5}$$

The expectation of the estimate is

$$\begin{aligned}\mathrm{E}[\hat{\mathbf{G}}]=&\mathrm{E}\Big\{[\sum_{l=1}^{q}\boldsymbol{\varphi}(t_{no_l})\boldsymbol{\varphi}^{\mathrm{T}}(t_{no_l})+\sum_{o=1}^{s}\boldsymbol{\varphi}(t_{ab_o})\boldsymbol{\varphi}^{\mathrm{T}}(t_{ab_o})]^{-1}\times\\&[\sum_{l=1}^{q}\boldsymbol{\varphi}(t_{no_l})y(t_{no_l})+\sum_{o=1}^{s}\boldsymbol{\varphi}(t_{ab_o})y(t_{ab_o})]\Big\}\\=&\mathrm{E}\Big\{[\sum_{l=1}^{q}\boldsymbol{\varphi}(t_{no_l})\boldsymbol{\varphi}^{\mathrm{T}}(t_{no_l})+\sum_{o=1}^{s}\boldsymbol{\varphi}(t_{ab_o})\boldsymbol{\varphi}^{\mathrm{T}}(t_{ab_o})]^{-1}\times\\&[\sum_{l=1}^{q}\boldsymbol{\varphi}(t_{no_l})\boldsymbol{\varphi}^{\mathrm{T}}(t_{no_l})\mathbf{G}+\boldsymbol{\varphi}(t_{no_l})v(t_{no_l})+\\&\sum_{o=1}^{s}\boldsymbol{\varphi}(t_{ab_o})\boldsymbol{\varphi}^{\mathrm{T}}(t_{ab_o})\mathbf{G}+\boldsymbol{\varphi}(t_{ab_o})\varpi(t_{ab_o})]\Big\}\\=&\mathbf{G}+\mathrm{E}\Big\{[\sum_{l=1}^{q}\boldsymbol{\varphi}(t_{no_l})\boldsymbol{\varphi}^{\mathrm{T}}(t_{no_l})+\sum_{o=1}^{s}\boldsymbol{\varphi}(t_{ab_o})\boldsymbol{\varphi}^{\mathrm{T}}(t_{ab_o})]^{-1}\times\\&[\boldsymbol{\varphi}(t_{no_l})v(t_{no_l})+\boldsymbol{\varphi}(t_{ab_o})\varpi(t_{ab_o})]\Big\}.\end{aligned}$$

Since the noise $v(t_{no_l})$ is Gaussian white, the above equation can be simplified as

$$\begin{aligned}\mathrm{E}[\hat{\mathbf{G}}]=\mathbf{G}+\mathrm{E}\Big\{&[\sum_{l=1}^{q}\boldsymbol{\varphi}(t_{no_l})\boldsymbol{\varphi}^{\mathrm{T}}(t_{no_l})+\sum_{o=1}^{s}\boldsymbol{\varphi}(t_{ab_o})\boldsymbol{\varphi}^{\mathrm{T}}(t_{ab_o})]^{-1}\times\\&[\boldsymbol{\varphi}(t_{ab_o})\varpi(t_{ab_o})]\Big\}.\end{aligned} \tag{6}$$

Those outliers $\varpi(t_{ab_o}), o=1,\cdots,s$ make the estimates $\hat{\mathbf{G}}$ biased.

Unlike the LS algorithm, the GD algorithm generates an estimation sequence. By designing the negative direction and its corresponding step-size, such a sequence can converge to the true values. Let $\hat{\mathbf{G}}_{k-1}$ be the estimate of $\mathbf{G}$ in iteration $k-1$. Using the GD algorithm to update the parameter vector yields the following iterative function,

$$\begin{aligned}\hat{\mathbf{G}}_k=&\hat{\mathbf{G}}_{k-1}+\lambda\sum_{l=1}^{q}\boldsymbol{\varphi}(t_{no_l})[y(t_{no_l})-\boldsymbol{\varphi}^{\mathrm{T}}(t_{no_l})\hat{\mathbf{G}}_{k-1}]+\\&\lambda\sum_{o=1}^{s}\boldsymbol{\varphi}(t_{ab_o})[y(t_{ab_o})-\boldsymbol{\varphi}^{\mathrm{T}}(t_{ab_o})\hat{\mathbf{G}}_{k-1}],\end{aligned} \tag{7}$$

where λ is the step-size, one can use the method in [15] to choose a suitable step-size, that is,

$$0<\lambda<\frac{2}{\lambda_{max}[\sum\limits_{l=1}^{q}\boldsymbol{\varphi}(t_{no_l})\boldsymbol{\varphi}^{\mathrm{T}}(t_{no_l})+\sum\limits_{o=1}^{s}\boldsymbol{\varphi}(t_{ab_o})\boldsymbol{\varphi}^{\mathrm{T}}(t_{ab_o})]},$$

wherein $\lambda\sum\limits_{o=1}^{s}\boldsymbol{\varphi}(t_{ab_o})[y(t_{ab_o})-\boldsymbol{\varphi}^{\mathrm{T}}(t_{ab_o})\hat{\mathbf{G}}_{k-1}]$ makes the estimates biased.

Clearly, if the systems are contaminated by outliers, the traditional LS and GD algorithms are both inefficient [8], [9], [31].

Remark 1: Two methods can eliminate the bias: (1) study the characteristics of the outliers and try to diminish the effect caused by these outliers [23]–[25]; (2) pick out all the abnormal data, and then estimate the parameters based on the normal data.

III. Iterative reweighted algorithm

The basic idea of iterative reweighted (IRE) algorithm is to assign different weights for each cost function, and those weights of the abnormal data have smaller values or equal to zero. Finally, all the cost functions of the abnormal data are picked out from the original cost function.

A. Weighted LS algorithm

Define the cost function as

$$\begin{aligned}J(\mathbf{G})=&\sum_{l=1}^{q}w(t_{no_l})[y(t_{no_l})-\boldsymbol{\varphi}^{\mathrm{T}}(t_{no_l})\mathbf{G}]^2+\\&\sum_{o=1}^{s}w(t_{ab_o})[y(t_{ab_o})-\boldsymbol{\varphi}^{\mathrm{T}}(t_{ab_o})\mathbf{G}]^2.\end{aligned} \tag{8}$$

At first, there is no quantifiable confidence of data. Therefore, all the initial weights are assigned as the same value, that is

$$w_0(t_{no_l})=w_0(t_{ab_o})=\frac{1}{L},\ l=1,\cdots,q,\ o=1,\cdots,s.$$

The parameters updated by the LS algorithm are formulated as follows

$$\hat{\mathbf{G}}_k=[\sum_{l=1}^{q}w_{k-1}(t_{no_l})\boldsymbol{\varphi}(t_{no_l})\boldsymbol{\varphi}^{\mathrm{T}}(t_{no_l})+$$

$$\sum_{o=1}^{s} w_{k-1}(t_{ab_o})\boldsymbol{\varphi}(t_{ab_o})\boldsymbol{\varphi}^{\mathrm{T}}(t_{ab_o})]^{-1} \times$$
$$[\sum_{l=1}^{q} w_{k-1}(t_{no_l})\boldsymbol{\varphi}(t_{no_l})y(t_{no_l}) +$$
$$\sum_{o=1}^{s} w_{k-1}(t_{ab_o})\boldsymbol{\varphi}(t_{ab_o})y(t_{ab_o})]. \tag{9}$$

Remark 2: When $w_{k-1}(t_{no_l}) = w_{k-1}(t_{ab_o}) = \frac{1}{L}$, the above equation is the same as Equation (3). To obtain unbiased estimates, one should update the weights in each iteration.

