Available online at www.sciencedirect.com
ScienceDirect
Automatica 43 (2007) 1–14
automatica
www.elsevier.com/locate/automatica

Performance analysis of multi-innovation gradient type identification methods☆

Feng Ding[a,*], Tongwen Chen[b]

[a]Control Science and Engineering Research Center, Southern Yangtze University, Wuxi 214122, PR China
[b]Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Canada T6G 2V4

Received 7 October 2005; received in revised form 4 July 2006; accepted 27 July 2006
Available online 16 October 2006

Abstract

It is well-known that the stochastic gradient (SG) identification algorithm has poor convergence rate. In order to improve the convergence rate, we extend the SG algorithm from the viewpoint of innovation modification and present multi-innovation gradient type identification algorithms, including a multi-innovation stochastic gradient (MISG) algorithm and a multi-innovation forgetting gradient (MIFG) algorithm. Because the multi-innovation gradient type algorithms use not only the current data but also the past data at each iteration, parameter estimation accuracy can be improved. Finally, the performance analysis and simulation results show that the proposed MISG and MIFG algorithms have faster convergence rates and better tracking performance than their corresponding SG algorithms.

Keywords: Recursive identification; Parameter estimation; Stochastic gradient; Convergence properties; Forgetting factors; Stochastic processes

1. Introduction

Let us begin with considering a time-invariant stochastic system described by a linear regression model:

$$y(t) = \boldsymbol{\varphi}^{\mathrm{T}}(t)\boldsymbol{\theta} + v(t), \tag{1}$$

where $y(t) \in \mathbb{R}^1$ is the system output, and $\boldsymbol{\varphi}(t) \in \mathbb{R}^n$ is the information vector consisting of the system observation (input–output) data, $v(t) \in \mathbb{R}^1$ is a stochastic noise with zero mean and $\boldsymbol{\theta} \in \mathbb{R}^n$ ($\boldsymbol{\theta}(t) \in \mathbb{R}^n$) is the (time-varying) parameter vector to be identified, the superscript T denotes the matrix transpose.

Assume that $y(t) = 0$, $\boldsymbol{\varphi}(t) = \mathbf{0}$ and $v(t) = 0$ for $t \leqslant 0$. $\{y(t), \boldsymbol{\varphi}(t)\}$ is the available measurement data. For convenience, we suppose that t is the current time, then $y(t)$ and $\boldsymbol{\varphi}(t)$ are called the current data, and $\{y(t-i), \boldsymbol{\varphi}(t-i) : i = 1, 2, \ldots, p-1\}$ called the past data.

For the time-invariant system in (1), defining and minimizing the cost function (Ljung, 1999),

$$J(\boldsymbol{\theta}) := \mathrm{E}[\|y(t) - \boldsymbol{\varphi}^{\mathrm{T}}(t)\boldsymbol{\theta}\|^2],$$

and using the stochastic gradient (SG) search principle, we may obtain a recursive SG identification algorithm (Goodwin & Sin, 1984; Ljung, 1999),

$$\hat{\boldsymbol{\theta}}(t) = \hat{\boldsymbol{\theta}}(t-1) + \frac{\boldsymbol{\varphi}(t)}{r(t)}[y(t) - \boldsymbol{\varphi}^{\mathrm{T}}(t)\hat{\boldsymbol{\theta}}(t-1)], \tag{2}$$

$$r(t) = r(t-1) + \|\boldsymbol{\varphi}(t)\|^2, \quad r(0) = 1. \tag{3}$$

Here, E denotes the expectation operator, the norm of the matrix $\boldsymbol{X}$ is defined by $\|\boldsymbol{X}\|^2 = \mathrm{tr}[\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}]$, $\hat{\boldsymbol{\theta}}(t)$ represents the estimate of $\boldsymbol{\theta}$ at time t and $1/r(t)$ is called the convergence factor or step-size.

Comparing with the recursive least-squares algorithm, the SG algorithm has very slow convergence rate; we think that the main reasons lie in the following:

- The error system corresponding to the parameter estimation error equation has $n-1$ eigenvalues on the unit circle,

☆ This paper was not presented at any IFAC meeting. This paper was recommended for publication in revised form by Associate Editor Brett Ninness under the direction of Editor Torsten Söderström.

* Corresponding author. Tel.: +86 510 88637783; fax: +86 510 85910652.
E-mail addresses: fding@sytu.edu.cn (F. Ding), tchen@ece.ualberta.ca (T. Chen).

0005-1098/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.automatica.2006.07.024

only one eigenvalue inside the unit circle. In fact, defining the parameter estimation error vector,

$$\tilde{\boldsymbol{\theta}}(t) := \hat{\boldsymbol{\theta}}(t) - \boldsymbol{\theta}$$

and using (2) and (3), it follows that

$$\begin{aligned}\tilde{\boldsymbol{\theta}}(t) &= \left[\boldsymbol{I} - \frac{\boldsymbol{\varphi}(t)\boldsymbol{\varphi}^{\mathrm{T}}(t)}{r(t)}\right]\tilde{\boldsymbol{\theta}}(t-1) + \frac{\boldsymbol{\varphi}(t)}{r(t)}v(t)\\ &=: \boldsymbol{H}(t)\tilde{\boldsymbol{\theta}}(t-1) + \frac{\boldsymbol{\varphi}(t)}{r(t)}v(t), \quad \boldsymbol{H}(t) := \boldsymbol{I} - \frac{\boldsymbol{\varphi}(t)\boldsymbol{\varphi}^{\mathrm{T}}(t)}{r(t)},\end{aligned}$$

where $\boldsymbol{I}$ stands for an identity matrix of appropriate sizes. Since $\boldsymbol{H}^{\mathrm{T}}(t)\boldsymbol{H}(t)$ has $n-1$ eigenvalues on the unit circle, the SG algorithm thus has a poor convergence rate based on the fact: for time-varying systems of the form, $\boldsymbol{x}(t) = \boldsymbol{H}(t)\boldsymbol{x}(t-1)$, $\boldsymbol{x}(t) \in \mathbb{R}^n$, $\boldsymbol{H}(t) \in \mathbb{R}^{n\times n}$, if all eigenvalues of $\boldsymbol{H}^{\mathrm{T}}(t)\boldsymbol{H}(t)$ are close to zero or with magnitude smaller than 1, then $\boldsymbol{x}(t)$ can (fast) converge to zero, otherwise, if some eigenvalues of $\boldsymbol{H}^{\mathrm{T}}(t)\boldsymbol{H}(t)$ are on the unit circle , then $\boldsymbol{x}(t)$ has a slow convergence rate.

- The SG algorithm does not make *sufficient* use of the available information (data) of systems because the algorithm uses only the *current* data $\{y(t), \boldsymbol{\varphi}(t)\}$ at *each* recursive/iterative computation, and does not use the *past* data $\{y(t-i), \boldsymbol{\varphi}(t-i): i = 1, 2, \ldots\}$.

Therefore, a natural question is how to extend the SG algorithm to achieve a fast convergence rate; this is the focus of this work.

Since the quantity $e(t) := y(t) - \boldsymbol{\varphi}^{\mathrm{T}}(t)\hat{\boldsymbol{\theta}}(t-1) \in \mathbb{R}^1$ in Eq. (2) is called the *innovation* (Ljung, 1999) and scalar-valued, we give the SG algorithm a new name—the single innovation (modification) SG identification algorithm. From the viewpoint of innovation modification, this paper extends the single innovation identification algorithm and presents multi-innovation identification methods. The proposed approaches use not only the current data $\{y(t), \boldsymbol{\varphi}(t)\}$ but also the past data $\{y(t-i), \boldsymbol{\varphi}(t-i): i = 1, 2, \ldots, p-1\}$, and the $\boldsymbol{H}$-matrix of the resulting estimation error equations has *all* eigenvalues inside the unit disc, and thus achieving fast convergence rates and improving the parameter estimation accuracy. Finally, we claim that the multi-innovation identification methods are different from the identification methods of multi-variable systems—see the discussion in Section 5.

It is well-known that the recursive least-squares (RLS) algorithm is based on all previous data, thus has faster convergence rate than the SG algorithm, but the SG algorithm requires less computational effort than the RLS algorithm. In order to enhance the convergence rate of the SG algorithm, we present multi-innovation stochastic gradient (MISG) algorithms based on finite previous data, i.e., the MISG approaches use not only the current data but also the past data at each iteration, thus parameter estimation accuracy can be improved. The MISG algorithms have advantages of the SG and RLS algorithms. This is a tradeoff between the two algorithms, i.e., the MISG algorithms have faster convergence rate than the SG algorithms and less computational burden than the RLS algorithms.

In the area of time-invariant system identification, earlier work on convergence exists: Ljung (1976) analyzed consistency of the RLS algorithm based on the assumptions that the noise is an independent and identically distributed (iid) random sequence with finite fourth-order moments and the input and output signals have finite non-zero power. Also, Lai and Wei (1982) obtained the convergence rate of the RLS parameter estimation by assuming that higher-order moments of the noises exist. Since then, most results of RLS or SG (based adaptive control) algorithms have made such assumptions, e.g., Lai and Wei (1986), Wei (1987), Ren and Kumar (1994), Guo (1995), and Kumar (2000). Recently, Ding and Chen (2005a) and Ding, Shi, and Chen (2006) studied in details the convergence properties of the RLS algorithms for time-invariant systems and for non-stationary ARMA processes but do not assume that the process noise is an iid sequence or higher-order moments exist. This paper explores the convergence properties of the MISG algorithms only assuming that the noise is a second-moment process with zero mean.

In the literature of time-varying systems, Guo and Ljung (1995a) discussed the exponential stability of the averaged (deterministic) equations corresponding to the homogeneous equations of the parameter estimation error systems of the RLS algorithms with a forgetting factor (RFFLS algorithms for short), and further Guo and Ljung (1995b) used the stochastic martingale theory to study the properties of the parameter estimation error covariance matrix of the RFFLS algorithms by assuming that the measured error $v(t)$ and the parameter drift $w(t)$ are of white noise character. Recently, Ding and Chen (2005a) derived in details the upper and lower bounds of the parameter estimation of the RFFLS algorithms and showed that only for deterministic systems, the RFFLS algorithms are exponentially convergent. In this paper, we present a SG algorithm and MISG algorithm with a forgetting factor, capable of tracking time-varying parameters, and analyze their parameter estimation error bounds.

The rest of the paper is organized as follows. Section 2 derives a MISG identification algorithm by extending the innovation modification technique. Sections 3 analyzes the convergence properties of the SG and MISG algorithms to show the advantages of the proposed MISG algorithm. Section 4 presents the MISG algorithm with a forgetting factor in order to track the time-varying parameters. Section 5 simply discusses the multi-variable version of the multi-innovation algorithms. Section 6 presents several illustrative examples for the results in this paper. Finally, concluding remarks are given in Section 7.

2. The MISG algorithm

In this section, we derive a MISG identification algorithm. The basic idea is to expand the scalar innovation $e(t)$ to an innovation vector (called also multi-innovation) (Ding &

Chen, 2006). Let

$$\boldsymbol{E}(p,t)=\begin{bmatrix} e(t) \\ e(t-1) \\ \vdots \\ e(t-p+1) \end{bmatrix} \in \mathbb{R}^p,$$

where the positive integer p denotes the innovation length, and

$$e(t-i)=y(t-i)-\boldsymbol{\varphi}^{\mathrm{T}}(t-i)\hat{\boldsymbol{\theta}}(t-i-1) \in \mathbb{R}^1.$$

In general, one thinks that the estimate $\hat{\boldsymbol{\theta}}(t-1)$ at time $t-1$ is closer to $\boldsymbol{\theta}$ than $\hat{\boldsymbol{\theta}}(t-i)$ at time $t-i$ $(i=2,3,4,\ldots,p-1)$. Thus, the innovation vector is taken more reasonably to be

$$\boldsymbol{E}(p,t)=\begin{bmatrix} y(t)-\boldsymbol{\varphi}^{\mathrm{T}}(t)\hat{\boldsymbol{\theta}}(t-1) \\ y(t-1)-\boldsymbol{\varphi}^{\mathrm{T}}(t-1)\hat{\boldsymbol{\theta}}(t-1) \\ \vdots \\ y(t-p+1)-\boldsymbol{\varphi}^{\mathrm{T}}(t-p+1)\hat{\boldsymbol{\theta}}(t-1) \end{bmatrix} \in \mathbb{R}^p.$$

