# Auxiliary Model-based Stochastic Gradient Algorithm for Multivariable Output Error Systems

DING Feng[1] LIU Xiao-Ping[2, 3]

**Abstract** The identification problem of multivariable output error systems is considered in this paper. By constructing an auxiliary model using available input-output data and by replacing the unknown inner variables in the information vector with the outputs of the auxiliary model, an auxiliary model-based stochastic gradient (AM-SG) identification algorithm is presented. Convergence analysis using the martingale convergence theorem indicates that the parameter estimates given by the AM-SG algorithm converge to their true values. The AM-SG algorithm with a forgetting factor is given to improve its convergence rate. The simulation results confirm the theoretical findings.

**Key words** Recursive identification, parameter estimation, stochastic gradient (SG), auxiliary model identification, multivariable systems, convergence property, martingale convergence theorem

**DOI** 10.3724/SP.J.1004.2010.00993

Parameter estimation has had wide applications in many areas, including signal processing, adaptive prediction and control, time-series analysis, process modeling, and so on. In the area of system identification, Zheng used the bias correction method in the identification of linear dynamic errors-in-variables systems[1], Yang et al. made comparisons of some bias compensation methods and other identification approaches for Box-Jenkins models[2], Zhang et al. presented a bias compensation recursive least squares identification for output error systems with colored noises[3], Zong et al. studied the iterative identification problem related to control design[4], Zhong et al. discussed the hierarchical optimization identification for linear state space systems[5].

The least squares identification algorithms have fast convergence rates. Recently, Wang presented an auxiliary model-based recursive extended least squares identification method for output error moving average systems[6]. The stochastic gradient (SG) parameter estimation algorithms have less computation load and have received much attention in self-tuning control and system identification. In the literature, many gradient-based identification approaches were reported. For example, Ding et al. proposed a hierarchical SG algorithm for multivariable systems[7] and a multi-innovation SG algorithm for linear regression model[8]. Wang et al. developed an auxiliary model-based multi-innovation generalized extended stochastic gradient (ESG) identification algorithm for Box-Jenkins models[9] using the multi-innovation identification theory[8], but no convergence analysis was carried out. Also, Wang et al. gave an ESG identification algorithm for Hammerstein-Wiener nonlinear ARMAX systems[10].

This paper studies the gradient-based identification approach and convergence for multivariable systems with output measurement noises. For such a system, the difficulty in identification is that the information vector contains unmeasurable variables. Our solution is to use the auxiliary model technique[11−12], to replace these unknown variables with the outputs of the auxiliary model, to present an auxiliary model-based stochastic gradient (AM-SG) identification algorithm, and further to analyze the convergence of the proposed algorithm. To improve the convergence rate of the gradient-based algorithm, an AM-SG algorithm with a forgetting factor (AM-FFSG algorithm) is given. Compared with the auxiliary model-based recursive least squares algorithm, the AM-FFSG algorithm has less computation burden. The simulation results indicate that if we choose the forgetting factor appropriately, the AM-FFSG algorithm can achieve faster convergence rate and the parameter estimation accuracy is closer to that of the least squares algorithm. The AM-SG algorithm with the unknown information vector for the output error systems in this paper differs from the standard SG algorithm in [13], which assumes that each entry of the information vector is known.

Briefly, the rest of the paper is organized as follows. Section 1 simply describes the identification problem to be discussed in the paper. Section 2 derives a basic identification algorithm for multivariable systems based on the auxiliary model technique. Sections 3 analyzes the performance of the proposed algorithm. Section 4 presents an illustrative example for the results in this paper. Finally, concluding remarks are given in Section 5.

## 1 Problem formulation

Consider a multivariable, i.e., multi-input multi-output (MIMO) output error system

$$\boldsymbol{x}(t) = G(z)\boldsymbol{u}(t) \tag{1}$$

$$\boldsymbol{y}(t) = \boldsymbol{x}(t) + \boldsymbol{v}(t) = G(z)\boldsymbol{u}(t) + \boldsymbol{v}(t) \tag{2}$$

which is different from the equation-error systems (CAR/ARX model) in [13], where $\boldsymbol{u}(t) \in \mathbf{R}^r$ is the system input vector, $\boldsymbol{x}(t) \in \mathbf{R}^m$ the system output vector (the true output or noise-free output), $\boldsymbol{y}(t) \in \mathbf{R}^m$ is the measurement of $\boldsymbol{x}(t)$ contaminated by the noise $\boldsymbol{v}(t) \in \mathbf{R}^m$, as depicted in Fig. 1, $G(z) \in \mathbf{R}^{m\times r}$ is the transfer matrix of the system with $z^{-1}$ representing the unit delay operator $z^{-1}$ $[z^{-1}\boldsymbol{u}(t) = \boldsymbol{u}(t-1)]$.

According to the matrix polynomial theory[14], any strictly proper rational fraction matrix can be decomposed into a matrix fraction description: $G(z) = A^{-1}(z)B(z)$, where $A(z)$ and $B(z)$ are polynomial matrices in $z^{-1}$ and defined as

Manuscript received December 17, 2008; accepted February 24, 2009

Supported by National Natural Science Foundation of China (60874020)

1. School of Communication and Control Engineering, Jiangnan University, Wuxi 214122, P. R. China 2. School of Mechatronics Engineering, Nanchang University, Nanchang 330031, P. R. China 3. Department of Systems and Computer Engineering, Carleton University, Ottawa K1S 5B6, Canada

$$A(z)=I+A_1z^{-1}+A_2z^{-2}+\cdots+A_{n_a}z^{-n_a}\in\mathbf{R}^{m\times m}$$
$$B(z)=B_1z^{-1}+B_2z^{-2}+\cdots+B_{n_b}z^{-n_b}\in\mathbf{R}^{m\times r}$$

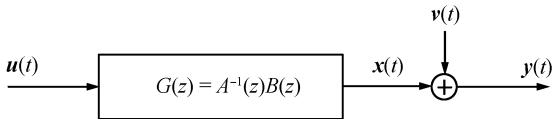


Fig. 1 The output-error system

Here, the inner variable $\boldsymbol{x}(t)$ is unknown, $\boldsymbol{y}(t)$ is the measurement output vector, and $\boldsymbol{v}(t)$ is the observation noise vector with zero mean. $\{\boldsymbol{u}(t),\boldsymbol{y}(t)\}$ are the measurement input-output data, and $A_i\in\mathbf{R}^{m\times m}$ and $B_i\in\mathbf{R}^{m\times r}$ are the parameter matrices to be identified. Assume that $n_a$ and $n_b$ are known and that $\boldsymbol{u}(t)=\boldsymbol{0}$, $\boldsymbol{y}(t)=\boldsymbol{0}$, and $\boldsymbol{v}(t)=\boldsymbol{0}$ for $t\le 0$. The system in (1) and (2) contains $m^2n_a+mrn_b=S_1$ parameters. The objective is to present an AM-SG algorithm to estimate the unknown parameter matrices $A_i$ and $B_i$ using the input-output data $\{\boldsymbol{u}(t),\boldsymbol{y}(t)\}$.