Actually, an ideal weight designing method is to assign smaller weights for the cost functions whose output data are contaminated by outliers. To this end, we first compute the residual errors of each data set once $\hat{\mathbf{G}}_k$ is obtained,

$$\varepsilon_k(t_{no_l}) = |y(t_{no_l}) - \boldsymbol{\varphi}^{\mathrm{T}}(t_{no_l})\hat{\mathbf{G}}_k|, l = 1, \cdots, q, \tag{10}$$
$$\varepsilon_k(t_{ab_o}) = |y(t_{ab_o}) - \boldsymbol{\varphi}^{\mathrm{T}}(t_{ab_o})\hat{\mathbf{G}}_k|, o = 1, \cdots, s. \tag{11}$$

Two methods can be applied to compute the weights:

Method 1:

Let

$$\bar{w}_k(t) = \frac{1}{\varepsilon_k(t)}$$

and

$$W_k = \sum_{t=1}^{L} \bar{w}_{k-1}(t).$$

Then, normalizing these weights yields

$$w_k(t) = \frac{\bar{w}_k(t)}{W_k}.$$

Method 2:

Let

$$\varepsilon_k^{max} = \max\{\varepsilon_k(t_{no_1}), \cdots, \varepsilon_k(t_{no_q}), \varepsilon_k(t_{ab_1}), \cdots, \varepsilon_k(t_{ab_s})\}.$$

Then, the weights for each data set can be assigned as

$$w_k(t) = \frac{\varepsilon_k^{max} - \varepsilon_k(t)}{m\varepsilon_k^{max} - \sum_{l=1}^{q} \varepsilon_k(t_{no_l}) - \sum_{o=1}^{s} \varepsilon_k(t_{ab_o})}. \tag{12}$$

The steps of the weighted LS (W-LS) algorithm are listed as follows:

W-LS algorithm

(1) Collect $y(t)$ and $\boldsymbol{\varphi}(t), t = 1, \cdots, L$

(2) Form $\hat{\mathbf{G}}_0 = [0, 0, \cdots, 0]^{\mathrm{T}}$ and $w_0(t) = 1/L, t = 1, \cdots, L$

(3) **for** $k = 1, 2, \cdots,$ do

(3-1) Update $\hat{\mathbf{G}}_k$ according to Equation (9)

(3-2) Compute $\varepsilon_k(t),\ t = 1, \cdots, L$ based on Equations (10) and (11)

(3-3) Update $w_k(t),\ t = 1, \cdots, L$ according to Equation (12)

end

until convergence

Remark 3: According to Equation (12), at least one model is picked out from the stacked model in each iteration. When the iterations become larger, some normal models will mistake for abnormal models. To deal with this dilemma, we can choose

$$\varepsilon_k^{max} = \max\{\varepsilon_k(t_{no_1}), \cdots, \varepsilon_k(t_{no_q}), \varepsilon_k(t_{ab_1}), \cdots, \varepsilon_k(t_{ab_s})\} + \varrho,$$

where ϱ is a small positive constant.

Remark 4: Different from the traditional LS algorithm, the W-LS algorithm cannot yield the parameter estimates in only one iteration because the weights are regarded as hidden variables. The parameters and weights are interactively estimated until both converge to their true values.

B. Weighted GD algorithm

Based on the cost function in (8), the direction of the weighted GD (W-GD) algorithm is computed by

$$d_k = \sum_{l=1}^{q} w(t_{no_l})\boldsymbol{\varphi}(t_{no_l})[y(t_{no_l}) - \boldsymbol{\varphi}^{\mathrm{T}}(t_{no_l})\hat{\mathbf{G}}_{k-1}] +$$
$$\sum_{o=1}^{s} w(t_{ab_o})\boldsymbol{\varphi}(t_{ab_o})[y(t_{ab_o}) - \boldsymbol{\varphi}^{\mathrm{T}}(t_{ab_o})\hat{\mathbf{G}}_{k-1}]. \tag{13}$$

Then, the parameter estimates are updated by

$$\hat{\mathbf{G}}_k = \hat{\mathbf{G}}_{k-1} + \lambda_k d_k, \tag{14}$$

where

$$0 < \lambda_k < \frac{2}{\lambda_{max}[M_k + N_k]},$$
$$M_k = \sum_{l=1}^{q} w_{k-1}(t_{no_l})\boldsymbol{\varphi}(t_{no_l})\boldsymbol{\varphi}^{\mathrm{T}}(t_{no_l}),$$
$$N_k = \sum_{o=1}^{s} w_{k-1}(t_{ab_o})\boldsymbol{\varphi}(t_{ab_o})\boldsymbol{\varphi}^{\mathrm{T}}(t_{ab_o}).$$

The weights $w_{k-1}(t_{no_l})$ and $w_{k-1}(t_{ab_o})$ are computed by Equation (12).

Remark 5: Unlike the GD algorithm, the W-GD algorithm requires computing the eigenvalues of the information matrix $[M_k + N_k]$ in each iteration because the weights are always changing. It will lead to heavy computational efforts [33], [34].

C. Iterative reweighted algorithm

In the iterative reweighted (IRE) algorithm, the cost function is defined as

$$J(\mathbf{G}) = \Big[\sum_{l=1}^{q} |y(t_{no_l}) - \boldsymbol{\varphi}^{\mathrm{T}}(t_{no_l})\mathbf{G}|^p +$$
$$\sum_{o=1}^{s} |y(t_{ab_o}) - \boldsymbol{\varphi}^{\mathrm{T}}(t_{ab_o})\mathbf{G}|^p\Big], \tag{15}$$

where $\|\alpha\|_p = (|\alpha_1|^p + |\alpha_2|^p + \cdots + |\alpha_n|^p)^{\frac{1}{p}} (\alpha \in \mathbb{R}^n)$ is a p-norm. Transform $J(\mathbf{G})$ into

$$J(\mathbf{G}) = \Big[\sum_{l=1}^{q} |y(t_{no_l}) - \boldsymbol{\varphi}^{\mathrm{T}}(t_{no_l})\mathbf{G}|^{p-2}[y(t_{no_l}) - \boldsymbol{\varphi}^{\mathrm{T}}(t_{no_l})\mathbf{G}]^2 +$$
$$\sum_{o=1}^{s} |y(t_{ab_o}) - \boldsymbol{\varphi}^{\mathrm{T}}(t_{ab_o})\mathbf{G}|^{p-2}[y(t_{ab_o}) - \boldsymbol{\varphi}^{\mathrm{T}}(t_{ab_o})\mathbf{G}]^2\Big]. \tag{16}$$

Compare Equation (16) with Equation (15), the weights can be assigned as

$$w_k(t_{no_l}) = |y(t_{no_l}) - \boldsymbol{\varphi}^{\mathrm{T}}(t_{no_l})\mathbf{G}|^{p-2},$$
$$w_k(t_{ab_o}) = |y(t_{ab_o}) - \boldsymbol{\varphi}^{\mathrm{T}}(t_{ab_o})\mathbf{G}|^{p-2}.$$

Since the parameter vector $\mathbf{G}$ is unknown, we can use its estimate in iteration $k-1$ to replace it,

$$w_k(t_{no_l}) = |y(t_{no_l}) - \boldsymbol{\varphi}^{\mathrm{T}}(t_{no_l})\hat{\mathbf{G}}_{k-1}|^{p-2},$$
$$w_k(t_{ab_o}) = |y(t_{ab_o}) - \boldsymbol{\varphi}^{\mathrm{T}}(t_{ab_o})\hat{\mathbf{G}}_{k-1}|^{p-2}.$$

A larger error $|y(t_{no_l}) - \boldsymbol{\varphi}^{\mathrm{T}}(t_{no_l})\hat{\mathbf{G}}_{k-1}|$ means that the output encounters an outlier, and then its corresponding weight should be assigned as a smaller value. Therefore, we usually assume that $0 < p < 2$. For different p, we have

1) When $p = 2$, the IRE algorithm is the same as the LS algorithm, and in this case, the parameter estimates are biased.
2) When $p = 1$, the weight is computed by
$$w_k(t) = \frac{1}{|y(t) - \boldsymbol{\varphi}^{\mathrm{T}}(t)\hat{\mathbf{G}}_{k-1}|},$$
which means that the IRE algorithm is equivalent to the W-LS algorithm.
3) When $0 < p < 1$, since the parameter estimates in the first few iterations are not accurate, we have no confidence of the residual errors. Therefore, in the first few iterations of the IRE algorithm, a small p is better.
4) When $1 < p < 2$, in the last few iterations, a large p is better.
5) When $p > 2$, the IRE algorithm is divergent.