By defining the information matrix $\boldsymbol{\Phi}(p,t)$ and stacked output vector $\boldsymbol{Y}(p,t)$ as

$$\boldsymbol{\Phi}(p,t)=[\boldsymbol{\varphi}(t),\boldsymbol{\varphi}(t-1),\ldots,\boldsymbol{\varphi}(t-p+1)] \in \mathbb{R}^{n\times p},$$
$$\boldsymbol{Y}(p,t)=[y(t),y(t-1),\ldots,y(t-p+1)]^{\mathrm{T}} \in \mathbb{R}^p,$$

the innovation vector $\boldsymbol{E}(p,t)$ may be expressed as

$$\boldsymbol{E}(p,t)=\boldsymbol{Y}(p,t)-\boldsymbol{\Phi}^{\mathrm{T}}(p,t)\hat{\boldsymbol{\theta}}(t-1).$$

Since $\boldsymbol{E}(1,t)=e(t)$, $\boldsymbol{\Phi}(1,t)=\boldsymbol{\varphi}(t)$ and $\boldsymbol{Y}(1,t)=y(t)$, the SG algorithm in (2) may be equivalently expressed as

$$\hat{\boldsymbol{\theta}}(t)=\hat{\boldsymbol{\theta}}(t-1)+\frac{\boldsymbol{\Phi}(1,t)}{r(t)}[\boldsymbol{Y}(1,t)-\boldsymbol{\Phi}^{\mathrm{T}}(1,t)\hat{\boldsymbol{\theta}}(t-1)].$$

Here, the multi-innovation length p is equal to 1. From here, we can derive the MISG algorithm with the innovation length p as follows:

$$\hat{\boldsymbol{\theta}}(t)=\hat{\boldsymbol{\theta}}(t-1)+\frac{\boldsymbol{\Phi}(p,t)}{r(t)}\boldsymbol{E}(p,t), \tag{4}$$
$$\boldsymbol{E}(p,t)=\boldsymbol{Y}(p,t)-\boldsymbol{\Phi}^{\mathrm{T}}(p,t)\hat{\boldsymbol{\theta}}(t-1), \tag{5}$$
$$r(t)=r(t-1)+\|\boldsymbol{\varphi}(t)\|^2, \quad r(0)=1, \tag{6}$$
$$\boldsymbol{\Phi}(p,t)=[\boldsymbol{\varphi}(t),\boldsymbol{\varphi}(t-1),\ldots,\boldsymbol{\varphi}(t-p+1)] \in \mathbb{R}^{n\times p}, \tag{7}$$
$$\boldsymbol{Y}(p,t)=[y(t),y(t-1),\ldots,y(t-p+1)]^{\mathrm{T}} \in \mathbb{R}^p. \tag{8}$$

Because $\boldsymbol{E}(p,t) \in \mathbb{R}^p$ in this algorithm is an innovation vector, namely, multi-innovation, we refer to the algorithm in (4)–(8) as the MISG algorithm. As $p=1$, the MISG algorithm reduces to the standard SG algorithm.

This MISG algorithm may also be approximately obtained by minimizing the cost function,

$$J(\boldsymbol{\theta}):=\mathrm{E}[\|\boldsymbol{Y}(p,t)-\boldsymbol{\Phi}^{\mathrm{T}}(p,t)\boldsymbol{\theta}\|^2]$$

and using the SG search principle. Ding, Xie, and Fang (1996) and Ding, Xiao, and Ding (2003) gave some special multi-innovation identification algorithms, e.g., the varying iterative interval MISG algorithm (V-MISG algorithm) and the V-MISG algorithm with a forgetting factor.

The MISG algorithm has the following properties:

- Comparing with the SG algorithm in (2)–(3) using only the current data $\{y(t),\boldsymbol{\varphi}(t)\}$, the MISG algorithm in (4)–(8) uses not only the current data $\{y(t),\boldsymbol{\varphi}(t)\}$ but also the past data $\{y(t-i),\boldsymbol{\varphi}(t-i):\ i=1,2,\ldots,p-1\}$, and thus has potential for better convergence properties for the parameter estimation.
- The MISG algorithm repeatedly utilizes the available data. In fact, at time t, the data that the MISG algorithm uses are $\{y(t-i),\boldsymbol{\varphi}(t-i):\ i=0,1,\ldots,p-1\}$; at time $t+1$, the data the MISG algorithm use are $\{y(t+1-i),\boldsymbol{\varphi}(t+1-i):\ i=0,1,\ldots,p-1\}$; thus in the neighboring two iterations, the data of the repeated utilization are $\{y(t-i),\boldsymbol{\varphi}(t-i):\ i=0,1,\ldots,p-2\}$. This is the key for the enhanced accuracy of the MISG algorithm.
- Increasing the innovation length p leads to smaller parameter estimation errors for the same length measurement data. In other words, a larger p results in a better model accuracy but the price we paid is a large computational effort. However, this increased computation is still tolerable and affordable.

These conclusions are confirmed from the theoretical analysis in the next section and examples later.

To initialize the MISG algorithm, the initial value $\hat{\boldsymbol{\theta}}(0)$ is generally taken to be a zero vector or a small real vector, e.g., $\hat{\boldsymbol{\theta}}(0)=10^{-6}\mathbf{1}_n$ with $\mathbf{1}_n$ being an n-dimensional column vector whose elements are 1.

To summarize, we list the steps involved in the MISG algorithm to recursively compute the parameter estimation vector $\hat{\boldsymbol{\theta}}(t)$ as t increases:

1. Collect the measurement data, form the information vector $\boldsymbol{\varphi}(t)$ and determine a data length L_e.
2. Let $t=1$: $\hat{\boldsymbol{\theta}}(0)=10^{-6}\mathbf{1}_n$ and $r(0)=1$.
3. Form $\boldsymbol{\Phi}(p,t)$ by (7) and $\boldsymbol{Y}(p,t)$ by (8).
4. Compute $r(t)$ by (6), $\boldsymbol{E}(p,t)$ by (5), and $\hat{\boldsymbol{\theta}}(t)$ by (4).
5. If $t=L_e$, then terminate the procedure and obtain the estimate $\hat{\boldsymbol{\theta}}(L_e)$ of the parameter vector $\boldsymbol{\theta}$; otherwise, increment t by 1 and go to step 3.

3. Convergence of the SG and MISG algorithms

Let us introduce some notation first. $\lambda_{\max}[\boldsymbol{X}]$ and $\lambda_{\min}[\boldsymbol{X}]$ represent the maximum and minimum eigenvalues of the symmetric matrix $\boldsymbol{X}$, respectively; for $g(t)\geqslant 0$, we write $f(t)=O(g(t))$ or $f(t)\sim g(t)$ to express $\lim_{t\to\infty} f(t)/g(t)=1$.

In order to derive the convergence properties of the SG and MISG algorithms, the following lemmas are required.

Lemma 1. *Let $\{x(t)\}$, $\{a_t\}$ and $\{b_t\}$ be non-negative real sequences satisfying*

$$x(t+1)\leqslant(1-a_t)x(t)+b_t, \quad t\geqslant 0,$$

with $a_t \in [0, 1)$ *and* $x(0) < \infty$. *Then*

$$\lim_{t\to\infty} x(t) \leqslant \lim_{t\to\infty} \frac{b_t}{a_t}$$

provided that the related limits exist.

The proof is easy and omitted here.

Lemma 2. *For the system in* (1) *and the algorithm in* (2)–(3), *if the information vector* $\boldsymbol{\varphi}(t)$ *is persistently exciting, that is, there exist constants* $0 < \alpha \leqslant \beta < \infty$ *and an integer* $N \geqslant n$ *such that the following persistent excitation condition holds* (Ljung, 1999):

$$\text{(A1)}\ \alpha \boldsymbol{I} \leqslant \frac{1}{N}\sum_{i=0}^{N-1} \boldsymbol{\varphi}(t+i)\boldsymbol{\varphi}^{\mathrm{T}}(t+i) \leqslant \beta \boldsymbol{I} \quad a.s.,\ t > 0,$$

then $r(t)$ *in* (3) *or* (6) *satisfies the inequality*

$$n\alpha(t-N+1)+1 \leqslant r(t) \leqslant n\beta(t+N-1)+1 \quad a.s.$$

(The proofs of Lemmas 2 to 4 and Theorems 1–3 in the sequel are given in the Appendix.)

3.1. The convergence of the SG algorithm

This subsection establishes the convergence of the SG algorithm.

Lemma 3. *For the system in* (1) *and the algorithm in* (2)–(3), *define the transition matrix*

$$\boldsymbol{L}(t+1, i) = \left[\boldsymbol{I} - \frac{\boldsymbol{\varphi}(t)\boldsymbol{\varphi}^{\mathrm{T}}(t)}{r(t)}\right]\boldsymbol{L}(t, i), \quad \boldsymbol{L}(i, i) = \boldsymbol{I}.$$

If Condition (A1) *holds, then*

$$\begin{aligned}\rho_t &:= \lambda_{\max}[\boldsymbol{L}^{\mathrm{T}}(t+N, t)\boldsymbol{L}(t+N, t)] \\ &\leqslant 1 - \frac{N\alpha^2}{(N+1)^2\delta_1[n\beta(t+2N-2)+1]} \quad a.s.,\ \delta_1 := nN\beta.\end{aligned}$$

Theorem 1. *For the system in* (1) *and the algorithm in* (2)–(3), *assume that Condition* (A1) *holds,* $\{v(t)\}$ *is a random noise sequence satisfying*

$$\begin{aligned}&\text{(A2)}\quad \mathrm{E}[v(t)] = 0;\ \mathrm{E}[v(t)v(i)] = 0,\\ &\qquad t \neq i;\ \mathrm{E}[v^2(t)] \leqslant \sigma_v^2 < \infty;\end{aligned}$$

and $\tilde{\boldsymbol{\theta}}(0)$ *is uncorrelated with* $\{v(t)\}$ *and* $\mathrm{E}[\|\tilde{\boldsymbol{\theta}}(0)\|^2] < \infty$. *Then the parameter estimation error* $\hat{\boldsymbol{\theta}}(t) - \boldsymbol{\theta}$ *satisfies*

$$\begin{aligned}&\lim_{t\to\infty} \mathrm{E}[\|\tilde{\boldsymbol{\theta}}(t)\|^2] \\ &\leqslant \lim_{t\to\infty} \frac{n^2N^3(N+1)^2\beta^2\sigma_v^2[n\beta(t+N-1)+1]}{\alpha^2[n\alpha(t-2N+2)+1]^2} = 0.\end{aligned}$$

3.2. The convergence of the MISG algorithm

Here we establish the convergence of the MISG algorithm.

Theorem 2. *For the system in* (1) *and the algorithm in* (4)–(8), *suppose that the conditions of Theorem* 1 *hold and take the innovation length* $p = N$ *for simplifying the proof. Then the parameter estimation error given by the MISG algorithm satisfies*

$$\lim_{t\to\infty} \mathrm{E}[\|\tilde{\boldsymbol{\theta}}(t)\|^2] \leqslant \lim_{t\to\infty} \frac{N\beta\sigma_v^2[n\beta(t-N+1)+1]}{\alpha[n\alpha(t-N+1)+1]^2} = 0.$$

From Theorems 1 and 2, we can draw the following conclusions:

- For time-invariant stochastic systems, the parameter estimation errors $\|\tilde{\boldsymbol{\theta}}(t)\|^2$ by the SG and MISG algorithms converge to zero at the rates of C_1/t and C_2/t (C_1 and C_2 can be found in the proofs of Theorems 1 and 2), respectively, but the MISG algorithm has higher parameter estimation accuracy than the SG algorithm because the scale factor $C_2 \ll C_1$, see Example 1.
- With the help of the parameter estimation error upper bounds in (A.6) and (A.8), we can find answers to the following important question in system identification:
  - Given a small and positive ε and using the SG and MISG identification algorithms, how large a data set one needs to use to guarantee that the parameter estimation errors are less than ε, i.e., $\mathrm{E}[\|\hat{\boldsymbol{\theta}}(t) - \boldsymbol{\theta}\|^2] \leqslant \varepsilon$?

Let L_e denote the data length. According to (A.6) and (A.8), the data length must satisfy

$$L_e \geqslant \frac{C_1}{\varepsilon} = \frac{nN^3(N+1)^2\beta^3\sigma_v^2}{\alpha^4\varepsilon}$$

for the SG algorithm and

$$L_e \geqslant \frac{C_2}{\varepsilon} = \frac{N\beta^2\sigma_v^2}{n\alpha^3\varepsilon}$$

for the MISG algorithm. Here, according to Condition (A1), we can compute the quantities α and β and find N. Notice that n is the number of the system parameters in $\boldsymbol{\theta}$ and known, the noise variance σ_v^2 is replaced by its estimate $\hat{\sigma}_v^2$, see (17).

4. The MISG algorithm with a forgetting factor

Note that the SG and MISG algorithms have no ability to track time-varying parameters because the algorithm gains, $\boldsymbol{\Phi}(p, t)/r(t)$ and $\boldsymbol{\varphi}(t)/r(t)$, approach zero as t increases. In order to improve the tracking performance of the SG and MISG algorithms, we introduce a forgetting factor in the SG and MISG algorithms to get the SG algorithm with a forgetting factor [forgetting gradient (FG) algorithm for short] and the MISG algorithm with a forgetting factor [multi-innovation forgetting gradient (MIFG) algorithm for short]. The following is to discuss these two algorithms.