The model in (1) and (2) may be equivalently written as an MIMO ARMAX model[15−16]:

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

This model can be identified by using the extended least squares (ELS) or ESG algorithm[16]. Using the ELS algorithm to estimate the parameters of such a special ARMAX model indeed requires identifying $m^2n_a$ number of more parameters than the actual model parameters. Although the noise model $D(z)$ equals $A(z)$, their estimates are different. In other words, the size of the parameter matrix increases, so this directly leads to a heavier computational burden. Therefore, exploring computationally efficient identification approach is the goal of this paper. The following is to derive auxiliary model identification algorithms with less computation.

## 2 Basic algorithms

Let us introduce some notations first. The symbol $I$ ($I_m$) stands for an identity matrix of appropriate size (of $m\times m$), the superscript T denotes the matrix transpose, the norm of a matrix $X$ is defined by $\|X\|^2=\mathrm{tr}[XX^{\mathrm{T}}]$, $\mathbf{1}_{m\times n}$ represents an $m\times n$ matrix whose elements are 1 and $\mathbf{1}_n=\mathbf{1}_{n\times 1}$, $\lambda_{\max}[X]$ and $\lambda_{\min}[X]$ represent the maximum and minimum eigenvalues of the square matrix $X$, respectively; for $g(t)\ge 0$, we write $f(t)=O(g(t))$ if there exist positive constants $\delta_1$ and $t_0$ such that $|f(t)|\le\delta_1 g(t)$ for $t\ge t_0$; $f(t)=o(g(t))$ represents $f(t)/g(t)\to 0$ as $t\to\infty$.

Let $n=mn_a+rn_b$. Define the parameter matrix $\theta$ and information vector $\boldsymbol{\varphi}_0(t)$ as

$$\theta^{\mathrm{T}}=[A_1,A_2,\cdots,A_{n_a},B_1,B_2,\cdots,B_{n_b}]\in\mathbf{R}^{m\times n}$$
$$\boldsymbol{\varphi}_0(t)=[-\boldsymbol{x}^{\mathrm{T}}(t-1),-\boldsymbol{x}^{\mathrm{T}}(t-2),\cdots,-\boldsymbol{x}^{\mathrm{T}}(t-n_a),\ \boldsymbol{u}^{\mathrm{T}}(t-1),\boldsymbol{u}^{\mathrm{T}}(t-2),\cdots,\boldsymbol{u}^{\mathrm{T}}(t-n_b)]^{\mathrm{T}}\in\mathbf{R}^n$$

Then, from (1) to (2), we have

$$\boldsymbol{x}(t)=\theta^{\mathrm{T}}\boldsymbol{\varphi}_0(t)$$
$$\boldsymbol{y}(t)=\theta^{\mathrm{T}}\boldsymbol{\varphi}_0(t)+\boldsymbol{v}(t)\tag{3}$$

Here, a difficulty arises in that the information vector $\boldsymbol{\varphi}_0(t)$ contains unknown $\boldsymbol{x}(t-i)$ so that the standard SG methods cannot be applied to (3) directly. The objective of this work is to establish an auxiliary model by using the available data $\{\boldsymbol{u}(t),\boldsymbol{y}(t)\}$, to present auxiliary model-based SG algorithms by using the output $\boldsymbol{x}_{\mathrm{a}}(t)$ of this auxiliary model in place of the unknown $\boldsymbol{x}(t)$, and to analyze the performance of the proposed algorithms.

Let $\hat{\theta}(t)$ be the estimate of $\theta$ at time $t$:

$$\hat{\theta}^{\mathrm{T}}(t)=[\hat{A}_1(t),\cdots,\hat{A}_{n_a}(t),\hat{B}_1(t),\cdots,\hat{B}_{n_b}(t)]$$

and use the entries of the estimate $\hat{\theta}(t)$ to form the polynomials:

$$\hat{A}(z)=I+\hat{A}_1(t)z^{-1}+\hat{A}_2(t)z^{-2}+\cdots+\hat{A}_{n_a}(t)z^{-n_a}$$
$$\hat{B}(z)=\hat{B}_1(t)z^{-1}+\hat{B}_2(t)z^{-2}+\cdots+\hat{B}_{n_b}(t)z^{-n_b}$$

In terms of $\hat{A}(z)$ and $\hat{B}(z)$, we construct an auxiliary model:

$$\boldsymbol{x}_{\mathrm{a}}(t)=G_{\mathrm{a}}(z)\boldsymbol{u}(t),\ G_{\mathrm{a}}(z)=\hat{A}^{-1}(z)\hat{B}(z)\tag{4}$$

where $G_{\mathrm{a}}(z)$ denotes the estimate of $G(z)$ and is used as the transfer function matrix of the the auxiliary model.