Remark 6: Based on the discussions mentioned above, we can choose a small p in the first few iterations, and then assign a larger p in the remaining iterations.

The weights of the IRE-LS and IRE-GD algorithms are updated by
$$w_{k-1}(t) = |y(t) - \boldsymbol{\varphi}^{\mathrm{T}}(t)\hat{\mathbf{G}}_{k-1}|^{p-2}. \tag{17}$$

Then, the steps of the IRE algorithms are listed as follows:

IRE algorithms

(1) Collect $y(t)$ and $\boldsymbol{\varphi}(t), t = 1, \cdots, L$
(2) Form $\hat{\mathbf{G}}_0 = [0, 0, \cdots, 0]^{\mathrm{T}}$, $w_0(t) = 1/L$
(3) **for** $k = 1, 2, \cdots$, do
  (3-1) Update $\hat{\mathbf{G}}_k$ based on Equation (9) or (14)
  (3-2) Compute $w_k(t),\ t = 1, \cdots, L$ according to Equation (17)
  (3-3) Normalize $w_k(t),\ t = 1, \cdots, L$
  (3-4) Compare $w_k(t),\ t = 1, \cdots, L$ with ϵ, if $w_k(t) < \epsilon$ ($\epsilon > 0$ is given a priori), let $w_k(t) = 0$
end
until convergence

Remark 7: To pick out the models which are contaminated by outliers, we can assign a small positive constant ϵ a priori. If the weight is smaller than ϵ, its corresponding model can be regarded as an abnormal model.

IV. Properties of iterative reweighted algorithms

In this section, we derive some properties of the IRE algorithms.

A. Convergence properties

The convergence properties of the IRE algorithms are given in the following theorems.

Theorem 1: Assume that the system with outliers is written by (1), and the number of the collected data sets is L (the number of the normal data sets is q, and $\frac{L}{2} < q \leqslant L$). The parameter estimates are updated using (9) and (17). Then, the IRE-LS algorithm is convergent, and the estimates $\hat{\mathbf{G}}_k$ are unbiased when $L \to \infty$.

Proof: Define the normal and abnormal weight sets as
$$W(no) = \{w(t_{no_1}), \cdots, w(t_{no_q})\},$$
$$W(ab) = \{w(t_{ab_1}), \cdots, w(t_{ab_s})\}.$$

Rewrite the cost function in iteration $k-1$ as
$$J(W_{k-1}(no), W_{k-1}(ab), \hat{\mathbf{G}}_{k-1}) = \sum_{l=1}^{q} w_{k-1}(t_{no_l})[y(t_{no_l}) - \boldsymbol{\varphi}^{\mathrm{T}}(t_{no_l})\hat{\mathbf{G}}_{k-1}]^2 + \sum_{o=1}^{s} w_{k-1}(t_{ab_o})[y(t_{ab_o}) - \boldsymbol{\varphi}^{\mathrm{T}}(t_{ab_o})\hat{\mathbf{G}}_{k-1}]^2,$$
which denotes the errors between the true outputs and predicted outputs in iteration $k-1$. To prove that the IRE-LS algorithm is convergent, we aim to obtain that the cost function in iteration k is less than or equal to the cost function in iteration $k-1$.

Fixing the weights $W_{k-1}(no)$ and $W_{k-1}(ab)$ and using the LS algorithm to update the parameters, we have
$$J(W_{k-1}(no), W_{k-1}(ab), \hat{\mathbf{G}}_k) \leqslant J(W_{k-1}(no), W_{k-1}(ab), \hat{\mathbf{G}}_{k-1}).$$

Let $\hat{\mathbf{G}}_k$ be fixed, then compute the weights. Clearly, the following inequalities hold
$$w_k(t_{no_l}) \geqslant w_{k-1}(t_{no_l}),\ l = 1, \cdots, q,$$
$$w_k(t_{ab_o}) \leqslant w_{k-1}(t_{ab_o}),\ o = 1, \cdots, s.$$

Since
$$[y(t_{no_l}) - \boldsymbol{\varphi}^{\mathrm{T}}(t_{no_l})\hat{\mathbf{G}}_k]^2 \leqslant [y(t_{ab_o}) - \boldsymbol{\varphi}^{\mathrm{T}}(t_{ab_o})\hat{\mathbf{G}}_{k-1}]^2,$$
$$l = 1, \cdots, q,\ o = 1, \cdots, s.$$

It gives rise to
$$J(W_k(no), W_k(ab), \hat{\mathbf{G}}_k) \leqslant J(W_{k-1}(no), W_{k-1}(ab), \hat{\mathbf{G}}_k).$$

Therefore, we can obtain
$$J(W_k(no), W_k(ab), \hat{\mathbf{G}}_k) \leqslant J(W_{k-1}(no), W_{k-1}(ab), \hat{\mathbf{G}}_{k-1}),$$
which shows that the IRE-LS algorithm is convergent.

Since the cost function is convex, once the IRE-LS algorithm is convergent, the weights of the abnormal data equal to zero. Then, the estimates are formulated as follows
$$\hat{\mathbf{G}}_k = [\sum_{l=1}^{q} w_{k-1}(no_l)\boldsymbol{\varphi}(no_l)\boldsymbol{\varphi}^{\mathrm{T}}(no_l)]^{-1}[\sum_{l=1}^{q} w_{k-1}(no_l)\boldsymbol{\varphi}(no_l)y(no_l)].$$

The expectation of $\hat{\mathbf{G}}_k$ is written by
$$\begin{aligned}
\mathrm{E}[\hat{\mathbf{G}}_k] &= \mathrm{E}\Big[[\sum_{l=1}^{q} w_{k-1}(no_l)\boldsymbol{\varphi}(no_l)\boldsymbol{\varphi}^{\mathrm{T}}(no_l)]^{-1} \times \\
&\quad [\sum_{l=1}^{q} w_{k-1}(no_l)\boldsymbol{\varphi}(no_l)[\boldsymbol{\varphi}^{\mathrm{T}}(no_l)\mathbf{G} + v(no_l)]]\Big] \\
&= \mathrm{E}\Big[[\sum_{l=1}^{q} w_{k-1}(no_l)\boldsymbol{\varphi}(no_l)\boldsymbol{\varphi}^{\mathrm{T}}(no_l)]^{-1} \times \\
&\quad [\sum_{l=1}^{q} w_{k-1}(no_l)\boldsymbol{\varphi}(no_l)\boldsymbol{\varphi}^{\mathrm{T}}(no_l)\mathbf{G}]\Big] \\
&\quad + \mathrm{E}\Big[[\sum_{l=1}^{q} w_{k-1}(no_l)\boldsymbol{\varphi}(no_l)\boldsymbol{\varphi}^{\mathrm{T}}(no_l)]^{-1} \times \\
&\quad [\sum_{l=1}^{q} w_{k-1}(no_l)\boldsymbol{\varphi}(no_l)v(no_l)]\Big] \\
&= \mathbf{G} + \mathrm{E}\Big[[\sum_{l=1}^{q} w_{k-1}(no_l)\boldsymbol{\varphi}(no_l)\boldsymbol{\varphi}^{\mathrm{T}}(no_l)]^{-1} \times \\
&\quad [\sum_{l=1}^{q} w_{k-1}(no_l)\boldsymbol{\varphi}(no_l)v(no_l)]\Big].
\end{aligned}$$

Since $v(no_l)$ is a Gaussian white noise with zero mean, and is independent of the information vector $\boldsymbol{\varphi}(no_l)$, the above equation can be simplified as
$$\mathrm{E}[\hat{\mathbf{G}}_k] = \mathbf{G}.$$

The proof is completed. ■

Remark 8: The IRE-LS algorithm should collect enough data to obtain unbiased estimates, this is untrue in engineering practices.