Consider the time-varying systems described by a linear regression model

$$y(t) = \boldsymbol{\varphi}^{\mathrm{T}}(t)\boldsymbol{\theta}(t) + v(t), \tag{9}$$

where the definitions of variables $y(t) \in \mathbb{R}^1$, $\boldsymbol{\varphi}(t) \in \mathbb{R}^n$ and $v(t) \in \mathbb{R}^1$ are the same as before, $\boldsymbol{\theta}(t) \in \mathbb{R}^n$ is the time-varying parameter vector to be identified.

The FG algorithm of identifying the time-varying parameter vector $\boldsymbol{\theta}(t)$ may be expressed as

$$\hat{\boldsymbol{\theta}}(t) = \hat{\boldsymbol{\theta}}(t-1) + \frac{\boldsymbol{\varphi}(t)}{r(t)}[y(t) - \boldsymbol{\varphi}^{\mathrm{T}}(t)\boldsymbol{\theta}(t-1)], \tag{10}$$

$$r(t) = \lambda r(t-1) + \|\boldsymbol{\varphi}(t)\|^2, \quad 0 < \lambda < 1, \quad r(0) > 0. \tag{11}$$

The MIFG algorithm of estimating $\boldsymbol{\theta}(t)$ may be expressed as

$$\hat{\boldsymbol{\theta}}(t) = \hat{\boldsymbol{\theta}}(t-1) + \frac{\boldsymbol{\Phi}(p,t)}{r(t)}[\boldsymbol{Y}(p,t) - \boldsymbol{\Phi}^{\mathrm{T}}(p,t)\hat{\boldsymbol{\theta}}(t-1)], \tag{12}$$

$$r(t) = \lambda r(t-1) + \|\boldsymbol{\varphi}(t)\|^2, \quad 0 < \lambda < 1, \quad r(0) > 0, \tag{13}$$

$$\boldsymbol{\Phi}(p,t) = [\boldsymbol{\varphi}(t), \boldsymbol{\varphi}(t-1), \ldots, \boldsymbol{\varphi}(t-p+1)] \in \mathbb{R}^{n\times p}, \tag{14}$$

$$\boldsymbol{Y}(p,t) = [y(t), y(t-1), \ldots, y(t-p+1)]^{\mathrm{T}} \in \mathbb{R}^p. \tag{15}$$

Obviously, when $p = 1$, we have MIFG = FG.

In engineering, the parameter estimation accuracy, e.g., measured by $\delta_a := \|\hat{\boldsymbol{\theta}}(t) - \boldsymbol{\theta}(t)\|^2$, is important. However, since the true parameter vector $\boldsymbol{\theta}(t)$ is unknown and is to be identified, the parameter estimation error δ_a is impossible to compute even if we obtain the estimation $\hat{\boldsymbol{\theta}}(t)$ by some identification algorithm. Therefore, we present the parameter estimation error upper bound to evaluate the parameter estimation accuracy indirectly (Ding & Chen, 2005a). The following lemma and theorem establish the parameter estimation error upper bound of the MIFG algorithm. The parameter estimation error upper bound of the FG algorithm can be done in a similar way as in Lemma 3 and Theorem 1, and hence is omitted here.

Lemma 4. *For the system in* (9) *and the MIFG algorithm in* (12)–(15), *assume that Condition* (A1) *holds, the information vector* $\boldsymbol{\varphi}(t)$ *has a lower bound like* $\|\boldsymbol{\varphi}(t)\|^2 \geqslant \alpha > 0$, *and* $r(0)$ *is chosen to satisfy*

$$\frac{\alpha}{1-\lambda} \leqslant r(0) \leqslant \frac{nN\beta}{1-\lambda}. \tag{16}$$

Then, for all $t > 0$, $r(t)$ *in* (13) *satisfies*

$$\frac{\alpha}{1-\lambda} \leqslant r(t) \leqslant \frac{nN\beta}{1-\lambda} \quad a.s., \ 0 < \lambda < 1.$$

Theorem 3. *For the system in* (9) *and the MIFG algorithm in* (12)–(15), *assume that* (A1) *holds,* $r(0)$ *is chosen by* (16), *the observation noise* $\{v(t)\}$ *and the parameter changing rate* $\{\boldsymbol{w}(t) := \boldsymbol{\theta}(t) - \boldsymbol{\theta}(t-1)\}$ *are stochastic sequences with zero mean, and the sequences* $\{v(t)\}$ *and* $\{\boldsymbol{w}(t)\}$ *satisfy*

(A3) $\mathrm{E}[v(t)] = 0, \quad \mathrm{E}[\boldsymbol{w}(t)] = \boldsymbol{0}$,

(A4) $\mathrm{E}[v^2(t)] \leqslant \sigma_v^2 < \infty, \quad \mathrm{E}[\|\boldsymbol{w}(t)\|^2] \leqslant \sigma_w^2 < \infty$.

Let the innovation length $p = N$ *and* $\mathrm{E}[\|\hat{\boldsymbol{\theta}}(0) - \boldsymbol{\theta}(0)\|^2] = \delta_0 < \infty$. *Then the parameter estimation error by the MIFG algorithm is mean square bounded*:

$$\begin{aligned}\mathrm{E}[\|\hat{\boldsymbol{\theta}}(t) - \boldsymbol{\theta}(t)\|^2] &\leqslant [\sqrt{1-\rho}]^t \delta_0 + \frac{3}{(1-\sqrt{1-\rho})^2} \\ &\quad \times \left[\frac{N^4\beta^2(1-\lambda)^2\sigma_w^2}{2\alpha^2} + \frac{N^2\beta(1-\lambda)^2\sigma_v^2}{\alpha^2} + \sigma_w^2\right] \\ &=: f_u(\lambda, t),\end{aligned}$$

where

$$0 < \rho := \frac{\alpha(1-\lambda)}{n\beta} < 1.$$

For large t, $0 < \lambda < 1$, $0 < \rho < 1$ and finite δ_0, since $[\sqrt{1-\rho}]^t\delta_0$ is very small and approaches zero as t goes to infinity, it can be neglected. Thus, we have an approximate upper bound:

$$\begin{aligned}f_u(\lambda, t) &\approx \frac{3}{(1-\sqrt{1-\rho})^2} \\ &\quad \times \left[\frac{N^4\beta^2(1-\lambda)^2\sigma_w^2}{2\alpha^2} + \frac{N^2\beta(1-\lambda)^2\sigma_v^2}{\alpha^2} + \sigma_w^2\right] \\ &= f(a_0, \lambda) =: f(\lambda).\end{aligned}$$

By experiments, we collect the measurement input–output data with data length $t = L_e \gg n$ to form the information vector $\boldsymbol{\varphi}(t)$. According to Condition (A1), we can compute the quantities α and β and find N. Notice that n is the number of the system parameters in $\boldsymbol{\theta}(t)$ and known, but the noise variances σ_v^2 and σ_w^2 are unknown. In order to obtain the estimation error upper bound $f(\lambda)$, we can compute the estimates $\hat{\sigma}_v^2$ and $\hat{\sigma}_w^2$ of σ_v^2 and σ_w^2 by

$$\begin{aligned}\hat{\sigma}_v^2 &= \frac{1}{L_e}\sum_{t=1}^{L_e}[y(t) - \boldsymbol{\varphi}^{\mathrm{T}}(t)\hat{\boldsymbol{\theta}}(L_e)]^2, \\ \hat{\sigma}_w^2 &= \frac{1}{L_e}\sum_{t=1}^{L_e}\|\hat{\boldsymbol{\theta}}(t) - \hat{\boldsymbol{\theta}}(t-1)\|^2.\end{aligned} \tag{17}$$

In practice, the true variances of noises are unavailable, so a way is to simply substitute the variances with their estimates to compute the error upper bound $f_u(\lambda, t)$.

From the above and Theorem 3, we can reach the following conclusions and corollaries:

- There exists a best forgetting factor λ such that the error upper bound $f(\lambda)$ is minimum. In other words, we may obtain the minimum estimation error upper bound $f(\lambda)$ by choosing an appropriate forgetting factor. In fact, letting

$$\frac{\mathrm{d}f(\lambda)}{\mathrm{d}\lambda} = 0$$

leads to the equation:

$$\frac{\alpha^2(1-\lambda)^2}{(n\beta)^2}-\frac{2\alpha(1-\lambda)}{n\beta}+2\sqrt{1-\frac{\alpha(1-\lambda)}{n\beta}}\,\frac{\alpha(1-\lambda)}{n\beta}-\frac{k_2}{k_1}=0, \tag{18}$$

where

$$k_1:=\frac{N^4\beta^4n^2\sigma_w^2}{2\alpha^4}+\frac{N^2\beta^3n^2\sigma_v^2}{\alpha^4},\quad k_2:=\sigma_w^2.$$

The equation in (18) has four solutions, and the solution $\lambda{=}\lambda_0$ which makes $f(\lambda)=\min$ is the best forgetting factor, and the corresponding minimum estimation error upper bound is $f(\lambda_0)$. This provides the theoretical guide of choosing the forgetting factor.

- A small noise variances σ_v^2 and/or small parameter changing rate σ_w^2 result in small estimation error upper bound.
- Large α and small β will generate a small estimation error upper bound. In other words, if α and β are closer, then the $\boldsymbol{\varphi}(t)$ are more stationary, and thus the stationarity of the input–output data can improve the parameter estimation accuracy.
- A small N that makes (A.1) hold can reduce the estimation error upper bound. That is, the sufficient richness of data can give good parameter estimation.

Corollary 1. *For a time-invariant deterministic system*

$$y(t)=\varphi^{\mathrm{T}}(t)\boldsymbol{\theta}$$

the estimation error converges to zero at exponentially fast rate, i.e.,

$$\mathrm{E}[\|\hat{\boldsymbol{\theta}}(t)-\boldsymbol{\theta}\|^2]\leqslant\left[1-\frac{\alpha(1-\lambda)}{n\beta}\right]^{2t}\delta_0\to 0,\quad t\geqslant N.$$

Corollary 2. *For a time-invariant stochastic system*

$$y(t)=\varphi^{\mathrm{T}}(t)\boldsymbol{\theta}+v(t),$$

we have

$$\begin{aligned}\lim_{t\to\infty}\mathrm{E}[\|\tilde{\boldsymbol{\theta}}(t)\|^2]&\leqslant\lim_{t\to\infty}[\sqrt{1-\rho}\,]^t\delta_0+\frac{1}{(1-\sqrt{1-\rho}\,)^2}\,\frac{N^2\beta(1-\lambda)^2\sigma_v^2}{\alpha^2}\\&=\left[1-\sqrt{1-\frac{\alpha(1-\lambda)}{n\beta}}\,\right]^{-2}\frac{N^2\beta(1-\lambda)^2\sigma_v^2}{\alpha^2}\\&=:f_1(\lambda).\end{aligned}$$

Thus, for time-invariant stochastic systems, the MIFG algorithm $(0<\lambda<1)$ *gives only a finite parameter estimation error upper bound* $f_1(\lambda)$*, but the MISG algorithm (take* $\lambda=1$ *in the MIFG algorithm) can give a consistent parameter estimation* $(\hat{\boldsymbol{\theta}}(t)\to\boldsymbol{\theta})$*. (See the preceding section.)*

Corollary 3. *For a time-varying deterministic system*

$$y(t)=\boldsymbol{\varphi}^{\mathrm{T}}(t)\boldsymbol{\theta}(t),$$

we have

$$\begin{aligned}\lim_{t\to\infty}\mathrm{E}[\|\tilde{\boldsymbol{\theta}}(t)\|^2]&\leqslant[\sqrt{1-\rho}\,]^t\delta_0+\frac{2}{(1-\sqrt{1-\rho}\,)^2}\left[\frac{N^4\beta^2(1-\lambda)^2\sigma_w^2}{2\alpha^2}+\sigma_w^2\right]\\&\leqslant 2\left[1-\sqrt{1-\frac{\alpha(1-\lambda)}{n\beta}}\,\right]^{-2}\frac{(N^4+1)\beta^2\sigma_w^2}{2\alpha^2}\\&=:f_2(\lambda).\end{aligned}$$

As $t\to\infty$*, the estimation error upper bound also approaches a constant* $f_2(\lambda)$*.*

Corollary 4. *For the time-varying system in* (9) *and the MIFG algorithm in* (12)–(15)*, we take a variable forgetting factor* $\lambda(t)$ *instead of the constant* λ *in* (13)*, and assume that the conditions of Theorem* 3 *hold. Then for* $0<\lambda_m\leqslant\lambda(t)\leqslant\lambda_M<1$*, we have*

$$\begin{aligned}\mathrm{E}[\|\tilde{\boldsymbol{\theta}}(t)\|^2]&\leqslant\left[\sqrt{1-\frac{\alpha(1-\lambda_M)}{n\beta}}\,\right]^t\delta_0+3\left[1-\sqrt{1-\frac{\alpha(1-\lambda_M)}{n\beta}}\,\right]^{-2}\\&\quad\times\left[\frac{N^4\beta^2(1-\lambda_m)^2\sigma_w^2}{2\alpha^2}+\frac{N^2\beta(1-\lambda_m)^2\sigma_v^2}{\alpha^2}+\sigma_w^2\right].\end{aligned}$$

Under the specific cases, the theorem and corollaries may be used to estimate estimation error upper bounds of the MIFG algorithm.