Equation (4) may also be written in the matrix form:

$$\boldsymbol{x}_{\mathrm{a}}(t)=\hat{\theta}^{\mathrm{T}}(t)\boldsymbol{\varphi}(t)$$
$$\boldsymbol{\varphi}(t)=[-\boldsymbol{x}_{\mathrm{a}}^{\mathrm{T}}(t-1),\cdots,-\boldsymbol{x}_{\mathrm{a}}^{\mathrm{T}}(t-n_a)\ \boldsymbol{u}^{\mathrm{T}}(t-1),\cdots,\boldsymbol{u}^{\mathrm{T}}(t-n_b)]^{\mathrm{T}}$$

If we use $\boldsymbol{\varphi}(t)$ to replace $\boldsymbol{\varphi}_0(t)$ in (3), the identification problem of $\theta$ can be solved. Using this idea, we can obtain an AM-SG algorithm of estimating the parameter matrix $\theta$ of the multivariable systems

$$\hat{\theta}(t)=\hat{\theta}(t-1)+\frac{\boldsymbol{\varphi}(t)}{r(t)}\boldsymbol{e}^{\mathrm{T}}(t)\tag{5}$$
$$\boldsymbol{e}(t)=\boldsymbol{y}(t)-\hat{\theta}^{\mathrm{T}}(t-1)\boldsymbol{\varphi}(t)\tag{6}$$
$$r(t)=r(t-1)+\|\boldsymbol{\varphi}(t)\|^2,\ r(0)=1\tag{7}$$
$$\boldsymbol{\varphi}(t)=[-\boldsymbol{x}_{\mathrm{a}}^{\mathrm{T}}(t-1),\cdots,-\boldsymbol{x}_{\mathrm{a}}^{\mathrm{T}}(t-n_a),\ \boldsymbol{u}^{\mathrm{T}}(t-1),\cdots,\boldsymbol{u}^{\mathrm{T}}(t-n_b)]^{\mathrm{T}}\tag{8}$$
$$\boldsymbol{x}_{\mathrm{a}}(t)=\hat{\theta}^{\mathrm{T}}(t)\boldsymbol{\varphi}(t)\tag{9}$$

The initial value is generally chosen to be a real matrix with smaller entries, e.g., $\hat{\theta}(0)=10^{-6}\mathbf{1}_{m\times n}$.

## 3 Main convergence results

We assume that $\{\boldsymbol{v}(t),\boldsymbol{F}_t\}$ is a martingale difference vector sequence defined on a probability space $\{\Omega,\boldsymbol{F},P\}$, where $\{\boldsymbol{F}_t\}$ is the $\sigma$ algebra sequence generated by the observation data up to and including time $t$[16]. The sequence $\{\boldsymbol{v}(t)\}$ satisfies:

Assumption 1. $\mathrm{E}[\boldsymbol{v}(t)|\boldsymbol{F}_{t-1}]=\boldsymbol{0}$, a.s.;

Assumption 2. $\mathrm{E}[\|\boldsymbol{v}(t)\|^2|\boldsymbol{F}_{t-1}]=\sigma^2r^{\epsilon}(t-1)$, $\sigma^2<\infty$, $\epsilon<1$, a.s..

where a.s. means almost surely.

**Lemma 1**[13]. For the algorithm in (5) ∼ (9), the following inequality holds:

$$\sum_{i=1}^{t}\frac{\|\boldsymbol{\varphi}(i)\|^2}{r(i)}\le\ln r(t),\ \text{a.s.}$$

**Theorem 1.** For the system in (3) and algorithm in (5) ∼ (9), define the data product moment matrix:

$$Q(t)=\sum_{i=1}^{t}\boldsymbol{\varphi}(i)\boldsymbol{\varphi}^{\mathrm{T}}(i)$$

and assume that Assumptions 1 and 2 hold, $A(z)$ is a strictly positive real matrix, $r(t)\to\infty$. Then, the parameter estimation matrix $\hat{\theta}(t)$ consistently converges to $\theta$.

The stochastic martingale theory is one of the main tools of analyzing the convergence of identification algorithms[11−12]. The following proves this theorem by formulating a martingale process and by using the martingale convergence theorem in [16] and the method in [13].

**Proof.** Define the parameter estimation error matrix:

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

Using (5) gives

$$\tilde{\theta}(t)=\tilde{\theta}(t-1)+\frac{\boldsymbol{\varphi}(t)}{r(t)}\boldsymbol{e}^{\mathrm{T}}(t) \tag{10}$$

Let

$$\begin{aligned}\tilde{\boldsymbol{y}}(t)&=-\tilde{\theta}^{\mathrm{T}}(t)\boldsymbol{\varphi}(t)\\ \boldsymbol{\eta}(t)&=\boldsymbol{y}(t)-\hat{\theta}^{\mathrm{T}}(t)\boldsymbol{\varphi}(t)\end{aligned} \tag{11}$$

Using (6), it follows that

$$\begin{aligned}\boldsymbol{\eta}(t)=\frac{r(t-1)}{r(t)}\boldsymbol{e}(t)=\boldsymbol{y}(t)-\boldsymbol{x}_{\mathrm{a}}(t)=\\ \boldsymbol{x}(t)+\boldsymbol{v}(t)-\boldsymbol{x}_{\mathrm{a}}(t)\end{aligned} \tag{12}$$

Taking the norm of both sides of (10) and using (11) yield

$$\begin{aligned}\|\tilde{\theta}(t)\|^2=\|\tilde{\theta}(t-1)\|^2-\frac{2\tilde{\boldsymbol{y}}^{\mathrm{T}}(t)[\boldsymbol{\eta}(t)-\boldsymbol{v}(t)]}{r(t-1)}+\\ \frac{2\boldsymbol{\varphi}^{\mathrm{T}}(t)\tilde{\theta}(t-1)\boldsymbol{v}(t)}{r(t-1)}+\frac{2\|\boldsymbol{\varphi}(t)\|^2}{r(t-1)r(t)}[\boldsymbol{e}(t)-\boldsymbol{v}(t)]^{\mathrm{T}}\boldsymbol{v}(t)+\\ \frac{2\|\boldsymbol{\varphi}(t)\|^2\|\boldsymbol{v}(t)\|^2}{r(t-1)r(t)}-\frac{\|\boldsymbol{\varphi}(t)\|^2}{r^2(t)}\|\boldsymbol{e}(t)\|^2\end{aligned} \tag{13}$$

From (12), we have

$$\begin{aligned}A(z)[\boldsymbol{\eta}(t)-\boldsymbol{v}(t)]=A(z)\boldsymbol{x}(t)-A(z)\boldsymbol{x}_{\mathrm{a}}(t)=\\ B(z)\boldsymbol{u}(t)-A(z)\boldsymbol{x}_{\mathrm{a}}(t)=\theta^{\mathrm{T}}\boldsymbol{\varphi}(t)-\boldsymbol{x}_{\mathrm{a}}(t)=\\ \theta^{\mathrm{T}}\boldsymbol{\varphi}(t)-\hat{\theta}^{\mathrm{T}}(t)\boldsymbol{\varphi}(t)=-\tilde{\theta}^{\mathrm{T}}(t)\boldsymbol{\varphi}(t)=\tilde{\boldsymbol{y}}(t)\end{aligned} \tag{14}$$