Theorem 2: For the system with outliers proposed in (1), assume that the number of the collected data sets is L, in which the number of the normal data sets is q, and $\frac{L}{2} < q \leqslant L$. The parameter estimates are updated using Equations (14) and (17). Then, the IRE-GD algorithm is convergent, and the estimates $\hat{\mathbf{G}}_k$ are unbiased when $L \to \infty$.

Proof: Since the direction is a negative direction, and the step-size λ_k satisfies

$$0 < \lambda_k < \frac{2}{\lambda_{max}[\sum\limits_{t=1}^{L} w_{k-1}(t)\boldsymbol{\varphi}(t)\boldsymbol{\varphi}^{\mathrm{T}}(t)]},$$

we have

$$J(W_{k-1}(no), W_{k-1}(ab), \hat{\mathbf{G}}_k) \leqslant J(W_{k-1}(no), W_{k-1}(ab), \hat{\mathbf{G}}_{k-1}).$$

The same way as the IRE-LS algorithm, we can obtain

$$J(W_k(no), W_k(ab), \hat{\mathbf{G}}_k) \leqslant J(W_{k-1}(no), W_{k-1}(ab), \hat{\mathbf{G}}_k),$$

and

$$J(W_k(no), W_k(ab), \hat{\mathbf{G}}_k) \leqslant J(W_{k-1}(no), W_{k-1}(ab), \hat{\mathbf{G}}_{k-1}).$$

It demonstrates that the IRE-GD algorithm is convergent.

Once the weights of the abnormal data converge to zero, the estimates using the IRE-GD algorithm are formulated as follows

$$\hat{\mathbf{G}}_k = \hat{\mathbf{G}}_{k-1} + \lambda_k d_k, \tag{18}$$

$$d_k = \sum_{l=1}^{q} w_{k-1}(no_l)\boldsymbol{\varphi}(no_l)[y(no_l) - \boldsymbol{\varphi}^{\mathrm{T}}(no_l)\hat{\mathbf{G}}_{k-1}], \tag{19}$$

$$0 < \lambda_k < \frac{2}{\lambda_{max}[\sum\limits_{l=1}^{q} w_{k-1}(no_l)\boldsymbol{\varphi}(no_l)\boldsymbol{\varphi}^{\mathrm{T}}(no_l)]}. \tag{20}$$

Subtracting the true value $\mathbf{G}$ on both sides of Equation (18) gives

$$\begin{aligned}\tilde{\mathbf{G}}_k &= \tilde{\mathbf{G}}_{k-1} + \lambda_k \sum_{l=1}^{q} w_{k-1}(no_l)\boldsymbol{\varphi}(no_l)[\boldsymbol{\varphi}^{\mathrm{T}}(no_l)\mathbf{G} + v(no_l) - \\ &\quad \boldsymbol{\varphi}^{\mathrm{T}}(no_l)\hat{\mathbf{G}}_{k-1}] \\ &= [\mathbf{I} - \lambda_k \sum_{l=1}^{q} w_{k-1}(no_l)\boldsymbol{\varphi}(no_l)\boldsymbol{\varphi}^{\mathrm{T}}(no_l)]\tilde{\mathbf{G}}_{k-1} + \\ &\quad \lambda_k \sum_{l=1}^{q} w_{k-1}(no_l)\boldsymbol{\varphi}(no_l)v(no_l),\end{aligned}$$

where $\tilde{\mathbf{G}}_k = \hat{\mathbf{G}}_k - \mathbf{G}$. According to Equation (20), we have

$$\|\mathbf{I} - \lambda_k \sum_{l=1}^{q} w_{k-1}(no_l)\boldsymbol{\varphi}(no_l)\boldsymbol{\varphi}^{\mathrm{T}}(no_l)\| < 1.$$

For the reason that the noise $v(no_l)$ is Gaussian white with zero mean, and is independent of $\boldsymbol{\varphi}(no_l)$. It gives rise to

$$\mathrm{E}[\tilde{\mathbf{G}}_k] = 0.$$

Therefore, the estimates are unbiased. ■

Remark 9: The IRE-LS algorithm performs a matrix inverse calculation in each iteration, while the IRE-GD algorithm should compute the eigenvalues of an information matrix. If the considered model has a high-order, both the methods have heavy computational efforts. The multi-direction method proposed in [33] can be used to reduce the computational efforts.

Remark 10: Since the weights of the models which are contaminated by outliers become smaller and smaller, the IRE-LS and IRE-GD algorithms proposed in this paper can alleviate the bad effect caused by the outliers. Thus, they can obtain unbiased estimates.

B. The number of the normal data

The number of the normal data plays an important role in the IRE algorithm convergence analyzing. Assume that the number of the normal data sets is q.

Case 1: $\frac{L}{2} < q \leqslant L$

In this case, the normal data play a more important role in estimating the parameters than the abnormal data, and the estimates asymptotically converge to the true values with increased numbers of k and q. In addition, the larger the number q is, the faster convergence rates the algorithms will have.

Case 2: $q < \frac{L}{2}$

When $q < \frac{L}{2}$, the IRE algorithm may be divergent. To deal with this problem, the EM or VB method proposed in [10], [25] is a good alternative.

Case 3: $q = \frac{L}{2}$

If the outlier is another kind of noise, the considered model can be regarded as a switching model which is constituted of two submodels: one is a normal model which is contaminated by a Gaussian white noise, and the other is an abnormal model whose noise is outlier. Then, the IRE algorithm is unavailable because these two kinds of models have the same number of data; On the other hand, if there are several kinds of outliers, and the number of the data for each abnormal model is less than $\frac{L}{2}$. In this case, the IRE algorithm is efficient.

V. Examples

For simplicity, in what follows, $\tau_k = \|\hat{\mathbf{G}}_k - \mathbf{G}\|/\|\mathbf{G}\|$ means the parameter estimation error in iteration k; $\sigma_k = \|w_k - w\|/\|w\|$ denotes the weight estimation error in iteration k.

A. Example 1

Consider the following model,

$$\begin{aligned}y(t) &= g_1 u(t-1) + g_2 u(t-2) + \cdots + g_8 u(t-8) + v(t), \\ \mathbf{G} &= [g_1, g_2, g_3, g_4, g_5, g_6, g_7, g_8]^{\mathrm{T}} = [2, -3, 2, 1, 1.5, 3, 0.8, -0.4]^{\mathrm{T}},\end{aligned}$$

where $u(t) \sim N(0, 1)$. We collect 1000 sets of data, where the noise sequence from $t = 1 : 500$ and $601 : 1000$ satisfies $v(t) \sim N(0, 0.1^2)$, while the other data from $t = 501 : 600$ are contaminated by outliers (10% percentage of data are contaminated by outliers).

Apply the traditional LS , GD, IRE-LS, and IRE-GD ($p = 1.2$) algorithms for the model. For fair comparison, the initial parameters $\boldsymbol{\theta}^0 = \mathbf{1}/10^6$ keep unchanging for all the algorithms ($\mathbf{1} = [1, 1, 1, 1, 1, 1, 1, 1]^{\mathrm{T}}$. The estimation errors τ_k versus k are shown in Fig. 1. The parameter estimates and the estimation errors are shown in Table I. To show the stability of the four algorithms, the boxplot of parameter estimates of different iterations are shown in Fig. 2 (the means and the outliers of the estimates of the 8 parameter elements).