These results assume that the noises $\{v(t)\}$ and $\{w(t)\}$ have constant variances σ_v^2 and σ_w^2. From the proofs, if $\{v(t)\}$ and $\{w(t)\}$ are non-stationary and have time-varying variances $\sigma_v^2(t)$ and $\sigma_w^2(t)$ with upper bounds

(A4′) $\mathrm{E}[v^2(t)]=\sigma_v^2(t)\leqslant\sigma_v^2<\infty$,
$\mathrm{E}[\|\boldsymbol{w}(t)\|^2]=\sigma_w^2(t)\leqslant\sigma_w^2<\infty$.

the results of the above theorems and corollaries still hold.

5. The multi-variable version of the multi-innovation algorithm

Consider a multi-input and multi-output (MIMO) system

$$\boldsymbol{A}(z)\boldsymbol{y}(t)=\boldsymbol{B}(z)\boldsymbol{u}(t)+\boldsymbol{v}(t), \tag{19}$$

where $\boldsymbol{u}(t)\in\mathbb{R}^r$ is the system input vector, $\boldsymbol{y}(t)\in\mathbb{R}^m$ the system output vector, $\boldsymbol{v}(t)\in\mathbb{R}^m$ a stochastic noise vector, $\boldsymbol{A}(z)$ and $\boldsymbol{B}(z)$ are polynomial matrices in the unit backward shift operator z^{-1} $[z^{-1}\boldsymbol{y}(t)=\boldsymbol{y}(t-1)]$, and

$$\begin{aligned}\boldsymbol{A}(z)&=\boldsymbol{I}+\boldsymbol{A}_1z^{-1}+\boldsymbol{A}_2z^{-2}+\cdots+\boldsymbol{A}_{n_a}z^{-n_a},\\\boldsymbol{B}(z)&=\boldsymbol{B}_1z^{-1}+\boldsymbol{B}_2z^{-2}+\cdots+\boldsymbol{B}_{n_b}z^{-n_b}.\end{aligned}$$

Table 1
Comparisons of the SG and MISG algorithms for SISO and MIMO systems

Systems	SISO system (1) $y(t)=\boldsymbol{\varphi}^{\mathrm{T}}(t)\boldsymbol{\theta}+v(t)$ $y(t)\in\mathbb{R}^1$, $\boldsymbol{\theta}\in\mathbb{R}^n$ $\boldsymbol{\varphi}(t)\in\mathbb{R}^n$, $v(t)\in\mathbb{R}^1$	MIMO system (20) $\boldsymbol{y}(t)=\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\varphi}(t)+\boldsymbol{v}(t)$ $\boldsymbol{y}(t)\in\mathbb{R}^m$, $\boldsymbol{\theta}\in\mathbb{R}^{n\times m}$ $\boldsymbol{\varphi}(t)\in\mathbb{R}^n$, $\boldsymbol{v}(t)\in\mathbb{R}^m$
SG	Algorithm (2)–(3) Gain vector: $\boldsymbol{L}(t)=\frac{\boldsymbol{\varphi}(t)}{r(t)}\in\mathbb{R}^n$ Single innovation: $e(t)=y(t)-\boldsymbol{\varphi}^{\mathrm{T}}(t)\hat{\boldsymbol{\theta}}(t)\in\mathbb{R}^1$	Algorithm (21)–(24) Gain vector: $\boldsymbol{L}(t)=\frac{\boldsymbol{\varphi}(t)}{r(t)}\in\mathbb{R}^n$ Single innovation row vector: $\boldsymbol{E}(t)=\boldsymbol{y}^{\mathrm{T}}(t)-\boldsymbol{\varphi}^{\mathrm{T}}(t)\hat{\boldsymbol{\theta}}(t-1)\in\mathbb{R}^{1\times m}$
MISG	Algorithm (4)–(8) Gain matrix: $\boldsymbol{L}(p,t)=\frac{\boldsymbol{\Phi}(p,t)}{r(t)}\in\mathbb{R}^{n\times p}$ Multi-innovation vector: $\boldsymbol{E}(p,t)\in\mathbb{R}^{p\times 1}$	Algorithm (25)–(30) Gain matrix: $\boldsymbol{L}(p,t)=\frac{\boldsymbol{\Phi}(p,t)}{r(t)}\in\mathbb{R}^{n\times p}$ Multi-innovation matrix: $\boldsymbol{E}(p,t)\in\mathbb{R}^{p\times m}$

Define the parameter matrix $\boldsymbol{\theta}$ and information vector $\boldsymbol{\varphi}(t)$ as

$$\begin{aligned}\boldsymbol{\theta}^{\mathrm{T}} &:= [\boldsymbol{A}_1,\boldsymbol{A}_2,\ldots,\boldsymbol{A}_{n_a},\boldsymbol{B}_1,\boldsymbol{B}_2,\ldots,\boldsymbol{B}_{n_b}]\in\mathbb{R}^{m\times n},\\ \boldsymbol{\varphi}(t) &:= [-\boldsymbol{y}^{\mathrm{T}}(t-1),-\boldsymbol{y}^{\mathrm{T}}(t-2),\ldots,-\boldsymbol{y}^{\mathrm{T}}(t-n_a),\\ &\quad \boldsymbol{u}^{\mathrm{T}}(t-1),\boldsymbol{u}^{\mathrm{T}}(t-2),\ldots,\boldsymbol{u}^{\mathrm{T}}(t-n_b)]^{\mathrm{T}}\in\mathbb{R}^n,\\ n &:= mn_a+rn_b.\end{aligned}$$

Then (19) can be rewritten as

$$\boldsymbol{y}(t)=\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\varphi}(t)+\boldsymbol{v}(t),$$

or

$$\boldsymbol{y}^{\mathrm{T}}(t)=\boldsymbol{\varphi}^{\mathrm{T}}(t)\boldsymbol{\theta}+\boldsymbol{v}^{\mathrm{T}}(t). \tag{20}$$

The SG algorithm of estimating the parameter matrix $\boldsymbol{\theta}$ in (20) is expressed as

$$\hat{\boldsymbol{\theta}}(t)=\hat{\boldsymbol{\theta}}(t-1)+\boldsymbol{L}(t)\boldsymbol{E}(t), \tag{21}$$

$$\boldsymbol{L}(t)=\frac{\boldsymbol{\varphi}(t)}{r(t)}\in\mathbb{R}^n \quad \text{(gain vector)}, \tag{22}$$

$$\boldsymbol{E}(t)=\boldsymbol{y}^{\mathrm{T}}(t)-\boldsymbol{\varphi}^{\mathrm{T}}(t)\hat{\boldsymbol{\theta}}(t-1)\in\mathbb{R}^{1\times m}, \tag{23}$$

$$r(t)=r(t-1)+\|\boldsymbol{\varphi}(t)\|^2, \quad r(0)=1. \tag{24}$$

Here, $\boldsymbol{E}(t)\in\mathbb{R}^{1\times m}$ represents a single innovation row vector and each element of $\boldsymbol{E}(t)$ is a scalar innovation corresponding to each output, and thus $\boldsymbol{E}(t)$ is a single innovation vector. We extend this single innovation vector $\boldsymbol{E}(t)\in\mathbb{R}^m$ to a multi-innovation matrix $\boldsymbol{E}(p,t)\in\mathbb{R}^{p\times m}$ to get the MISG algorithm of estimating the parameter matrix $\boldsymbol{\theta}$ as follows:

$$\hat{\boldsymbol{\theta}}(t)=\hat{\boldsymbol{\theta}}(t-1)+\boldsymbol{L}(p,t)\boldsymbol{E}(p,t), \tag{25}$$

$$\boldsymbol{L}(p,t)=\frac{\boldsymbol{\Phi}(p,t)}{r(t)}\in\mathbb{R}^{n\times p} \quad \text{(gain matrix)}, \tag{26}$$

$$\boldsymbol{E}(p,t)=\boldsymbol{Y}(p,t)-\boldsymbol{\Phi}^{\mathrm{T}}(p,t)\hat{\boldsymbol{\theta}}(t-1)\in\mathbb{R}^{p\times m}, \tag{27}$$

$$r(t)=r(t-1)+\|\boldsymbol{\varphi}(t)\|^2, \quad r(0)=1, \tag{28}$$

$$\boldsymbol{\Phi}(p,t)=[\boldsymbol{\varphi}(t),\boldsymbol{\varphi}(t-1),\ldots,\varphi(t-p+1)]\in\mathbb{R}^{n\times p}, \tag{29}$$

$$\boldsymbol{Y}(p,t)=[\boldsymbol{y}(t),\boldsymbol{y}(t-1),\ldots,\boldsymbol{y}(t-p+1)]^{\mathrm{T}}\in\mathbb{R}^{p\times m}. \tag{30}$$

Here, $\boldsymbol{E}(p,t)\in\mathbb{R}^{p\times m}$ denotes a multi-innovation matrix.

From here, we can see that the multi-innovation identification algorithms do not imply the identification algorithms of multi-variable systems. For convenience, Table 1 states the differences of the variables of the single innovation SG and MISG algorithms for single-input, single-output (SISO) and MIMO systems.

6. Examples

Several illustrative examples are given in this section. The first is a time-invariant system; the second and third are time-varying systems.

Example 1. Consider a time-invariant system:

$$\begin{aligned}&A(z)y(t)=B(z)u(t)+v(t),\\ &A(z)=1+a_1z^{-1}+a_2z^{-2}=1-1.60z^{-1}+0.80z^{-2},\\ &B(z)=b_1z^{-1}+b_2z^{-2}=0.309z^{-1}+0.529z^{-2},\end{aligned}$$

where $u(t)$ and $y(t)$ are the system input and output, respectively. Define

$$\begin{aligned}\boldsymbol{\theta} &:= [a_1,a_2,b_1,b_2]^{\mathrm{T}}=[-1.60,0.80,0.309,0.529]^{\mathrm{T}},\\ \boldsymbol{\varphi}(t) &:= [-y(t-1),-y(t-2),u(t-1),u(t-2)]^{\mathrm{T}},\end{aligned}$$

then this example system can be written as the form of (1). In simulation, the inputs $\{u(t)\}$ is taken as an uncorrelated persistent excitation signal sequence with zero mean and unit variance $\sigma_u^2=1.00^2$, and $\{v(t)\}$ as a white noise sequence with zero mean and variance $\sigma_v^2=0.50^2$. Applying the SG algorithm and MISG algorithm to estimate the parameters of this system, the parameter estimates and their errors with different innovation length are shown in Tables 2 and 3, and the parameter estimation errors $\delta=\|\hat{\boldsymbol{\theta}}(t)-\boldsymbol{\theta}\|/\|\boldsymbol{\theta}\|$ vs. t are shown in Fig. 1 with $p=1,2,3,5,8$, where the noise-to-signal ratio of the system is $\delta_{\mathrm{ns}}=61.27\%$. For comparison, Fig. 1 also plots the estimation error curve of the RLS algorithm.