Since $A(z)$ is strictly positive real, referring to Appendix C in [16], the following inequality holds:

$$S(t)=\sum_{i=1}^{t}\frac{2\tilde{\boldsymbol{y}}^{\mathrm{T}}(i)[\boldsymbol{\eta}(i)-\boldsymbol{v}(i)]}{r(t-1)}\geq 0,\text{ a.s.}$$

Let $W(t)=\|\tilde{\theta}(t)\|^2+S(t)$. Adding both sides of (13) by $S(t)$ gives

$$\begin{aligned}W(t)=W(t-1)+\frac{2\boldsymbol{\varphi}^{\mathrm{T}}(t)\tilde{\theta}(t-1)\boldsymbol{v}(t)}{r(t-1)}+\\ \frac{2\|\boldsymbol{\varphi}(t)\|^2}{r(t-1)r(t)}[\boldsymbol{e}(t)-\boldsymbol{v}(t)]^{\mathrm{T}}\boldsymbol{v}(t)+\\ \frac{2\|\boldsymbol{\varphi}(t)\|^2\|\boldsymbol{v}(t)\|^2}{r(t-1)r(t)}-\frac{\|\boldsymbol{\varphi}(t)\|^2}{r^2(t)}\|\boldsymbol{e}(t)\|^2\end{aligned}$$

Since $S(t-1)$, $\boldsymbol{\varphi}^{\mathrm{T}}(t)\tilde{\theta}(t-1)$, $r(t-1)$, $\boldsymbol{\varphi}(t)$, $r(t)$, and $\boldsymbol{e}(t)-\boldsymbol{v}(t)$ are uncorrelated with $\boldsymbol{v}(t)$, and $\boldsymbol{F}_{t-1}$ measurable, taking the conditional expectation of both sides of the above equation with respect to $\boldsymbol{F}_{t-1}$ and using Assumptions 1 and 2 yield

$$\begin{aligned}\mathrm{E}[W(t)|\boldsymbol{F}_{t-1}]=W(t-1)+\frac{2\|\boldsymbol{\varphi}(t)\|^2\sigma^2 r^{\epsilon}(t-1)}{r(t-1)r(t)}-\\ \mathrm{E}\left[\frac{\|\boldsymbol{\varphi}(t)\|^2}{r^2(t)}\|\boldsymbol{e}(t)\|^2|\boldsymbol{F}_{t-1}\right],\text{a.s.}\end{aligned} \tag{15}$$

The summation of the second term on the right-hand side of the above equation from $t=1$ to $t=\infty$ is finite[16] , i.e.,

$$\sigma^2\sum_{t=1}^{\infty}\frac{\|\boldsymbol{\varphi}(t)\|^2}{[r(t-1)]^{1-\epsilon}r(t)}<\infty,\text{a.s.},\ 1-\epsilon>0$$

Applying the martingale convergence theorem (Lemma D.5.3 in [16]) to (15) to get that $W(t)$ almost surely (a.s.) converges to a finite random variable, say, $C$, i.e.,

$$\lim_{t\to\infty}\|\tilde{\theta}(t)\|^2+S(t)=C<\infty,\text{a.s.} \tag{16}$$

and also

$$\sum_{t=1}^{\infty}\frac{\|\boldsymbol{\varphi}(t)\|^2}{r^2(t)}\|\boldsymbol{e}(t)\|^2<\infty,\text{a.s.} \tag{17}$$

Hence,

$$\begin{aligned}\sum_{t=1}^{\infty}\frac{\|\tilde{\boldsymbol{y}}(t)\|^2}{r(t-1)}<\infty,\text{a.s.}\\ \sum_{t=1}^{\infty}\frac{\|\boldsymbol{\eta}(t)-\boldsymbol{v}(t)\|^2}{r(t-1)}<\infty,\text{a.s.}\end{aligned} \tag{18}$$

Using the Kronecker lemma (Lemma D.5.5 in [16]), it follows that

$$\begin{aligned}\lim_{t\to\infty}\frac{1}{r(t-1)}\sum_{i=1}^{t}\|\tilde{\boldsymbol{y}}(i)\|^2=0,\text{a.s.}\\ \lim_{t\to\infty}\frac{1}{r(t-1)}\sum_{i=1}^{t}\|\boldsymbol{\eta}(i)-\boldsymbol{v}(i)\|^2=0,\text{a.s.}\end{aligned}$$

Equation (16) shows that the parameter estimation error is consistent bounded. From (10), we have

$$\tilde{\theta}(t)=\tilde{\theta}(t-i)+\sum_{j=0}^{i-1}\frac{\boldsymbol{\varphi}(t-j)}{r(t-j)}\boldsymbol{e}^{\mathrm{T}}(t-j),\quad i\geq 1 \tag{19}$$

Thus, we have

$$\begin{aligned}\sum_{t=i}^{\infty}\|\tilde{\theta}(t)-\tilde{\theta}(t-i)\|^2=\sum_{t=i}^{\infty}\|\hat{\theta}(t)-\hat{\theta}(t-i)\|^2=\\ \sum_{t=i}^{\infty}\left\|\sum_{j=0}^{i-1}\frac{\boldsymbol{\varphi}(t-j)}{r(t-j)}\boldsymbol{e}^{\mathrm{T}}(t-j)\right\|^2\leq\\ i\sum_{j=0}^{i-1}\sum_{t=i}^{\infty}\frac{\|\boldsymbol{\varphi}(t-j)\|^2}{r^2(t-j)}\|\boldsymbol{e}(t-j)\|^2<\infty,\text{a.s.},\ i<\infty\end{aligned}$$