In addition, use the IRE-LS and IRE-GD ($p = 1.2$) algorithms for the model: (1) the data from $601 : 1000$ are contaminated by outliers (40% percentage of data are contaminated by outliers); (2) the data from $701 : 1000$ are contaminated by outliers (30% percentage of data are contaminated by outliers). The estimation errors are shown in Fig. 3. The weight estimates are shown in Fig. 4 (30% percentage of data are contaminated by outliers). To show the tracking ability of the algorithm, assume that the true weights of the submodel from $t = 1 : 600$ are equal to 1 (not contaminated by outliers), while the others are equal to zero (contaminated by outliers). Assign a threshold $\xi = 0.0008$, if the weight estimate is larger than ξ, let it be equal to 1; otherwise, let it be equal to 0. The weight estimation errors σ_k versus k are shown in Fig. 5.

Based on this example, we can obtain:

1) the estimates of the IRE based methods are more accurate than the estimates of the traditional GD and LS algorithms, as shown in Fig. 1 and Table I;
2) the more data are contaminated by outliers, the slower convergence rates the algorithms will have, as shown in Fig. 3;

TABLE I
PARAMETER ESTIMATES AND ESTIMATION ERRORS

	k	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	τ (%)
LS	30	1.91612	-2.42922	1.81939	0.83119	1.37610	2.94540	0.79785	-0.19097	12.31815
	60	1.91612	-2.42922	1.81939	0.83119	1.37610	2.94540	0.79785	-0.19097	12.31815
	90	1.91612	-2.42922	1.81939	0.83119	1.37610	2.94540	0.79785	-0.19097	12.31815
	120	1.91612	-2.42922	1.81939	0.83119	1.37610	2.94540	0.79785	-0.19097	12.31815
	150	1.91612	-2.42922	1.81939	0.83119	1.37610	2.94540	0.79785	-0.19097	12.31815
GD	30	1.69562	-1.44599	1.55921	0.84896	1.15562	2.36507	0.73064	0.15782	34.43099
	60	1.87523	-2.13386	1.74584	0.83660	1.28801	2.77920	0.75134	-0.07615	18.76374
	90	1.91112	-2.33946	1.79822	0.83270	1.34298	2.89692	0.77631	-0.15283	14.24583
	120	1.91676	-2.40165	1.81317	0.83156	1.36409	2.93099	0.78918	-0.17822	12.90908
	150	1.91695	-2.42067	1.81752	0.83126	1.37185	2.94103	0.79460	-0.18669	12.50256
IRE-LS	30	1.99186	-2.98206	1.98528	0.98034	1.51099	3.00173	0.80380	-0.39295	0.62645
	60	1.99464	-2.99252	1.98611	0.98612	1.50897	2.99936	0.80625	-0.39262	0.46321
	90	1.99462	-2.99293	1.98663	0.98646	1.50936	2.99912	0.80587	-0.39303	0.45147
	120	1.99462	-2.99294	1.98664	0.98646	1.50936	2.99912	0.80587	-0.39303	0.45130
	150	1.99460	-2.99304	1.98673	0.98656	1.50942	2.99908	0.80582	-0.39311	0.44877
IRE-GD	30	1.82633	-2.01520	1.70540	0.87703	1.24623	2.62316	0.73372	-0.00879	22.08552
	60	1.95190	-2.60010	1.86532	0.90217	1.39404	2.91932	0.77360	-0.23730	8.84030
	90	1.96697	-2.73024	1.90342	0.91356	1.44149	2.97400	0.79206	-0.29718	5.92230
	120	1.96781	-2.76003	1.91332	0.91729	1.45508	2.98494	0.79800	-0.31361	5.24575
	150	1.96746	-2.76669	1.91579	0.91836	1.45869	2.98730	0.79962	-0.31807	5.08909
	True Values	2.00000	-3.00000	2.00000	1.00000	1.50000	3.00000	0.80000	-0.40000	

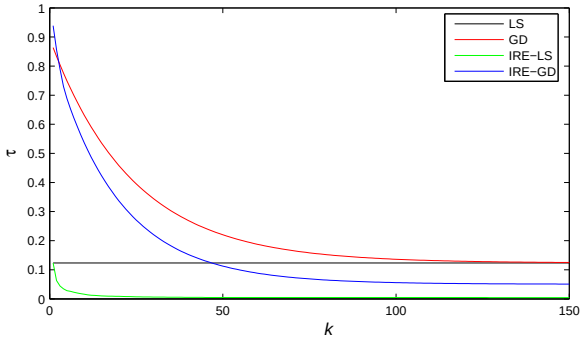


Fig. 1. Parameter estimation errors τ versus k

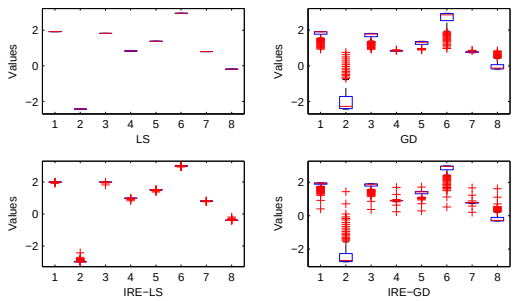


Fig. 2. Parameter estimates using the four algorithms for 150 iterations

3) the LS algorithms (traditional LS and IRE-LS algorithms) are more stable than the GD algorithms (traditional GD and IRE-GD algorithms), this can be shown in Fig. 2;
4) the IRE-LS algorithm has more accurate weight estimates than those of the IRE-GD algorithm, this can be shown in Figs. 4 and 5.

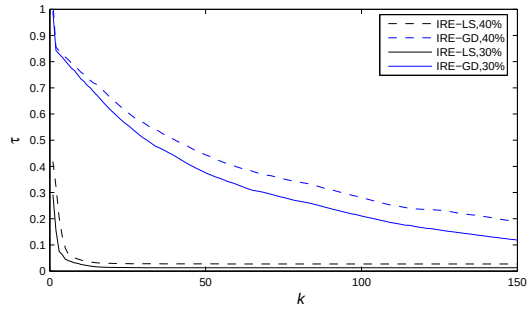


Fig. 3. Parameter estimation errors τ versus k

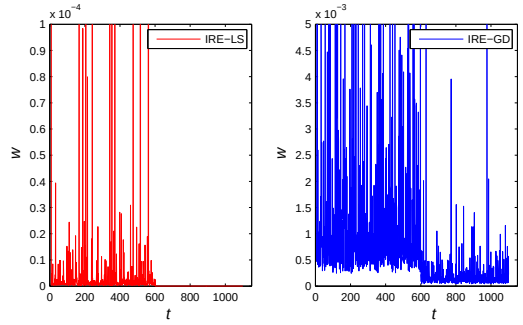


Fig. 4. Weight estimates

B. Example 2

In this example, we use the same model as in Example 1. However, the normal data and the abnormal data have the same number: (1) there is only one kind of outlier, e.g., $q = \frac{L}{2}$ for normal data, and $s = \frac{L}{2}$ for abnormal data; (2) there are several kinds of outliers, e.g., $q = \frac{L}{2}$ for normal data, and $s_i < \frac{L}{2}, \sum_{i=2}^{m} s_i = \frac{L}{2}, i = 2, \cdots, m$ for abnormal data.

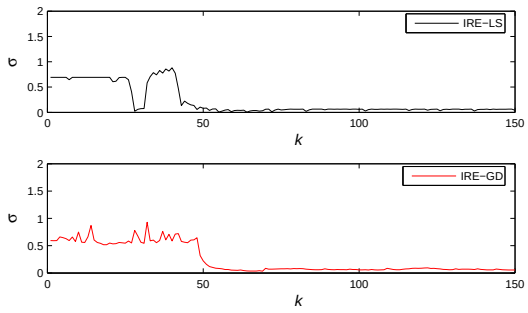


Fig. 5. Weight estimation errors σ versus k

Use the IRE-LS and IRE-GD ($p = 1.5$) algorithms to identify the model. The estimation errors are depicted in Fig. 6.