From Tables 2 and 3 and Fig. 1, we can see that the MISG estimates with $p\geqslant 2$ have higher accuracy than the standard SG estimates, the parameter estimation errors by the MISG algorithm become smaller and smaller as the innovation length p increases and go to zero with t increasing. This confirms

Table 2
The SG estimates and errors ($\sigma_v^2 = 0.50^2$)

t	a_1	a_2	b_1	b_2	δ (%)
100	−0.74698	−0.05187	0.04739	0.03718	70.23421
200	−0.77582	−0.01730	0.05240	0.05986	67.58660
300	−0.77784	0.00672	0.05255	0.06973	66.51767
500	−0.81194	0.02151	0.05996	0.08823	64.41216
1000	−0.84693	0.05143	0.06761	0.10922	61.71967
1500	−0.87014	0.07191	0.07450	0.12398	59.87700
2000	−0.88303	0.08312	0.07773	0.13183	58.87245
2500	−0.89483	0.09440	0.08043	0.13829	57.93481
3000	−0.90194	0.10179	0.08218	0.14245	57.34344
True values	−1.60000	0.80000	0.30900	0.52900	

Table 3
The MISG estimates and errors ($\sigma_v^2 = 0.50^2$)

p	t	a_1	a_2	b_1	b_2	δ (%)
2	100	−1.05483	0.24582	−0.03541	0.30286	46.53018
	200	−1.08585	0.28946	−0.01710	0.32624	43.36857
	300	−1.08736	0.31692	−0.01046	0.33310	42.19645
	500	−1.12761	0.33375	0.00666	0.35226	39.68895
	1000	−1.16595	0.36993	0.02515	0.37094	36.59920
	1500	−1.19071	0.39327	0.03806	0.38482	34.56431
	2000	−1.20448	0.40542	0.04512	0.39192	33.47091
	2500	−1.21715	0.41815	0.05024	0.39790	32.45212
	3000	−1.22436	0.42662	0.05376	0.40157	31.81539
5	100	−1.38165	0.57678	0.16809	0.49441	18.20973
	200	−1.40091	0.61985	0.19247	0.50508	15.53127
	300	−1.40872	0.63526	0.19661	0.50252	14.68105
	500	−1.43787	0.64842	0.21220	0.51504	12.82711
	1000	−1.46838	0.67465	0.22574	0.52138	10.58043
	1500	−1.48441	0.68834	0.23358	0.52745	9.38914
	2000	−1.49450	0.69487	0.23990	0.53031	8.68319
	2500	−1.50289	0.70441	0.24262	0.53326	8.01916
	3000	−1.50614	0.71168	0.24476	0.53508	7.62238
8	100	−1.56479	0.75915	0.20129	0.43085	8.21780
	200	−1.55085	0.78023	0.24047	0.44926	6.22680
	300	−1.55066	0.77909	0.24568	0.45013	6.05334
	500	−1.56505	0.77185	0.26131	0.47217	4.58559
	1000	−1.57593	0.78208	0.27077	0.48353	3.51971
	1500	−1.58003	0.78314	0.27492	0.49139	3.01923
	2000	−1.58537	0.78291	0.28158	0.49594	2.56415
	2500	−1.58857	0.78864	0.28268	0.50027	2.22991
	3000	−1.58802	0.79372	0.28397	0.50339	2.02449
True values		−1.60000	0.80000	0.30900	0.52900	

Theorems 1 and 2. Also, as the innovation length p increases, the MISG estimates are approaching the RLS estimates.

Example 2. Consider a time-varying system with jump changing parameters:

$$y(t) + a(t)y(t-1) = b(t)u(t-1) + v(t),$$

where $a(t) = 0.80$ and

$$b(t) = \begin{cases} 1.2, & 0 \leqslant t \leqslant 700,\ 1401 \leqslant t < 2100, \ldots, \\ 1.6, & 701 \leqslant t \leqslant 1400,\ 2101 \leqslant t < 2800, \ldots. \end{cases}$$

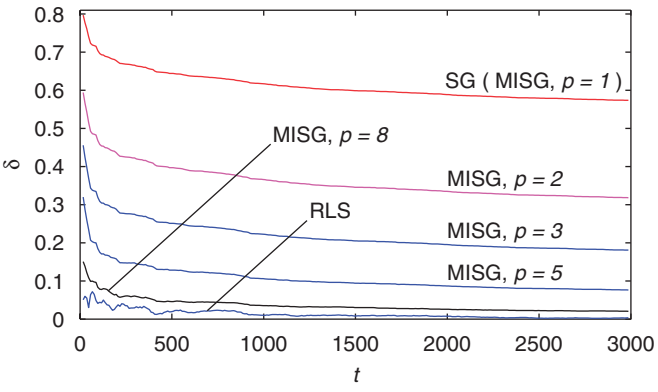

Fig. 1. The parameter estimation errors δ vs. t ($\sigma^2 = 0.50^2$).

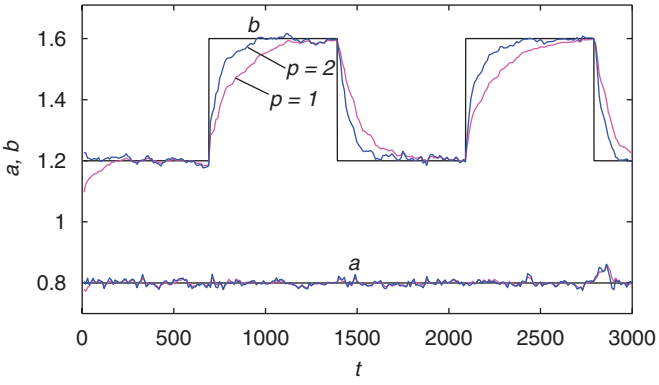

Fig. 2. The parameter estimates vs. t ($p = 2$).

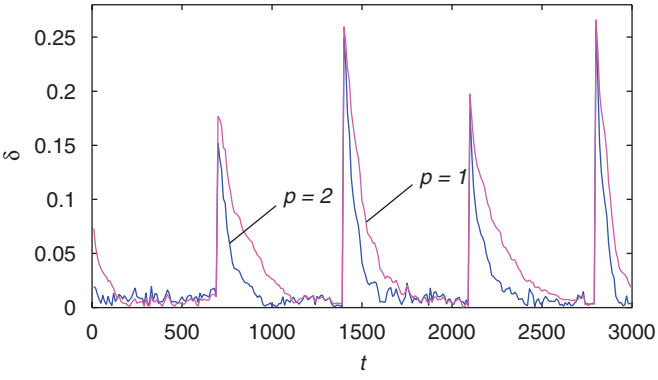

Fig. 3. The parameter estimation errors δ vs. t ($p = 2$).

The simulation conditions are as before, the noise variance here is $\sigma_v^2 = 0.10^2$. Applying the FG algorithm (take $p = 1$ in the MIFG algorithm) and MIFG algorithm with $\lambda = 0.95$ to estimate the parameters of this system, the parameter estimates and estimation error curves vs. t are shown in Figs. 2–5 for different innovation length $p = 2$ and $p = 3$.

From Figs. 2–5, it is clear that the MIFG algorithm with $p \geqslant 2$ has a faster convergence rate than the FG algorithm or the MIFG algorithm with $p = 1$ but has large estimation error variance. Therefore, a compromise/tradeoff is to change p into

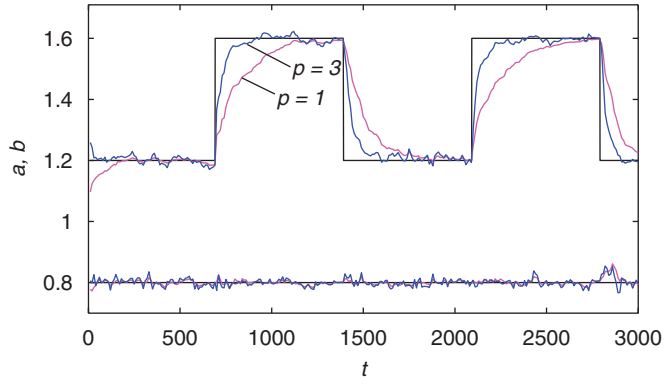

Fig. 4. The parameter estimates vs. t ($p = 3$).

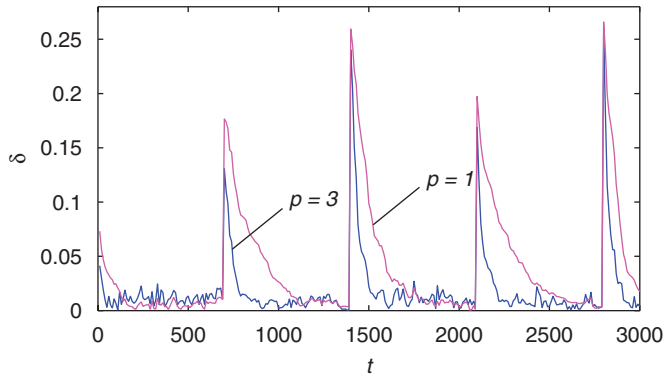

Fig. 5. The parameter estimation errors δ vs. t ($p = 3$).

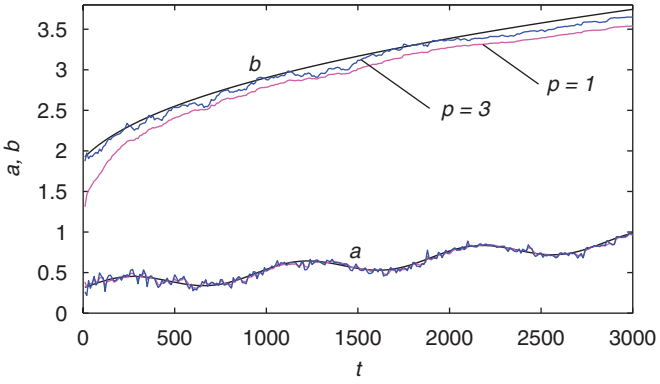

Fig. 6. The parameter estimates vs. t ($p = 3$).

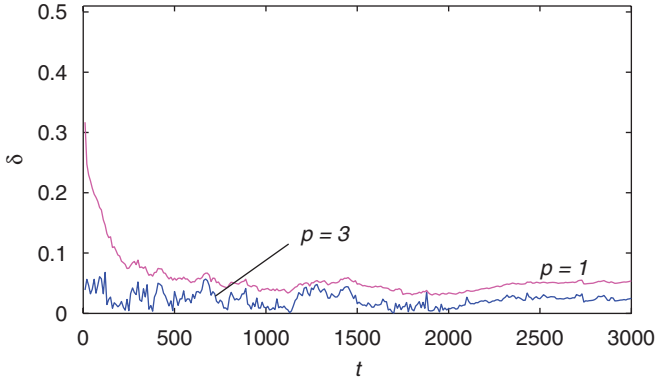

Fig. 7. The parameter estimation errors δ vs. t ($p = 3$).

a smaller value when the parameter estimation changing rate is small.

Example 3. Consider a time-varying parameter system:

$$y(t) + a(t)y(t-1) = b(t)u(t-1) + v(t),$$

where

$$a(t) = 0.55 + 0.00025t + 0.1\sin(0.00225\pi t),$$

$$b(t) = 1.25 + 0.1\sqrt[5]{(t+100)^2}.$$

The simulation conditions are as before, the noise variance here is $\sigma_v^2 = 0.50^2$. Applying the FG algorithm and MIFG algorithm with $\lambda = 0.95$ to estimate the parameters of this system, the parameter estimates and estimation error curves vs. t are shown in Figs. 6 and 7 for different innovation length $p = 3$.

The simulation results in Figs. 6 and 7 all show the advantages of the proposed MIFG algorithm.

7. Conclusions

Extending the concept of the innovation modification, we presented several new algorithms, including the MISG algorithm, the MIFG algorithm. The algorithms developed have faster convergence rates and can improve the parameter estimation accuracy. The simulation results confirm the theoretical findings.

Acknowledgments

This work was supported by the Natural Sciences and Engineering Research Council of Canada and the National Natural Science Foundation of China (Nos. 60574051, 60528007).

Appendix A. Proofs

In this appendix we shall prove the main results of this paper by formulating a martingale process and using stochastic process theory and the martingale convergence theorem (Ding & Chen, 2004, 2005b–e).