$$\sum_{t=1}^{\infty}\frac{\|\boldsymbol{e}(t)-\boldsymbol{v}(t)\|^2}{r(t-1)}=\sum_{t=1}^{\infty}\frac{\|\boldsymbol{y}(t)-\hat{\theta}^{\mathrm{T}}(t-1)\boldsymbol{\varphi}(t)-\boldsymbol{v}(t)\|^2}{r(t-1)}=$$
$$\sum_{t=1}^{\infty}\frac{\|\boldsymbol{y}(t)-\hat{\theta}^{\mathrm{T}}(t)\boldsymbol{\varphi}(t)-\boldsymbol{v}(t)+[\hat{\theta}^{\mathrm{T}}(t)-\hat{\theta}^{\mathrm{T}}(t-1)]\boldsymbol{\varphi}(t)\|^2}{r(t-1)}=$$
$$\sum_{t=1}^{\infty}\frac{\|\boldsymbol{\eta}(t)-\boldsymbol{v}(t)+[\tilde{\theta}^{\mathrm{T}}(t)-\tilde{\theta}^{\mathrm{T}}(t-1)]\boldsymbol{\varphi}(t)\|^2}{r(t-1)}\leq$$

$$\sum_{t=1}^{\infty}\frac{2\|\boldsymbol{\eta}(t)-\boldsymbol{v}(t)\|^2}{r(t-1)}+\sum_{t=1}^{\infty}\frac{2\|[\tilde{\theta}^{\mathrm{T}}(t)-\tilde{\theta}^{\mathrm{T}}(t-1)]\boldsymbol{\varphi}(t)\|^2}{r(t-1)}\leq$$
$$\sum_{t=1}^{\infty}\frac{2\|\boldsymbol{y}(t)-\boldsymbol{v}(t)\|^2}{r(t-1)}+2\sum_{t=1}^{\infty}\frac{\|\boldsymbol{\varphi}(t)\|^2}{r(t-1)}\|\tilde{\theta}(t)-\tilde{\theta}(t-1)\|^2\leq$$
$$\sum_{t=1}^{\infty}\frac{2\|\boldsymbol{\eta}(t)-\boldsymbol{v}(t)\|^2}{r(t-1)}+2C_1\sum_{t=1}^{\infty}\|\tilde{\theta}(t)-\tilde{\theta}(t-1)\|^2=$$
$$C_2<\infty,\text{a.s.},\ C_1<\infty$$

Here, we have assumed that $\|\boldsymbol{\varphi}(t)\|^2\leq C_1 r(t-1)$. Using the Kronecker lemma gives

$$\lim_{t\to\infty}\frac{1}{r(t-1)}\sum_{j=1}^{t}\|\boldsymbol{e}(j)-\boldsymbol{v}(j)\|^2=0,\text{a.s.}$$

From (19), we have

$$\tilde{\theta}(t-i)=\tilde{\theta}(t)-\sum_{j=0}^{i-1}\frac{\boldsymbol{\varphi}(t-j)}{r(t-j)}\boldsymbol{e}^{\mathrm{T}}(t-j)\tag{20}$$

Replacing $t$ in (11) by $t-i$ yields

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

Using (20), we have

$$\boldsymbol{\varphi}^{\mathrm{T}}(t-i)\tilde{\theta}(t)=-\tilde{\boldsymbol{y}}^{\mathrm{T}}(t-i)+\boldsymbol{\varphi}^{\mathrm{T}}(t-i)\sum_{j=0}^{i-1}\frac{\boldsymbol{\varphi}(t-j)}{r(t-j)}\boldsymbol{e}^{\mathrm{T}}(t-j)$$

To some extent, the rest of the proof is similar to that of [13]. Squaring and using the relation, $(a+b)^2\leq 2(a^2+b^2)$, yield

$$\|\boldsymbol{\varphi}^{\mathrm{T}}(t-i)\tilde{\theta}(t)\|^2\leq 2\|\tilde{\boldsymbol{y}}(t-i)\|^2+2\|\boldsymbol{\varphi}(t-i)\|^2\times\left\|\sum_{j=0}^{i-1}\frac{\boldsymbol{\varphi}(t-j)}{r(t-j)}\{[\boldsymbol{e}(t-j)-\boldsymbol{v}(t-j)]+\boldsymbol{v}(t-j)\}^{\mathrm{T}}\right\|^2$$

Since $\boldsymbol{e}(t-j)-\boldsymbol{v}(t-j)$ is uncorrelated with $\boldsymbol{v}(t-j)$ and is $\boldsymbol{F}_{t-1}$ measurable, taking the conditional expectation on both sides with respect to $\boldsymbol{F}_{t-1}$ and using Assumptions 1 and 2 give

$$\mathrm{E}[\|\boldsymbol{\varphi}^{\mathrm{T}}(t-i)\tilde{\theta}(t)\|^2|\boldsymbol{F}_{t-1}]\leq 2\|\tilde{\boldsymbol{y}}(t-i)\|^2+2\|\boldsymbol{\varphi}(t-i)\|^2\sum_{j=0}^{i-1}\frac{\|\boldsymbol{\varphi}(t-j)\|^2}{r^2(t-j)}\times\{\|\boldsymbol{e}(t-j)-\boldsymbol{v}(t-j)\|^2+\sigma^2 r^{\epsilon}(t-1)\}$$

Summing for $i$ from 0 to $t-1$ on both sides and dividing by $r(t)$ yield

$$\frac{\mathrm{E}\{\mathrm{tr}[\tilde{\theta}^{\mathrm{T}}(t)Q(t)\tilde{\theta}(t)]|\boldsymbol{F}_{t-1}\}}{r(t)}=\frac{2}{r(t)}\sum_{i=1}^{t}\|\tilde{\boldsymbol{y}}(i)\|^2+S_1(t)+S_2(t)$$

where

$$S_1(t)=2\sum_{i=1}^{t-1}\frac{\|\boldsymbol{\varphi}(t-i)\|^2}{r(t)}\sum_{j=0}^{i-1}\frac{\|\boldsymbol{\varphi}(t-j)\|^2}{r^2(t-j)}\sigma^2 r^{\epsilon}(t-1)$$
$$S_2(t)=2\sum_{i=1}^{t-1}\frac{\|\boldsymbol{\varphi}(t-i)\|^2}{r(t)}\sum_{j=0}^{i-1}\frac{\|\boldsymbol{\varphi}(t-j)\|^2}{r^2(t-j)}\times\|\boldsymbol{e}(t-j)-\boldsymbol{v}(t-j)\|^2$$