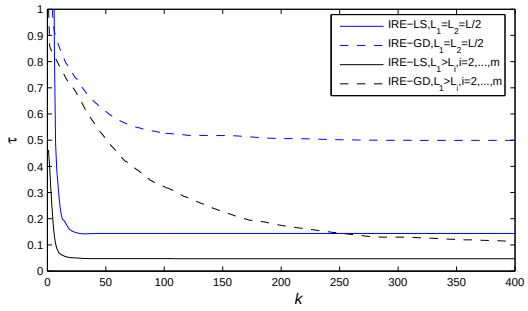


Fig. 6. Parameter estimation errors τ versus k

This example verifies the results of Case 3 in Section IV. B.

C. Example 3

A continuous stirred tank heater (CSTH) is taken as an example in this subsection [22]. Its structure is shown in Fig. 7. The steam valve position is $u(t)$. A temperature sensor located at the bottom of tank can collect the temperature of the water in the outflow pipe and transmit them over a communication network to the control center. Due to the complexity of the network, the outputs (temperature) $y(t)$ in the control center usually encounter noises, and some are even contaminated by outliers. The CSTH model can be written by

$$y(t) = 0.06012u(t-3) + 0.05390u(t-4) + 0.04832u(t-5) + 0.04332u(t-6) + v(t).$$

In simulation, we collect $L = 820$ sets of data. The input satisfies $u(t) \sim N(0,1)$. The noise $v(t)$ from $t = 1:500$ and $t = 601:750$ is a Gaussian white noise satisfies $v(t) \sim N(0, 0.1^2)$. The output data at other sampling instants are $y(501:600) = (rand(100,1)) \times 10$ and $y(751:820) = (rand(70,1)) \times 10$. The simulation data are shown in Fig. 8.

TABLE II
OPERATING CONDITIONS

Variable	Value
Level	12mA (20.48cm)
Cold valve	12.96mA (9.0383×10^{-5} m^3/s)
Steam valve	12.57mA
Temperature	10.5mA (42.52°C)

The parameter estimation errors using the IRE-LS and IRE-GD ($p = 1.8$) algorithms are depicted in Fig. 9. The weight estimates are shown in Fig. 10. Use the true parameters to recover the true outputs

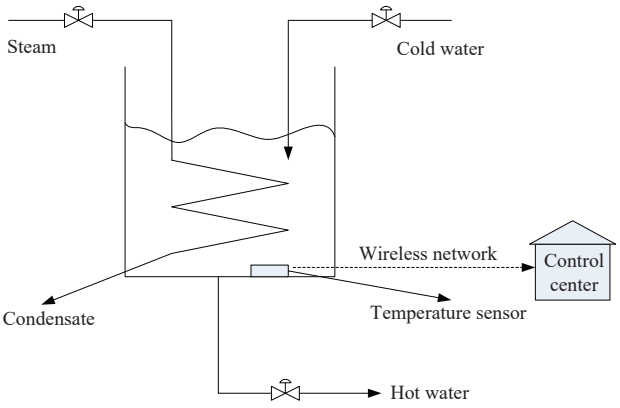


Fig. 7. CSTH system

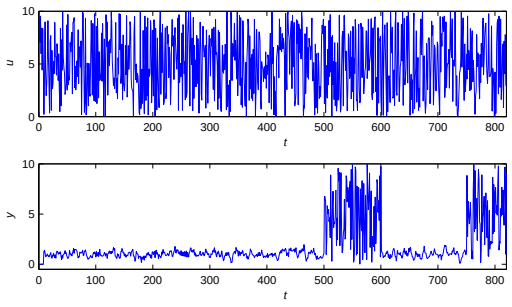


Fig. 8. Simulation data

($t = 500:600$), and the estimated parameters (IRE-LS) to compute the predicted outputs ($t = 500:600$). These two kinds of outputs are shown in Fig. 11.

Using the Monte Carlo method for the CSTH system (50 sets of noises), the estimation errors are shown in Fig. 12.

In addition, assume that the output data from $501:600$ and $751:820$ are $y(501:600) = (rand(100,1)) \times 15$ and $y(751:820) = 0$. The simulation data are shown in Fig. 13. The parameter estimation errors using the IRE-LS and IRE-GD ($p = 1.8$) algorithms are depicted in Fig. 14. Use the true parameters to recover the true outputs ($t = 500:600$), and the estimated parameters (IRE-LS) to compute the predicted outputs ($t = 500:600$). These two kinds of outputs are shown in Fig. 15.

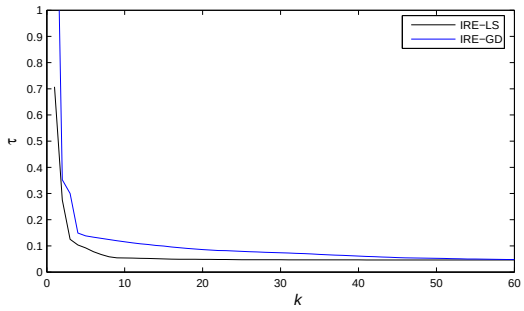


Fig. 9. Parameter estimation errors τ versus k

From this example, we have the following findings: (1) both the IRE-GD and IRE-LS algorithms are convergent, as shown in Fig. 9; (2) the weight estimates show that both the two algorithms can exactly catch the outlier/fault dynamics, this is demonstrated in Fig.

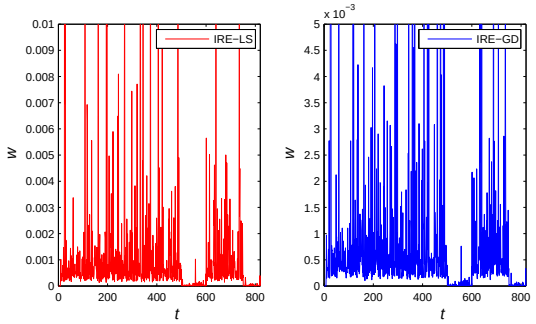


Fig. 10. Weight estimates

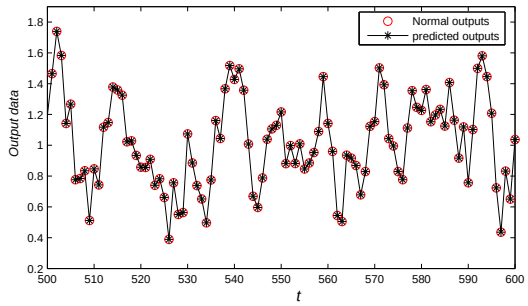


Fig. 11. True and predicted outputs

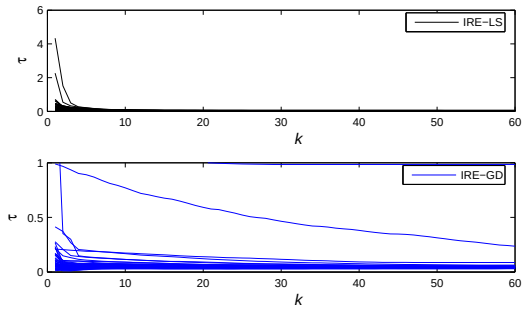


Fig. 12. Parameter estimation errors τ versus k (50 sets of noises)

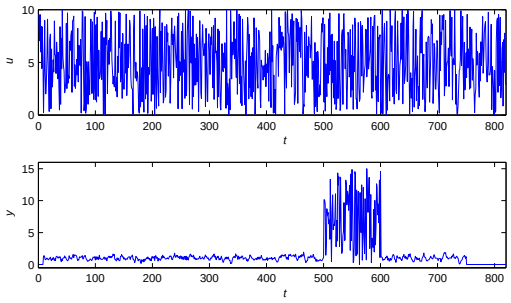


Fig. 13. Simulation data

10; (3) using the parameter estimates to predict the outputs, which can recover the true outputs during the outlier/fault instants, see Fig. 11; (4) compared with the IRE-GD algorithm, the IRE-LS algorithm is more robust to the noises, as shown in Fig. 12; (5) if the number of the abnormal data is much smaller than that of the normal data, the IRE algorithms are always effective.