Proof of Lemma 2. Taking the trace of Condition (A1), it is easy to get

$$nN\alpha \leqslant \sum_{i=0}^{N-1} \|\boldsymbol{\varphi}(t+i)\|^2 \leqslant nN\beta =: \delta_1 \quad a.s., \tag{A.1}$$

and $\|\varphi(t)\|^2 \leqslant \delta_1$, *a.s.* Let $\lfloor x \rfloor$ denote the greatest integer not greater than x. From (3), successive substitution gives

$$\begin{aligned} r(t) &= r(t-1) + \|\varphi(t)\|^2 = \sum_{j=1}^{t} \|\varphi(j)\|^2 + r(0) \\ &\leqslant \sum_{j=0}^{\lfloor (t-1)/N \rfloor} \sum_{i=1}^{N} \|\varphi(jN+i)\|^2 + r(0) \\ &\leqslant \sum_{j=0}^{\lfloor (t-1)/N \rfloor} \delta_1 + r(0) \leqslant \left(\left\lfloor \frac{t-1}{N} \right\rfloor + 1 \right) \delta_1 + 1 \\ &\leqslant n\beta(t+N-1) + 1 \quad a.s., \end{aligned}$$

$$\begin{aligned} r(t) &\geqslant \sum_{j=0}^{\lfloor t/N \rfloor - 1} \sum_{i=1}^{N} \|\varphi(jN+i)\|^2 + r(0) \\ &\geqslant \sum_{j=0}^{\lfloor t/N \rfloor - 1} nN\alpha + r(0) \geqslant \left\lfloor \frac{t}{N} \right\rfloor nN\alpha + 1 \\ &\geqslant n\alpha(t-N+1) + 1 \quad a.s. \end{aligned}$$

This proves Lemma 2. □

Proof of Lemma 3. Let v_0 be the unit eigenvector, of the matrix $\boldsymbol{L}^{\mathrm{T}}(t+N,t)\boldsymbol{L}(t+N,t)$, corresponding to the greatest eigenvalue ρ_t, and form the difference equation (Chen & Guo, 1987),

$$\boldsymbol{x}_{i+1} = \left[\boldsymbol{I} - \frac{\varphi(i)\varphi^{\mathrm{T}}(i)}{r(i)} \right] \boldsymbol{x}_i = \boldsymbol{L}(i+1,i)\boldsymbol{x}_i, \quad \boldsymbol{x}_t = \boldsymbol{v}_0. \tag{A.2}$$

Using the properties of the transition matrix, $\boldsymbol{L}(t,i)\boldsymbol{L}(i,s) = \boldsymbol{L}(t,s)$, it follows that

$$\begin{aligned} &\boldsymbol{x}_{t+N} = \boldsymbol{L}(t+N,t)\boldsymbol{x}_t = \boldsymbol{L}(t+N,t)\boldsymbol{v}_0, \\ &\|\boldsymbol{x}_{t+N}\|^2 = \boldsymbol{v}_0^{\mathrm{T}}\boldsymbol{L}^{\mathrm{T}}(t+N,t)\boldsymbol{L}(t+N,t)\boldsymbol{v}_0 = \boldsymbol{v}_0^{\mathrm{T}}\rho_t\boldsymbol{v}_0 = \rho_t. \end{aligned}$$

Taking the norm of both sides of (A.2) gives

$$\begin{aligned} \boldsymbol{x}_{i+1}^{\mathrm{T}}\boldsymbol{x}_{i+1} &= \boldsymbol{x}_i^{\mathrm{T}} \left[\boldsymbol{I} - \frac{\varphi(i)\varphi^{\mathrm{T}}(i)}{r(i)} \right]^2 \boldsymbol{x}_i \\ &\leqslant \boldsymbol{x}_i^{\mathrm{T}} \left[\boldsymbol{I} - \frac{\varphi(i)\varphi^{\mathrm{T}}(i)}{r(i)} \right] \boldsymbol{x}_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_i - \frac{\|\varphi^{\mathrm{T}}(i)\boldsymbol{x}_i\|^2}{r(i)}. \end{aligned}$$

Thus

$$\frac{\|\varphi^{\mathrm{T}}(i)\boldsymbol{x}_i\|^2}{r(i)} \leqslant \|\boldsymbol{x}_i\|^2 - \|\boldsymbol{x}_{i+1}\|^2.$$

Replacing i by $t+i$ and summing for i from $i=0$ to $i=N-1$ yield

$$\sum_{i=0}^{N-1} \frac{\|\varphi^{\mathrm{T}}(t+i)\boldsymbol{x}_{t+i}\|^2}{r(t+i)} \leqslant \|\boldsymbol{x}_t\|^2 - \|\boldsymbol{x}_{t+N}\|^2 = 1 - \rho_t. \tag{A.3}$$

For any $i \in [0, N-1]$, using the formula $(\sum a_i b_i)^2 \leqslant (\sum a_i^2)(\sum b_i^2)$, from (A.2) and (A.3), we have

$$\begin{aligned} \|\boldsymbol{x}_{t+i} - \boldsymbol{v}_0\| &= \left\| \sum_{j=0}^{i-1} \frac{\varphi(t+j)\varphi^{\mathrm{T}}(t+j)}{r(t+j)} \boldsymbol{x}_{t+j} \right\| \\ &\leqslant \left[\sum_{j=0}^{i-1} \frac{\|\varphi(t+j)\|^2}{r(t+j)} \right]^{1/2} \left[\sum_{j=0}^{i-1} \frac{\|\varphi^{\mathrm{T}}(t+j)\boldsymbol{x}_{t+j}\|^2}{r(t+j)} \right]^{1/2} \\ &\leqslant \sqrt{i(1-\rho_t)} \leqslant \sqrt{N(1-\rho_t)}. \end{aligned} \tag{A.4}$$

Here, we have used $\|\varphi(t)\|^2/r(t) < 1$. Pre-multiplying (A.1) by $\boldsymbol{v}_0^{\mathrm{T}}$ and post-multiplying (A.1) by $\boldsymbol{v}_0$, and using Lemma 2, (A.3) and (A.4) and $\|\varphi(t)\|^2 \leqslant \delta_1$, *a.s.*, noting that $r(t)$ is non-decreasing, we have

$$\begin{aligned} \alpha N &\leqslant \boldsymbol{v}_0^{\mathrm{T}} \sum_{i=0}^{N-1} \varphi(t+i)\varphi^{\mathrm{T}}(t+i)\boldsymbol{v}_0 \\ &\leqslant \sqrt{r(t+N-1)}\, \boldsymbol{v}_0^{\mathrm{T}} \sum_{i=0}^{N-1} \frac{\varphi(t+i)\varphi^{\mathrm{T}}(t+i)}{\sqrt{r(t+i)}} \boldsymbol{v}_0 \\ &\leqslant \sqrt{r(t+N-1)} \left\| \sum_{i=0}^{N-1} \frac{\varphi(t+i)\varphi^{\mathrm{T}}(t+i)}{\sqrt{r(t+i)}} (\boldsymbol{v}_0 - \boldsymbol{x}_{t+i} + \boldsymbol{x}_{t+i}) \right\| \\ &\leqslant \sqrt{r(t+N-1)} \left[\left\| \sum_{i=0}^{N-1} \frac{\varphi(t+i)\varphi^{\mathrm{T}}(t+i)}{\sqrt{r(t+i)}} (\boldsymbol{v}_0 - \boldsymbol{x}_{t+i}) \right\| \right. \\ &\quad \left. + \left\| \sum_{i=0}^{N-1} \frac{\varphi(t+i)\varphi^{\mathrm{T}}(t+i)\boldsymbol{x}_{t+i}}{\sqrt{r(t+i)}} \right\| \right] \\ &\leqslant \sqrt{r(t+N-1)} \left\{ \sum_{i=0}^{N-1} \sqrt{\delta_1}\, \|\boldsymbol{x}_{t+i} - \boldsymbol{v}_0\| \right. \\ &\quad \left. + \left[\sum_{i=0}^{N-1} \|\varphi(t+i)\|^2 \right]^{1/2} \left[\sum_{i=0}^{N-1} \frac{\|\varphi^{\mathrm{T}}(t+i)\boldsymbol{x}_{t+i}\|^2}{r(t+i)} \right]^{1/2} \right\} \\ &\leqslant \sqrt{n\beta(t+2N-2)+1} \left[N\sqrt{\delta_1}\sqrt{N(1-\rho_t)} + \sqrt{N\delta_1}\sqrt{1-\rho_t} \right] \\ &= \sqrt{n\beta(t+2N-2)+1}\, (N+1)\sqrt{N\delta_1(1-\rho_t)} \quad a.s. \end{aligned}$$

Solving ρ_t leads to the conclusion of Lemma 3. □

Proof of Theorem 1. Define the parameter estimation error vector

$$\tilde{\boldsymbol{\theta}}(t) := \hat{\boldsymbol{\theta}}(t) - \boldsymbol{\theta}.$$

Using (2) and (1), we have

$$\begin{aligned}\tilde{\boldsymbol{\theta}}(t) &= \tilde{\boldsymbol{\theta}}(t-1) + \frac{\boldsymbol{\varphi}(t)}{r(t)}[-\boldsymbol{\varphi}^{\mathrm{T}}(t)\tilde{\boldsymbol{\theta}}(t-1) + v(t)] \\ &= \left[\boldsymbol{I} - \frac{\boldsymbol{\varphi}(t)\boldsymbol{\varphi}^{\mathrm{T}}(t)}{r(t)}\right]\tilde{\boldsymbol{\theta}}(t-1) + \frac{\boldsymbol{\varphi}(t)}{r(t)}v(t) \\ &= \boldsymbol{L}(t+1,t)\tilde{\boldsymbol{\theta}}(t-1) + \frac{\boldsymbol{\varphi}(t)}{r(t)}v(t) \\ &= \boldsymbol{L}(t+1,t-N+1)\tilde{\boldsymbol{\theta}}(t-N) \\ &\quad + \sum_{i=0}^{N-1}\boldsymbol{L}(t+1,t-i+1)\frac{\boldsymbol{\varphi}(t-i)}{r(t-i)}v(t-i).\end{aligned}$$

Taking the norm gives

$$\begin{aligned}\|\tilde{\boldsymbol{\theta}}(t)\|^2 &= \tilde{\boldsymbol{\theta}}^{\mathrm{T}}(t-N)\boldsymbol{L}^{\mathrm{T}}(t+1,t-N+1) \\ &\quad \times \boldsymbol{L}(t+1,t-N+1)\tilde{\boldsymbol{\theta}}(t-N) \\ &\quad + 2\tilde{\boldsymbol{\theta}}^{\mathrm{T}}(t-N)\boldsymbol{L}^{\mathrm{T}}(t+1,t-N+1) \\ &\quad \times \sum_{i=0}^{N-1}\boldsymbol{L}(t+1,t-i+1)\frac{\boldsymbol{\varphi}(t-i)}{r(t-i)}v(t-i) \\ &\quad + \left\|\sum_{i=0}^{N-1}\boldsymbol{L}(t+1,t-i+1)\frac{\boldsymbol{\varphi}(t-i)}{r(t-i)}v(t-i)\right\|^2 \\ &\leqslant \tilde{\boldsymbol{\theta}}^{\mathrm{T}}(t-N)\boldsymbol{L}^{\mathrm{T}}(t+1,t-N+1) \\ &\quad \times \boldsymbol{L}(t+1,t-N+1)\tilde{\boldsymbol{\theta}}(t-N) \\ &\quad + 2\tilde{\boldsymbol{\theta}}^{\mathrm{T}}(t-N)\boldsymbol{L}^{\mathrm{T}}(t+1,t-N+1) \\ &\quad \times \sum_{i=0}^{N-1}\boldsymbol{L}(t+1,t-i+1)\frac{\boldsymbol{\varphi}(t-i)}{r(t-i)}v(t-i) \\ &\quad + N\sum_{i=0}^{N-1}\left\|\boldsymbol{L}(t+1,t-i+1)\frac{\boldsymbol{\varphi}(t-i)}{r(t-i)}v(t-i)\right\|^2.\end{aligned} \tag{A.5}$$

For any $i \geqslant 1$, the greatest eigenvalue of $\boldsymbol{L}^{\mathrm{T}}(t+1,t-i+1)\boldsymbol{L}(t+1,t-i+1)$ is smaller than or equal to 1. Let $T(t) = \mathrm{E}[\|\tilde{\boldsymbol{\theta}}(t)\|^2]$. Taking the expectation of both sides of (A.5) and using (A.2) and Lemmas 2 and 3 lead to

$$\begin{aligned}T(t) &\leqslant \rho_{t-N+1}T(t-N) \\ &\quad + N\sum_{i=0}^{N-1}\mathrm{E}\left[\left\|\boldsymbol{L}(t+1,t-i+1)\frac{\boldsymbol{\varphi}(t-i)}{r(t-i)}v(t-i)\right\|^2\right] \\ &\leqslant \rho_{t-N+1}T(t-N) + N\sum_{i=0}^{N-1}\mathrm{E}\left[\left\|\frac{\boldsymbol{\varphi}(t-i)}{r(t-i)}v(t-i)\right\|^2\right] \\ &\leqslant \rho_{t-N+1}T(t-N) + N\sum_{i=0}^{N-1}\mathrm{E}\left[\frac{\|\boldsymbol{\varphi}(t-i)\|^2}{r^2(t-i)}v^2(t-i)\right] \\ &\leqslant \rho_{t-N+1}T(t-N) + N\sum_{i=0}^{N-1}\frac{\delta_1\sigma_v^2}{[n\alpha(t-N+1-i)+1]^2} \\ &\leqslant \left(1 - \frac{N\alpha^2}{(N+1)^2\delta_1[n\beta(t+N-1)+1]}\right)T(t-N) \\ &\quad + \frac{N^2\delta_1\sigma_v^2}{[n\alpha(t-2N+2)+1]^2}.\end{aligned}$$