Using Lemma 1, we have

$$S_1(t)=\frac{2}{r(t)}\sum_{i=2}^{t}\frac{[r(i-1)-r(0)]\|\boldsymbol{\varphi}(i)\|^2}{r^2(i)}\sigma^2 r^{\epsilon}(t-1)\leq\frac{2}{[r(t)]^{1-\epsilon}}\sum_{i=2}^{t}\frac{\|\boldsymbol{\varphi}(i)\|^2}{r(i)}\sigma^2\leq\frac{2\sigma^2\ln r(t)}{[r(t)]^{1-\epsilon}}\to 0,\text{a.s.}$$
$$S_2(t)=\frac{2}{r(t)}\sum_{i=2}^{t-1}\frac{[r(i-1)-r(0)]\|\boldsymbol{\varphi}(i)\|^2}{r^2(i)}\|\boldsymbol{e}(i)-\boldsymbol{v}(i)\|^2\leq\frac{2}{r(t)}\sum_{i=2}^{t}\frac{\|\boldsymbol{\varphi}(i)\|^2}{r(i)}\|\boldsymbol{e}(i)-\boldsymbol{v}(i)\|^2\leq\frac{2}{r(t)}\sum_{i=2}^{t}\|\boldsymbol{e}(i)-\boldsymbol{v}(i)\|^2\to 0,\text{a.s. as } t\to\infty$$

Hence,

$$\|\tilde{\theta}(t)\|^2=o\left(\frac{r(t)}{\lambda_{\min}[Q(t)]}\right),\text{a.s.}$$

This proves Theorem 1. □

The AM-SG algorithm has low computational burden, but its convergence is slow, just like the SG algorithm of scalar systems in [16]. To improve the convergence rate and tracking performance, we introduce a forgetting factor $\lambda$ in the AM-SG algorithm to have the AM-SG algorithm with a forgetting factor (the AM-FFSG algorithm for short) as follows:

$$\hat{\theta}(t)=\hat{\theta}(t-1)+\frac{\boldsymbol{\varphi}(t)}{r(t)}[\boldsymbol{y}^{\mathrm{T}}(t)-\boldsymbol{\varphi}^{\mathrm{T}}(t)\hat{\theta}(t-1)]\tag{21}$$
$$r(t)=\lambda r(t-1)+\|\boldsymbol{\varphi}(t)\|^2,\ 0<\lambda<1,\ r(0)=1\tag{22}$$
$$\boldsymbol{\varphi}(t)=[-\boldsymbol{\varphi}^{\mathrm{T}}(t-1)\hat{\theta}(t-1),\cdots,-\boldsymbol{\varphi}^{\mathrm{T}}(t-n_a)\hat{\theta}(t-n_a),\boldsymbol{u}^{\mathrm{T}}(t-1),\cdots,\boldsymbol{u}^{\mathrm{T}}(t-n_b)]^{\mathrm{T}}\tag{23}$$

When $\lambda=1$, the AM-FFSG algorithm reduces to the AM-SG algorithm; when $\lambda=0$, the AM-FFSG algorithm is the auxiliary model projection algorithm.

For comparison, the following equation gives the auxiliary model-based recursive least squares (AM-RLS) algorithm for estimating $\theta$:

$$\hat{\theta}(t)=\hat{\theta}(t-1)+P(t)\boldsymbol{\varphi}(t)[\boldsymbol{y}^{\mathrm{T}}(t)-\boldsymbol{\varphi}^{\mathrm{T}}(t)\hat{\theta}(t-1)]\tag{24}$$
$$P(t)=P(t-1)-\frac{P(t-1)\boldsymbol{\varphi}(t)\boldsymbol{\varphi}^{\mathrm{T}}(t)P(t-1)}{1+\boldsymbol{\varphi}^{\mathrm{T}}(t)P(t-1)\boldsymbol{\varphi}(t)}\tag{25}$$

## 4 Simulation tests

Consider the following 2-input and 2-output system (the output error system):

$$\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} -0.50 & 0.30 \\ 0.30 & -0.70 \end{bmatrix} \begin{bmatrix} x_1(t-1) \\ x_2(t-1) \end{bmatrix} = \begin{bmatrix} 2.00 & 0.80 \\ 0.60 & 1.50 \end{bmatrix} \begin{bmatrix} u_1(t-1) \\ u_2(t-1) \end{bmatrix}$$

$$\begin{bmatrix} y_1(t) \\ y_2(t) \end{bmatrix} = \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} v_1(t) \\ v_2(t) \end{bmatrix}$$

In simulation, the inputs $\{u_1(t)\}$ and $\{u_2(t)\}$ are taken as two independent persistent excitation sequences with zero mean and unit variances, and $v_1(t)$ and $v_2(t)$ as two white noise sequences with zero mean and variances $\sigma_1^2$ and $\sigma_2^2$. We apply the AM-SG, AM-FFSG, and AM-RLS algorithms to estimate the parameters of this system. The parameter estimates are shown in Tables 1 $\sim$ 3 and the estimation errors $\delta$ vs. $t$ are shown in Figs. 2 and 3, where $\delta = \|\hat{\theta}(t) - \theta\|/\|\theta\| \times 100\,\%$ is the parameter estimation error.

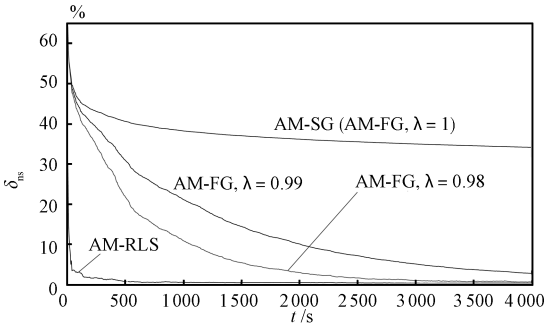


Fig. 2 The estimation errors $\delta$ vs. $t$ with different forgetting factors ($\sigma_1^2 = 0.20^2$ and $\sigma_2^2 = 0.20^2$)

Changing the noise variances $\sigma_1^2$ and $\sigma_2^2$ can adjust the noise-to-signal ratios $\delta_{\rm ns}(1)$ and $\delta_{\rm ns}(2)$ of two output channels. When $\sigma_1^2 = 0.20^2$ and $\sigma_2^2 = 0.20^2$, the noise-to-signal ratios are $\delta_{\rm ns}(1) = 7.61\,\%$ and $\delta_{\rm ns}(2) = 7.99\,\%$; when $\sigma_1^2 = 1.00^2$ and $\sigma_2^2 = 1.00^2$, the noise-to-signal ratios are $\delta_{\rm ns}(1) = 38.06\,\%$ and $\delta_{\rm ns}(2) = 39.96\,\%$.