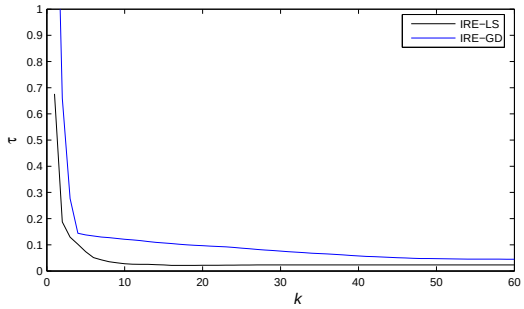


Fig. 14. Parameter estimation errors τ versus k

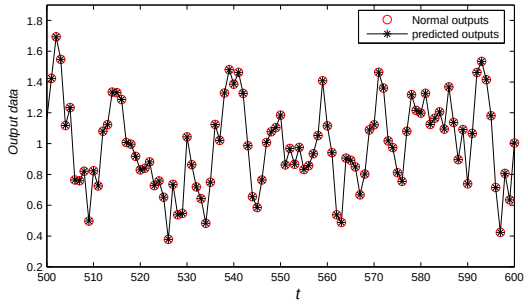


Fig. 15. True and predicted outputs

D. Example 4

Consider the following nonlinear model

$$y(t) = g_1 u^2(t-1) + g_2 u(t-1)u(t-2) + g_3 u(t-2)u(t-3) + g_4 u^2(t-4),$$
$$\mathbf{G} = [g_1, g_2, g_3, g_4]^{\mathrm{T}} = [1.2, 0.86, -0.64, 1.12]^{\mathrm{T}},$$

where $u(t) \sim N(0,1)$. We collect 1000 sets of data, and the noise sequence from $t = 1:800$ satisfies $v(t) \sim N(0, 0.1^2)$, while the other data from $t = 801:1000$ are contaminated by outliers.

Use the IRE-LS and IRE-GD algorithms for this model ($p = 1.3$). The parameter estimation errors are shown in Fig. 16.

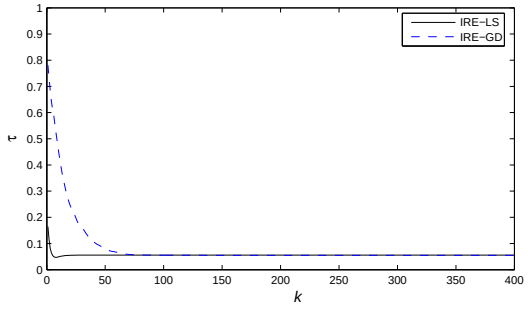


Fig. 16. Parameter estimation errors τ versus k

This example shows that the IRE-LS and IRE-GD algorithms are both effective for nonlinear models with outliers.

E. Example 5

A dryer system is considered in this subsection, where $u(t)$ is the capacity of the dryer, and $y(t)$ is the outlet temperature of the dryer. The sampling period $\Delta t = 0.08$ sec. Using the command 'load dryer2' in Matlab, we can generate 1000 sets of input-output data.

When $t = 1 : 800$, the data are contaminated by a Gaussian white noise, and the noise satisfies $v(t) \sim N(0, 0.01)$; when $t = 801 : 1000$, the data are contaminated by outliers. The simulation data are shown in Fig. 17.

Apply the IRE-LS and IRE-GD algorithms to identify the model ($p = 1.2$), and then predict the outputs $t = 801 : 1000$. The predicted outputs and the true outputs are shown in Fig. 18.

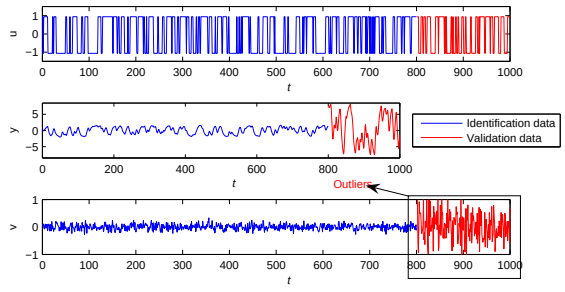


Fig. 17. Simulation data

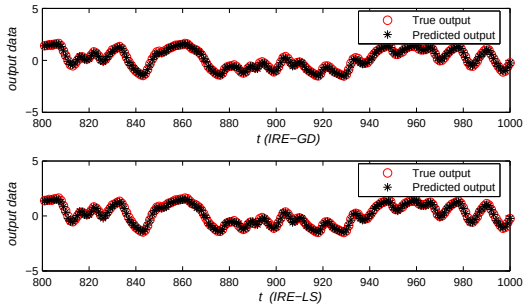


Fig. 18. True and predicted outputs

This example shows that the IRE-LS and IRE-GD algorithms are both effective for the dryer system.

VI. Conclusions

In this paper, we propose two iterative reweighted algorithms for systems contaminated by outliers. These two algorithms interactively update the parameter estimates and weight estimates, and take several advantages over the traditional identification algorithms:

1) compared with the GD and LS algorithms, the proposed algorithms proposed in this paper can obtain unbiased parameter estimates;
2) compared with the EM and VB algorithms, the proposed algorithms do not require any prior knowledge of the outliers, thus can be widely used in engineering practices;
3) the proposed algorithms can exactly catch the dynamics of the model by observing the weight estimates, thus can be extended to fault diagnosis and fault detection.

Therefore, the proposed algorithms are powerful and flexible tools for engineering and applied problems.

It is noteworthy that although the proposed algorithms have these advantages, there still remain some challenging and interesting topics. For example, how to choose the optimal p when computing the weights? and can the algorithms be applied to systems in which most of the data are contaminated by outliers?

Acknowledgments

The authors would like to thank the Associate Editor and the anonymous reviewers for their constructive and helpful comments and suggestions to improve the quality of this paper.

References

[1] N. Sammaknejad, Y. Zhao, B. Huang, A review of the expectation maximization algorithm in data-driven process identification, Journal of Process Control (2019) (73) 123-136.

[2] D.Q. Wang, S. Zhang, M. Gan, J.L. Qiu, A novel EM identification method for Hammerstein systems with missing output data, IEEE Transactions on Industrial Informatics (2020) 16 (4) 2500-2508.

[3] M. Gan, H.T. Zhu, G.Y. Chen, C.L.P Chen, Weighted generalized cross validation based regularization for broad learning system, IEEE Transactions on Cybernetics (2020). DOI: 10.1109/TCYB.2020.3015749.

[4] G.Y. Chen, M. Gan, C.L.P. Chen, L. Chen, A two-stage estimation algorithm based on variable projection method for GPS positioning, IEEE Transactions on Instrumentation Measurement (2018) 67 (11) 2518-2525.

[5] N. Lin, R. Chi, B. Huang, Event-triggered ILC for optimal consensus at specified data points of Heterogeneous networked agents with switching yopologies, IEEE Transactions on Cybernetics (2021). DOI: 10.1109/TCYB.2021.3054421

[6] F. Amjad, S.K. Varanasi, B. Huang, Kalman filter based convolutional neural network for robust tracking of froth-middling interface in a primary separation vessel in presence of occlusions, IEEE Transactions on Instrumentation Measurement (2021). DOI: 10.1109/TIM.2021.3060598

[7] N. Reamaroon, M.W. Sjoding, K. Lin, et al., Accounting for label uncertainty in machine learning for detection of acute respiratory distress syndrome, IEEE Journal of Biomedical and Health Informatics (2019) 23 (1) 407-415.