Using Lemma 1, it is easy to get

$$\begin{aligned}&\lim_{t\to\infty}\mathrm{E}[\|\tilde{\boldsymbol{\theta}}(t)\|^2] \\ &\leqslant \lim_{t\to\infty}\frac{N^2\delta_1\sigma_v^2}{[n\alpha(t-2N+2)+1]^2} \\ &\quad \times \frac{(N+1)^2\delta_1[n\beta(t+N-1)+1]}{N\alpha^2} \\ &\leqslant \lim_{t\to\infty}\frac{n^2N^3(N+1)^2\beta^2\sigma_v^2[n\beta(t+N-1)+1]}{\alpha^2[n\alpha(t-2N+2)+1]^2} \\ &\sim \frac{nN^3(N+1)^2\beta^3\sigma_v^2}{\alpha^4}\frac{1}{t} =: C_1\frac{1}{t},\end{aligned} \tag{A.6}$$

where

$$C_1 := \frac{nN^3(N+1)^2\beta^3\sigma_v^2}{\alpha^4}.$$

The proof of Theorem 1 is completed. □

Proof of Theorem 2. Define the noise vector

$$\boldsymbol{V}(p,t) := [v(t), v(t-1), \ldots, v(t-p+1)] \in \mathbb{R}^p.$$

Subtracting both sides of (4) and using (5), (7), (8) and (1) yield

$$\begin{aligned}\tilde{\boldsymbol{\theta}}(t) &= \tilde{\boldsymbol{\theta}}(t-1) + \frac{\boldsymbol{\Phi}(p,t)}{r(t)}[-\boldsymbol{\Phi}^{\mathrm{T}}(p,t)\tilde{\boldsymbol{\theta}}(t-1) + \boldsymbol{V}(p,t)] \\ &= \left[\boldsymbol{I} - \frac{\boldsymbol{\Phi}(p,t)\boldsymbol{\Phi}^{\mathrm{T}}(p,t)}{r(t)}\right]\tilde{\boldsymbol{\theta}}(t-1) + \frac{\boldsymbol{\Phi}(p,t)\boldsymbol{V}(p,t)}{r(t)}.\end{aligned}$$

Taking the norm of both sides gives

$$\begin{aligned}\|\tilde{\boldsymbol{\theta}}(t)\|^2 &= \left\|\left[\boldsymbol{I} - \frac{\boldsymbol{\Phi}(p,t)\boldsymbol{\Phi}^{\mathrm{T}}(p,t)}{r(t)}\right]\tilde{\boldsymbol{\theta}}(t-1)\right\|^2 \\ &\quad + 2\tilde{\boldsymbol{\theta}}^{\mathrm{T}}(t-1)\left[\boldsymbol{I} - \frac{\boldsymbol{\Phi}(p,t)\boldsymbol{\Phi}^{\mathrm{T}}(p,t)}{r(t)}\right]\frac{\boldsymbol{\Phi}(p,t)\boldsymbol{V}(p,t)}{r(t)} \\ &\quad + \left\|\frac{\boldsymbol{\Phi}(p,t)\boldsymbol{V}(p,t)}{r(t)}\right\|^2 \\ &\leqslant \lambda_{\max}\left[\boldsymbol{I} - \frac{\boldsymbol{\Phi}(p,t)\boldsymbol{\Phi}^{\mathrm{T}}(p,t)}{r(t)}\right]\|\tilde{\boldsymbol{\theta}}(t-1)\|^2 \\ &\quad + 2\tilde{\boldsymbol{\theta}}^{\mathrm{T}}(t-1)\left[\boldsymbol{I} - \frac{\boldsymbol{\Phi}(p,t)\boldsymbol{\Phi}^{\mathrm{T}}(p,t)}{r(t)}\right]\frac{\boldsymbol{\Phi}(p,t)\boldsymbol{V}(p,t)}{r(t)} \\ &\quad + \frac{\|\boldsymbol{\Phi}(p,t)\boldsymbol{V}(p,t)\|^2}{r^2(t)}.\end{aligned} \tag{A.7}$$

Using Lemma 1 and noting $p = N$, from Condition (A.1), we have

$$\boldsymbol{I} - \frac{\boldsymbol{\Phi}(p,t)\boldsymbol{\Phi}^{\mathrm{T}}(p,t)}{r(t)} \leqslant \left[1 - \frac{N\alpha}{n\beta(t-N+1)+1}\right]\boldsymbol{I} \quad a.s.,$$

$$\begin{aligned}\mathrm{E}[\|\boldsymbol{\Phi}(p,t)\boldsymbol{V}(p,t)\|^2] &\leqslant \mathrm{E}\{\lambda_{\max}[\boldsymbol{\Phi}(p,t)\boldsymbol{\Phi}^{\mathrm{T}}(p,t)]\|\boldsymbol{V}(p,t)\|^2\} \\ &\leqslant p\beta\mathrm{E}[\|\boldsymbol{V}(p,t)\|^2] = p^2\beta\sigma_v^2 = N^2\beta\sigma_v^2.\end{aligned}$$

Thus, taking the expectation of both sides of (A.7) and using (A.2) give

$$\begin{aligned}&\mathrm{E}[\|\tilde{\boldsymbol{\theta}}(t)\|^2] \\ &\leqslant \left[1 - \frac{N\alpha}{n\beta(t-N+)+1}\right]\mathrm{E}[\|\tilde{\boldsymbol{\theta}}(t-1)\|^2] \\ &\quad +2\mathrm{E}\left\{\tilde{\boldsymbol{\theta}}^{\mathrm{T}}(t-1)\left[\boldsymbol{I} - \frac{\boldsymbol{\Phi}(p,t)\boldsymbol{\Phi}^{\mathrm{T}}(p,t)}{r(t)}\right]\frac{\boldsymbol{\Phi}(p,t)\boldsymbol{V}(p,t)}{r(t)}\right\} \\ &\quad + \frac{\mathrm{E}[\boldsymbol{\Phi}(p,t)\boldsymbol{V}(p,t)\|^2]}{[n\alpha(t-N+1)+1]^2} \\ &\leqslant \left[1 - \frac{N\alpha}{n\beta(t-N+1)+1}\right]\mathrm{E}[\|\tilde{\boldsymbol{\theta}}(t-1)\|^2] \\ &\quad + \frac{N^2\beta\sigma_v^2}{[n\alpha(t-N+1)+1]^2}.\end{aligned}$$

Using Lemma 1, we have

$$\begin{aligned}&\lim_{t\to\infty}\mathrm{E}[\|\tilde{\boldsymbol{\theta}}(t)\|^2] \\ &\leqslant \lim_{t\to\infty}\frac{N^2\beta\sigma_v^2}{[n\alpha(t-N+1)+1]^2}\,\frac{n\beta(t-N+1)+1}{N\alpha} \\ &\leqslant \lim_{t\to\infty}\frac{N\beta\sigma_v^2[n\beta(t-N+1)+1]}{\alpha[n\alpha(t-N+1)+1]^2} \\ &\sim \frac{N\beta^2\sigma_v^2}{n\alpha^3}\frac{1}{t} =: C_2\frac{1}{t}, \qquad (A.8)\end{aligned}$$

where

$$C_2 := \frac{N\beta^2\sigma_v^2}{n\alpha^3}.$$

This proves Theorem 2. □

Proof of Lemma 4. From (13) and using (A.1), we have

$$\begin{aligned}r(t) &= \lambda r(t-1) + \|\boldsymbol{\varphi}(t)\|^2 = \sum_{j=1}^{t}\lambda^{t-j}\|\boldsymbol{\varphi}(j)\|^2 + \lambda^t r(0) \\ &\leqslant \sum_{j=1}^{t}\lambda^{t-j}\left[\sum_{i=0}^{N-1}\|\boldsymbol{\varphi}(j+i)\|^2\right] + \lambda^t r(0) \\ &\leqslant \sum_{i=1}^{t}\lambda^{t-i}[nN\beta] + \lambda^t r(0) \\ &= \frac{nN\beta}{1-\lambda} + \left[r(0) - \frac{nN\beta}{1-\lambda}\right]\lambda^t \leqslant \frac{nN\beta}{1-\lambda} \quad a.s.,\end{aligned}$$

$$\begin{aligned}r(t) &\geqslant \sum_{j=1}^{t}\lambda^{t-j}\alpha + \lambda^t r(0) = \frac{\alpha(1-\lambda^t)}{1-\lambda} + \lambda^t r(0) \\ &= \frac{\alpha}{1-\lambda} + \lambda^t\left[r(0) - \frac{\alpha}{1-\lambda}\right] \geqslant \frac{\alpha}{1-\lambda} \quad a.s.\end{aligned}$$

This proves Lemma 4. □

Proof of Theorem 3. Since $\boldsymbol{\theta}(t) = \boldsymbol{\theta}(t-1) + \boldsymbol{w}(t)$, we have

$$\boldsymbol{\theta}(t-1) = \boldsymbol{\theta}(t) - \boldsymbol{w}(t),$$

$$\boldsymbol{\theta}(t-j) = \boldsymbol{\theta}(t-1) - \sum_{l=1}^{j-1}\boldsymbol{w}(t-l).$$

Define the noise vectors,

$$\boldsymbol{V}(p,t) := \begin{bmatrix} v(t) \\ v(t-1) \\ \vdots \\ v(t-p+1)\end{bmatrix} \in \mathbb{R}^p,$$

$$\boldsymbol{W}(p,t) := \begin{bmatrix} 0 \\ \boldsymbol{\varphi}^{\mathrm{T}}(t-1)\boldsymbol{w}(t-1) \\ \boldsymbol{\varphi}^{\mathrm{T}}(t-2)[\boldsymbol{w}(t-1)+\boldsymbol{w}(t-2)] \\ \vdots \\ \boldsymbol{\varphi}^{\mathrm{T}}(t-p+1)\sum_{j=1}^{p-1}\boldsymbol{w}(t-j)\end{bmatrix} \in \mathbb{R}^p,$$

and the parameter estimation error vector

$$\tilde{\boldsymbol{\theta}}(t) := \hat{\boldsymbol{\theta}}(t) - \boldsymbol{\theta}(t).$$

By using (12), it follows that

$$\begin{aligned}\tilde{\boldsymbol{\theta}}(t) &= \hat{\boldsymbol{\theta}}(t) - [\boldsymbol{\theta}(t-1) + \boldsymbol{w}(t)] \\ &= \tilde{\boldsymbol{\theta}}(t-1) + \frac{\boldsymbol{\Phi}(p,t)}{r(t)}[-\boldsymbol{\Phi}^{\mathrm{T}}(p,t)\tilde{\boldsymbol{\theta}}(t-1) \\ &\quad - \boldsymbol{W}(p,t) + \boldsymbol{V}(p,t)] - \boldsymbol{w}(t) \\ &= \left[\boldsymbol{I} - \frac{\boldsymbol{\Phi}(p,t)\boldsymbol{\Phi}^{\mathrm{T}}(p,t)}{r(t)}\right]\tilde{\boldsymbol{\theta}}(t-1) \\ &\quad + \frac{\boldsymbol{\Phi}(p,t)[-\boldsymbol{W}(p,t)+\boldsymbol{V}(p,t)]}{r(t)} - \boldsymbol{w}(t). \qquad (A.9)\end{aligned}$$

Using (A.1), (A.3) and (A.4), we have

$$\mathrm{E}[\|\boldsymbol{\Phi}(p,t)\boldsymbol{V}(p,t)\|^2] \leqslant p\beta\mathrm{E}[\|\boldsymbol{V}(p,t)\|^2] \leqslant p^2\beta\sigma_v^2 = N^2\beta\sigma_v^2,$$

$$\begin{aligned}&\mathrm{E}[\|\boldsymbol{\Phi}(p,t)\boldsymbol{W}(p,t)\|^2] \\ &\leqslant p\beta\mathrm{E}[\|\boldsymbol{W}(p,t)\|^2] \\ &= p\beta\mathrm{E}\{\|\boldsymbol{\varphi}^{\mathrm{T}}(t-1)\boldsymbol{w}(t-1)\|^2 \\ &\quad + \|\boldsymbol{\varphi}^{\mathrm{T}}(t-2)[\boldsymbol{w}(t-1)+\boldsymbol{w}(t-2)]\|^2 \\ &\quad + \cdots + \|\boldsymbol{\varphi}^{\mathrm{T}}(t-p+1) \\ &\quad \times [\boldsymbol{w}(t-1)+\boldsymbol{w}(t-2)+\cdots+\boldsymbol{w}(t-p+1)]\|^2\} \\ &\leqslant p^2\beta^2\mathrm{E}[\|\boldsymbol{w}(t-1)\|^2 + \|\boldsymbol{w}(t-1)+\boldsymbol{w}(t-2)\|^2 \\ &\quad + \cdots + \|\boldsymbol{w}(t-1)+\boldsymbol{w}(t-2)+\cdots+\boldsymbol{w}(t-p+1)\|^2] \\ &\leqslant \frac{(p-1)p^3\beta^2\sigma_w^2}{2} \leqslant \frac{p^4\beta^2\sigma_w^2}{2} = \frac{N^4\beta^2\sigma_w^2}{2}.\end{aligned}$$