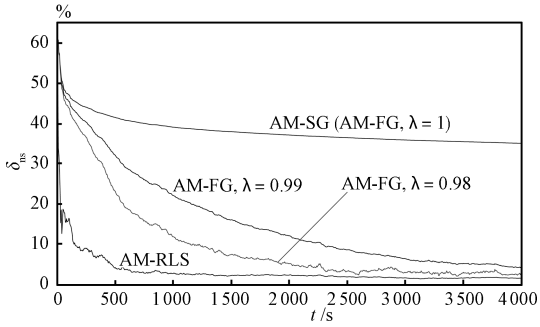


Fig. 3 The estimation errors $\delta$ vs. $t$ with different forgetting factors ($\sigma_1^2 = 1.00^2$ and $\sigma_2^2 = 1.00^2$)

From the simulation results of Tables 1 $\sim$ 3 and Figs. 2 and 3, we can draw the following conclusions: 1) A lower noise level leads to a faster rate of convergence of the parameter estimates to the true parameters; 2) As long as an appropriate forgetting factor is chosen, the faster convergence rate can be achieved and the smaller estimation errors may be obtained; 3) The estimation errors $\delta$ become smaller (in general) as the data length $t$ increases. In other words, increasing data length generally results in smaller parameter estimation errors; 4) If we choose an appropriate forgetting factor, the parameter estimation error of the AM-FFSG algorithm is very close to that of the AM-RLS algorithm; 5) These show the effectiveness of the proposed algorithms.

Table 1 The AM-SG estimates and errors ($\sigma_1^2 = 0.20^2$ and $\sigma_2^2 = 0.20^2$)

| $t$ | $a_{11}$ | $a_{12}$ | $b_{11}$ | $b_{12}$ | $a_{21}$ | $a_{22}$ | $b_{21}$ | $b_{22}$ | $\delta$ (%) |
|---|---|---|---|---|---|---|---|---|---|
| 100 | −0.59922 | 0.17140 | 1.48708 | −0.30071 | 0.33764 | −0.71091 | 0.31219 | 1.11802 | 46.03369 |
| 200 | −0.60059 | 0.17094 | 1.51063 | −0.24293 | 0.31021 | −0.69470 | 0.32900 | 1.13286 | 43.72964 |
| 500 | −0.59547 | 0.16413 | 1.54290 | −0.16607 | 0.30202 | −0.68306 | 0.34396 | 1.16092 | 40.65912 |
| 1 000 | −0.59658 | 0.16547 | 1.56834 | −0.10671 | 0.29320 | −0.68347 | 0.35884 | 1.18001 | 38.27283 |
| 2 000 | −0.59529 | 0.16723 | 1.59320 | −0.05864 | 0.29220 | −0.69037 | 0.37232 | 1.19823 | 36.23175 |
| 3 000 | −0.59712 | 0.16900 | 1.60756 | −0.02985 | 0.29159 | −0.69500 | 0.38084 | 1.20899 | 35.02159 |
| 4 000 | −0.59702 | 0.17187 | 1.61694 | −0.00982 | 0.29265 | −0.69681 | 0.38608 | 1.21616 | 34.18509 |
| True values | −0.50000 | 0.30000 | 2.00000 | 0.80000 | 0.30000 | −0.70000 | 0.60000 | 1.50000 | |

Table 2 The AM-FFSG estimates and errors with $\lambda = 0.99$ and $\lambda = 0.98$ ($\sigma_1^2 = 0.20^2$ and $\sigma_2^2 = 0.20^2$)

| $\lambda$ | $t$ | $a_{11}$ | $a_{12}$ | $b_{11}$ | $b_{12}$ | $a_{21}$ | $a_{22}$ | $b_{21}$ | $b_{22}$ | $\delta$ (%) |
|---|---|---|---|---|---|---|---|---|---|---|
| | 100 | −0.60777 | 0.17692 | 1.51146 | −0.25885 | 0.32854 | −0.70781 | 0.31572 | 1.14094 | 44.21498 |
| | 200 | −0.61377 | 0.18746 | 1.55035 | −0.16457 | 0.29378 | −0.70023 | 0.34521 | 1.16656 | 40.41276 |
| | 500 | −0.57355 | 0.18923 | 1.65060 | 0.06882 | 0.30173 | −0.68181 | 0.39410 | 1.24891 | 30.90156 |
| 0.99 | 1 000 | −0.55986 | 0.22650 | 1.75541 | 0.30436 | 0.28222 | −0.69776 | 0.45650 | 1.32641 | 21.15147 |
| | 2 000 | −0.52576 | 0.26546 | 1.88925 | 0.55857 | 0.29103 | −0.70166 | 0.53029 | 1.42339 | 10.09261 |
| | 3 000 | −0.51682 | 0.28012 | 1.94452 | 0.67591 | 0.29855 | −0.69931 | 0.57385 | 1.46477 | 5.07976 |
| | 4 000 | −0.50970 | 0.29168 | 1.97003 | 0.73266 | 0.30409 | −0.69537 | 0.58706 | 1.48132 | 2.74364 |
| | 100 | −0.61258 | 0.18144 | 1.53585 | −0.21435 | 0.31580 | −0.69947 | 0.32049 | 1.16407 | 42.30653 |
| | 200 | −0.61699 | 0.20207 | 1.59079 | −0.07974 | 0.27566 | −0.70259 | 0.36541 | 1.20042 | 36.86039 |
| | 500 | −0.54834 | 0.23167 | 1.75031 | 0.28612 | 0.31004 | −0.68227 | 0.44480 | 1.32795 | 21.77915 |
| 0.98 | 1 000 | −0.53044 | 0.26963 | 1.87190 | 0.54686 | 0.28936 | −0.70362 | 0.51873 | 1.41278 | 10.87786 |
| | 2 000 | −0.50726 | 0.28972 | 1.96794 | 0.72901 | 0.29620 | −0.69964 | 0.57536 | 1.48445 | 2.94576 |
| | 3 000 | −0.50522 | 0.29662 | 1.98799 | 0.77670 | 0.30052 | −0.69631 | 0.60452 | 1.50000 | 0.96454 |
| | 4 000 | −0.50621 | 0.29923 | 1.99316 | 0.78825 | 0.30504 | −0.69347 | 0.60168 | 1.49924 | 0.60126 |
| True values | | −0.50000 | 0.30000 | 2.00000 | 0.80000 | 0.30000 | −0.70000 | 0.60000 | 1.50000 | |