[8] G.C. Goodwin, K.S. Sin, *Adaptive Filtering, Prediction and Control*, Englewood CliPs, NJ: Prentice-Hall, (1984).

[9] T. Söderström, P. Stoica, *Systen Identification*, Englewood Cliffs, NJ: Prentice-Hall, (1989).

[10] X. Liu, X.Q. Yang, P.B. Zhu, Y.B. Wang, Robust multimodel identification of LPV systems with missing observations based on t-distribution, IEEE Transactions on Systems, Man, and Cybernetics: Systems (2021) 51 (8) 5254-5260.

[11] J.M. Li, F. Ding, Synchronous optimization schemes for dynamic systems through the kernel-based nonlinear observer canonical form, IEEE Transactions on Instrumentation and Measurement (2022) 71 3210952.

[12] C.P. Yu, M. Verhaegen, Subspace identification of individual systems operating in a network (SI^{2}ON), IEEE Transactions on Automatic Control 63 (4) (2019) 1120-1125.

[13] T.S. Chen, S.A. Martin, et al., System identification via sparse multiple kernel-based regularization using sequential convex optimization techniques, IEEE Transactions on Automatic Control (2014) 59 (11) 2933-2945.

[14] J. Chen, Q.M. Zhu, Y.J. Liu, Modified Kalman filtering based multi-step-length gradient iterative algorithm for ARX models with random missing outputs, Automatica (2020). DOI: 10.1016/j.automatica.2020.109034

[15] H. Ma, Y. Wang, Z.C. Ji, F. Ding, A novel three-stage quality oriented data-driven nonlinear industrial process monitoring strategy, IEEE Transactions on Instrumentation and Measurement (2022) 71 3524711.

[16] D.Q. Wang, Q.H. Fan, Y. Ma, An interactive maximum likelihood estimation method for multivariable Hammerstein systems, Journal of the Franklin Institute (2020) 357 (17) 12986-13005.

[17] J.X. Ma, B. Huang, F. Ding, Iterative identification of Hammerstein parameter varying systems with parameter uncertainties based on the variational bayesian approach, IEEE Transactions on Systems, Man, and Cybernetics: Systems (2020) 50 (3) 1035-1045.

[18] Y.H. Zhou, X. Zhang, F. Ding, Partially-coupled nonlinear parameter optimization algorithm for a class of multivariate hybrid models, Applied Mathematics and Computation 414 (2022) 126663.

[19] F. Ding, G. Liu, X.P. Liu, Partially coupled stochastic gradient identification methods for non-uniformly sampled systems, IEEE Transactions on Automatic Control 55 (8) (2010) 1976-1981.

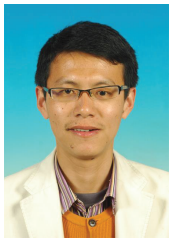
[20] H.P. Li, Y. Shi, W.S. Yan, F.Q. Liu, Receding horizon consensus of general linear multi-agent systems with input constraints: An inverse optimality approach, Automatica 91 (2018) 10-16.

[21] C.P. Yu, J. Chen, S.K. Li, M. Verhaegen, Identification of affinely parameterized state-space models with unknown inputs, Automatica 122 (2020). DOI: 10.1016/j.automatica.2020.109271

[22] J. Chen, M. Gan, Q.M. Zhu, Y.W. Mao, Varying infimum gradient descent algorithm for agent-server systems with uncertain communication network, IEEE Transactions on Instrumentation and Measurement (2021). DOI: 10.1109/TIM.2021.3070602

[23] X.Q. Yang, X. Liu, Z. Li, Multimodel approach to robust identification of multiple-input single-output nonlinear time-delay systems, IEEE Transactions on Industrial Informatics (2020) 16 (4) 2413-2422.

[24] M. Yu, T.Y. Zhang, X.Q. Yang, Identification of ARX system based on shifted asymmetric Laplace distribution, Chinese Control Conference, 27-30 July (2019).
[25] Y. Zhao, A. Fatehi, B. Huang, Robust estimation of ARX models with time varying time delays using variational Bayesian, IEEE Transactions on Cybernetics (2018) 48 (2) 532-542.
[26] P.L. Shui, L.X. Shi, H. Yu, Y.T. Huang, Iterative maximum likelihood and outlier-robust bipercentile estimation of parameters of compound-gaussian clutter with inverse Gaussian texture, IEEE Signal Processing Letters (2016) 23 (11) 1572-1576.
[27] J.X. Ma, J. Chen, W.L. Xiong, F. Ding, Expectation maximization estimation algorithm for Hammerstein models with non-Gaussian noise and random time delay from dual-rate sampled-data, Digital Signal Processing (2018) 73 135-144.
[28] Y. Lu, B. Huang, Robust multiple-model LPV approach to nonlinear process identification using mixture t distributions, Journal of Process Control (2014) 24, (9) 1472-1488.
[29] M. Zorzi, Empirical Bayesian learning in AR graphical models, Automatica 109 (2019) 108516.
[30] I. Daubechies, R. DeVore, M. Fornasier, C.S. Gunturk, Iteratively reweighted least squares minimization for sparse recovery, Communications on Pure and Applied Mathematics 63 (1) (2010) 1-38.
[31] F. Ding, *System Identification– New Theory and Methods*, Beijing: Science Press, (2013).
[32] J. Guo, B.Q. Mu, L.Y. Wang, G. Yin, L.J. Xu, Decision-based system identification and adaptive resource allocation, IEEE Transactions on Automatic Control 62 (5) (2017) 2166-2179.
[33] J. Chen, M. Gan, J.X. Ma, Multi-direction gradient iterative algorithm: a unified framework for gradient iterative and least squares algorithms, IEEE Transactions on Automatic Control (2021). DOI:10.1109/TAC.2021.3132262
[34] Y. Saad, *Iterative Methods for Sparse Linear Systems*, Society for Industrial and Applied Mathematics, (2003).



Jing Chen received his B.Sc. degree in the School of Mathematical Science and M.Sc. degree in the School of Information Engineering from Yangzhou University (Yanghzou, China) in 2003 and 2006, respectively, and received his Ph.D. degree in the School of Internet of Things Engineering, Jiangnan University (Wuxi, China) in 2013. He is currently a Professor in the School of Science, Jiangnan University (Wuxi, China). He is a Colleges and Universities Blue Project Middle-Aged Academic Leader (Jiangsu, China). His research interests include processing control and system identification.

Manfeng Hu received his Ph.D. degree in The Light Industry Technology and Engineering from Jiangnan University in 2008. He is currently a Professor in the School of Science, Jiangnan University (Wuxi, China). His research interests include network-based control, robust control theory and networked control under various communication protocols.

Yanjun Liu received the B.Sc. degree from Jiangsu University of Technology (Changzhou, China) in 2003, the M.Sc. degree and the Ph.D. degree from Jiangnan University (Wuxi, China) in 2009 and 2012, respectively. She is currently an associate Professor in the School of Internet of Things Engineering, Jiangnan University. Her research interests are system identification and parameter estimation.

Quanmin Zhu is Professor in control systems at the Department of Engineering Design and Mathematics, University of the West of England, Bristol, UK. He obtained his MSc in Harbin Institute of Technology, China in 1983 and PhD in Faculty of Engineering, University of Warwick, UK in 1989. His main research interest is in the area of nonlinear system modelling, identification, and control. He has published over 250 papers on these topics, edited various books with Springer, Elsevier, and the other publishers, and provided consultancy to various industries. Currently, Professor Zhu is acting as Editor of International Journal of Modelling, Identification and Control, Editor of International Journal of Computer Applications in Technology, Academic Editor of Complexity, Hindawi, Member of various journal editorial boards, and Editor of Elsevier book series of Emerging Methodologies and Applications in Modelling, Identification and Control. He is the founder and president of a series annual International Conference on Modelling, Identification and Control.