Hence, using Lemma 4 yields

$$\mathrm{E}\left[\frac{\|\boldsymbol{\Phi}(p,t)\boldsymbol{V}(p,t)\|^2}{r^2(t)}\right]\leqslant\frac{N^2\beta(1-\lambda)^2\sigma_v^2}{\alpha^2},$$
$$\mathrm{E}\left[\frac{\|\boldsymbol{\Phi}(p,t)\boldsymbol{W}(p,t)\|^2}{r^2(t)}\right]\leqslant\frac{N^4\beta^2(1-\lambda)^2\sigma_w^2}{2\alpha^2},$$

and

$$\boldsymbol{I}-\frac{\boldsymbol{\Phi}(p,t)\boldsymbol{\Phi}^{\mathrm{T}}(p,t)}{r(t)}\leqslant\left[1-\frac{\alpha(1-\lambda)}{n\beta}\right]\boldsymbol{I}=:(1-\rho)\boldsymbol{I}.$$

Taking the norm of both sides of (A.9) and using the inequality $\|\boldsymbol{x}+\boldsymbol{y}\|^2\leqslant(1+a)\|\boldsymbol{x}\|^2+(1+a^{-1})\|\boldsymbol{y}\|^2$ $(a>0)$ yield

$$\begin{aligned}\|\tilde{\boldsymbol{\theta}}(t)\|^2&\leqslant(1+a)\left\|\left[\boldsymbol{I}-\frac{\boldsymbol{\Phi}(p,t)\boldsymbol{\Phi}^{\mathrm{T}}(p,t)}{r(t)}\right]\tilde{\boldsymbol{\theta}}(t-1)\right\|^2\\&\quad+(1+a^{-1})\left\|\frac{\boldsymbol{\Phi}(p,t)[-\boldsymbol{W}(p,t)+\boldsymbol{V}(p,t)]}{r(t)}-\boldsymbol{w}(t)\right\|^2\\&\leqslant(1+a)(1-\rho)\|\tilde{\boldsymbol{\theta}}(t-1)\|^2\\&\quad+3(1+a^{-1})\left[\frac{\|\boldsymbol{\Phi}(p,t)\boldsymbol{W}(p,t)\|^2}{r^2(t)}\right.\\&\quad\left.+\frac{\|\boldsymbol{\Phi}(p,t)\boldsymbol{V}(p,t)\|^2}{r^2(t)}+\|\boldsymbol{w}(t)\|^2\right].\end{aligned}$$

Taking the expectation gives

$$\begin{aligned}\mathrm{E}[\|\tilde{\boldsymbol{\theta}}(t)\|^2]&\leqslant(1+a)(1-\rho)\mathrm{E}[\|\tilde{\boldsymbol{\theta}}(t-1)\|^2]\\&\quad+3(1+a^{-1})\left[\frac{N^4\beta^2(1-\lambda)^2\sigma_w^2}{2\alpha^2}\right.\\&\quad\left.+\frac{N^2\beta(1-\lambda)^2\sigma_v^2}{\alpha^2}+\sigma_w^2\right].\end{aligned}\tag{A.10}$$

We take a to satisfy

$$0<a<\frac{\rho}{1-\rho}.$$

That is, $0<(1+a)(1-\rho)<1$; successive substitution in (A.10) gives

$$\begin{aligned}\mathrm{E}[\|\tilde{\boldsymbol{\theta}}(t)\|^2]&\leqslant[(1+a)(1-\rho)]^t\mathrm{E}[\|\tilde{\boldsymbol{\theta}}(0)\|^2]\\&\quad+3(1+a^{-1})\sum_{i=0}^{t-1}[(1+a)(1-\rho)]^i\\&\quad\times\left[\frac{N^4\beta^2(1-\lambda)^2\sigma_w^2}{2\alpha^2}+\frac{N^2\beta(1-\lambda)^2\sigma_v^2}{\alpha^2}+\sigma_w^2\right]\\&\leqslant[(1+a)(1-\rho)]^t\delta_0\\&\quad+3(1+a^{-1})\frac{1-[(1+a)(1-\rho)]^t}{1-(1+a)(1-\rho)}\\&\quad\times\left[\frac{N^4\beta^2(1-\lambda)^2\sigma_w^2}{2\alpha^2}+\frac{N^2\beta(1-\lambda)^2\sigma_v^2}{\alpha^2}+\sigma_w^2\right]\\&\leqslant[(1+a)(1-\rho)]^t\delta_0\\&\quad+\frac{3(1+a^{-1})}{1-(1+a)(1-\rho)}\left[\frac{N^4\beta^2(1-\lambda)^2\sigma_w^2}{2\alpha^2}\right.\\&\quad\left.+\frac{N^2\beta(1-\lambda)^2\sigma_v^2}{\alpha^2}+\sigma_w^2\right]\\&=:[(1+a)(1-\rho)]^t\delta_0+f(a,\lambda),\end{aligned}\tag{A.11}$$

where

$$\begin{aligned}f(a,\lambda)&:=g(a)\left[\frac{N^4\beta^2(1-\lambda)^2\sigma_w^2}{2\alpha^2}+\frac{N^2\beta(1-\lambda)^2\sigma_v^2}{\alpha^2}+\sigma_w^2\right],\\g(a)&:=\frac{3(1+a^{-1})}{1-(1+a)(1-\rho)}.\end{aligned}\tag{A.12}$$

In the above derivation, we have introduced an intermediate variable a and function $g(a)$. In order to obtain a minimum estimation error upper bound, minimizing the expression on the right-hand side of (A.11) and (A.12), we must determine a best a such that $g(a)$ is minimum, and thus so is $f(a,\lambda)$ for a. Let the first-order derivative of $g(a)$ with respect to a be zero, i.e.,

$$\frac{\mathrm{d}g(a)}{\mathrm{d}a}=0$$

which leads to

$$(1-\rho)a^2+2(1-\rho)a-\rho=0,$$

or

$$(1-\rho)(a+1)^2=1.$$

Its solutions are

$$a=\pm\frac{1}{\sqrt{1-\rho}}-1.$$

Since a is a positive number, the best a value is

$$a=a_0=\frac{1}{\sqrt{1-\rho}}-1,$$

and the corresponding minimum $g(a)$ is

$$\min g(a)=g(a_0)=\frac{3}{(1-\sqrt{1-\rho})^2}.$$

Thus, we have

$$\begin{aligned}f(a_0,\lambda)&=\frac{3}{(1-\sqrt{1-\rho})^2}\left[\frac{N^4\beta^2(1-\lambda)^2\sigma_w^2}{2\alpha^2}\right.\\&\quad\left.+\frac{N^2\beta(1-\lambda)^2\sigma_v^2}{\alpha^2}+\sigma_w^2\right]\end{aligned}$$

and

$$\begin{aligned}\mathrm{E}[\|\tilde{\boldsymbol{\theta}}(t)\|^2]&\leqslant[(1+a_0)(1-\rho)]^t\delta_0+f(a_0,\lambda)\\&\leqslant[\sqrt{1-\rho}\,]^t\delta_0+\frac{3}{(1-\sqrt{1-\rho})^2}\\&\quad\times\left[\frac{N^4\beta^2(1-\lambda)^2\sigma_w^2}{2\alpha^2}+\frac{N^2\beta(1-\lambda)^2\sigma_v^2}{\alpha^2}+\sigma_w^2\right].\end{aligned}$$

This directly gives the conclusion of Theorem 3. □

References

Chen, H. F., & Guo, L. (1987). Adaptive control via consistent estimation for deterministic systems. *International Journal of Control*, *45*(6), 2183–2202.

Ding, F., & Chen, T. (2004). Combined parameter and output estimation of dual-rate systems using an auxiliary model. *Automatica*, *40*(10), 1739–1748.

Ding, F., & Chen, T. (2005a). Performance bounds of forgetting factor least squares algorithm for time-varying systems with finite measurement data. *IEEE Transactions on Circuits and Systems—I: Regular Papers*, *52*(3), 555–566.

Ding, F., & Chen, T. (2005b). Parameter estimation of dual-rate stochastic systems by using an output error method. *IEEE Transactions on Automatic Control*, *50*(9), 1436–1441.

Ding, F., & Chen, T. (2005c). Identification of Hammerstein nonlinear ARMAX systems. *Automatica*, *41*(9), 1479–1489.

Ding, F., & Chen, T. (2005d). Hierarchical gradient-based identification of multivariable discrete-time systems. *Automatica*, *41*(2), 315–325.

Ding, F., & Chen, T. (2005e). Hierarchical least squares identification methods for multivariable systems. *IEEE Transactions on Automatic Control*, *50*(3), 397–402.

Ding, F., & Chen, T., (2006). Multi-innovation stochastic gradient identification methods. *Proceedings of the sixth world congress on intelligent control and automation (WCICA2006)*, June 21–23, 2006, Dalian, China (pp. 1501–1505).

Ding, F., Shi, Y., & Chen, T. (2006). Performance analysis of estimation algorithms of non-stationary ARMA processes. *IEEE Transactions on Signal Processing*, *54*(3), 1041–1053.

Ding, F., Xiao, D. Y., & Ding, T. (2003). Multi-innovation stochastic gradient identification methods. *Control Theory and Applications*, *20*(6), 870–874.

Ding, F., Xie, X. M., & Fang, C. Z. (1996). Multi-innovation identification methods for time-varying systems. *Acta Automatica Sinica*, *22*(1), 85–91.

Goodwin, G. C., & Sin, K. S. (1984). *Adaptive filtering, prediction and control*. Englewood Cliffs, NJ: Prentice-Hall.

Guo, L. (1995). Convergence and logarithm laws of self-tuning regulators. *Automatica*, *31*(3), 435–450.

Guo, L., & Ljung, L. (1995a). Exponential stability of general tracking algorithms. *IEEE Transactions on Automatic Control*, *40*(8), 1376–1387.

Guo, L., & Ljung, L. (1995b). Performance analysis of general tracking algorithms. *IEEE Transactions on Automatic Control*, *40*(8), 1388–1402.

Kumar, K. (2000). On the identification of autoregressive moving average models. *Control and Intelligent Systems*, *28*(2), 41–46.

Lai, T. L., & Wei, C. Z. (1982). Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *Annals of Statistics*, *10*(1), 154–166.

Lai, T. L., & Wei, C. Z. (1986). Extended least squares and their applications to adaptive control and prediction in linear systems. *IEEE Transactions on Automatic Control, 31*(10), 898–906.

Ljung, L. (1976). Consistency of the least-squares identification method. *IEEE Transactions on Automatic Control*, *21*(5), 779–781.

Ljung, L. (1999). *System identification: Theory for the user*. (2nd ed.), Englewood Cliffs, NJ: Prentice-Hall.

Ren, W., & Kumar, P. K. (1994). Stochastic adaptive prediction and model reference control. *IEEE Transactions on Automatic Control*, *39*(10), 2047–2060.

Wei, C. Z. (1987). Adaptive prediction by least squares prediction in stochastic regression models. *Annals of Statistics*, *15*(4), 1667–1682.

Feng Ding was born in Guangshui, Hubei Province. He received the B.Sc. degree from the Hubei University of Technology (Wuhan, China) in 1984, and the M.Sc. and Ph.D. degrees in automatic control both from the Department of Automation, Tsinghua University in 1991 and 1994, respectively.

From 1984 to 1988, he was an Electrical Engineer at the Hubei Pharmaceutical Factory, Xiangfan, China. From 1994 to 2002, he was with the Department of Automation at the Tsinghua University, Beijing, China. From 2002 to 2005, he was a Post-Doctoral Fellow/Research Associate at the University of Alberta, Edmonton, Canada. From 2004, he is a Professor in the Control Science and Engineering Research Center at the Southern Yangtze University, Wuxi, China, and his current research interests include model identification and adaptive control. He co-authored the book *Adaptive Control Systems* (Tsinghua University Press, Beijing, 2002), and published over 100 papers on modeling and identification as the first author.

Tongwen Chen received the B.Eng. degree in Automation and Instrumentation from Tsinghua University (Beijing) in 1984, and the M.A.Sc. and Ph.D. degrees in Electrical Engineering from the University of Toronto in 1988 and 1991, respectively. From 1991 to 1997, he was an Assistant/Associate Professor in the Department of Electrical and Computer Engineering at the University of Calgary, Canada. Since 1997, he has been with the Department of Electrical and Computer Engineering at the University of Alberta, Edmonton, Canada, and is presently a Professor. He held visiting positions at the Hong Kong University of Science and Technology, Tsinghua University, and Kumamoto University.

His research interests include computer and network based control systems, process control, multirate digital signal processing, and their applications to industrial problems. He co-authored with B.A. Francis the book *Optimal Sampled-Data Control Systems* (Springer, 1995). Dr. Chen received a McCalla Professorship for 2000–2001 and a Killam Professorship for 2006–2007, both from the University of Alberta, and a Fellowship from the Japan Society for the Promotion of Science for 2004. He was elected an IEEE Fellow in 2006. He has served as an Associate Editor for several international journals, including IEEE Transactions on Automatic Control, Automatica, and Systems and Control Letters. He is a registered Professional Engineer in Alberta, Canada.