Table 3 The AM-RLS estimates and errors ($\sigma_1^2 = 0.20^2$ and $\sigma_2^2 = 0.20^2$)

| $t$ | $a_{11}$ | $a_{12}$ | $b_{11}$ | $b_{12}$ | $a_{21}$ | $a_{22}$ | $b_{21}$ | $b_{22}$ | $\delta$ (%) |
|---|---|---|---|---|---|---|---|---|---|
| 100 | −0.50892 | 0.28087 | 2.02473 | 0.73734 | 0.29262 | −0.69484 | 0.54565 | 1.48989 | 3.15273 |
| 200 | −0.50403 | 0.30380 | 2.00898 | 0.77285 | 0.28644 | −0.70340 | 0.57779 | 1.47845 | 1.56517 |
| 500 | −0.50247 | 0.30326 | 2.00902 | 0.78733 | 0.29335 | −0.69865 | 0.59075 | 1.49674 | 0.70049 |
| 1 000 | −0.49955 | 0.30134 | 2.00597 | 0.79384 | 0.29455 | −0.70053 | 0.59248 | 1.49277 | 0.51227 |
| 2 000 | −0.50093 | 0.29993 | 1.99891 | 0.78932 | 0.29720 | −0.70082 | 0.59432 | 1.49749 | 0.44703 |
| 3 000 | −0.50041 | 0.29945 | 1.99776 | 0.79155 | 0.29832 | −0.70002 | 0.60076 | 1.49984 | 0.31355 |
| 4 000 | −0.50115 | 0.29986 | 1.99770 | 0.79179 | 0.30042 | −0.70001 | 0.60161 | 1.49924 | 0.30776 |
| True values | −0.50000 | 0.30000 | 2.00000 | 0.80000 | 0.30000 | −0.70000 | 0.60000 | 1.50000 | |

## 5 Conclusions

Using the auxiliary model technique, the auxiliary model-based stochastic gradient algorithms are presented for MIMO systems. The convergence of the proposed algorithm is analyzed by using the martingale convergence theorem. The simulation results show that the proposed algorithms are effective. The proposed methods can be extended to output error moving average (OEMA) systems[17] and non-uniform sampled systems[18].

### References


1 Zheng W X. A bias correction method for identification of linear dynamic errors-in-variables models. *IEEE Transactions on Automatic Control*, 2002, **47**(7): 1142−1147

2 Yang Hui-Zhong, Zhang Yong. Comparisons of bias compensation methods and other identification approaches for Box-Jenkins models. *Control Theory and Applications*, 2007, **24**(2): 215−222 (in Chinese)

3 Zhang Yong, Yang Hui-Zhong. Bias compensation recursive least squares identification for output error systems with colored noises. *Acta Automatica Sinica*, 2007, **33**(10): 1053−1060 (in Chinese)

4 Zong Qun, Dou Li-Qian, Liu Wen-Jing. Iterative identification and control design based on Vinnicombe distance. *Acta Automatica Sinica*, 2008, **34**(11): 1431−1436 (in Chinese)

5 Zhong Lu-Sheng, Song Zhi-Huan. Hierarchical optimization identification of LTI state-space systems by projected gradient search. *Acta Automatica Sinica*, 2008, **34**(6): 711−715 (in Chinese)

6 Wang Dong-Qing. Recursive extended least squares identification method based on auxiliary models. *Control Theory and Applications*, 2009, **26**(1): 51−56 (in Chinese)

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

8 Ding F, Chen T. Performance analysis of multi-innovation gradient type identification methods. *Automatica*, 2007, **43**(1): 1−14

9 Wang Dong-Qing, Ding Feng. Auxiliary models based multi-innovation generalized extended stochastic gradient algorithms. *Control and Decision*, 2008, **23**(9): 999−1003 (in Chinese)

10 Wang D Q, Ding F. Extended stochastic gradient identification algorithms for hammerstein-wiener ARMAX systems. *Computers and Mathematics with Applications*, 2008, **56**(12): 3157−3164

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

12 Ding F, Chen T. Hierarchical least squares identification methods for multivariable systems. *IEEE Transactions on Automatic Control*, 2005, **50**(3): 397−401

13 Ding F, Yang H Z, Liu F. Performance analysis of stochastic gradient algorithms under weak conditions. *Science in China Series F: Information Sciences*, 2008, **51**(9): 1269−1280

14 Kailath T. *Linear Systems*. New Jersey: Prentice-Hall, 1980

15 Solo V. The convergence of AML. *IEEE Transactions on Automatic Control*, 1979, **24**(6): 958−962

16 Goodwin G C, Sin K S. *Adaptive Filtering, Prediction and Control*. New Jersey: Prentice-Hall, 1984

17 Ding F, Liu P X, Liu G. Auxiliary model based multi-innovation extended stochastic gradient parameter estimation with colored measurement noises. *Signal Processing*, 2009, **89**(10): 1883−1890

18 Ding F, Qiu L, Chen T. Reconstruction of continuous-time systems from their non-uniformly sampled discrete-time systems. *Automatica*, 2009, **45**(2): 324−332



**DING Feng** Received his B. Sc. degree from Hubei University of Technology in 1984, and his 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 Hubei Pharmaceutical Factory. From 1994 to 2002, he was with the Department of Automation, Tsinghua University and was a post-doctoral fellow and research associate at University of Alberta, Canada from 2002 to 2005.

He was a visiting professor in the Department of Systems and Computer Engineering, Carleton University, Canada from May to December 2008 and a research associate in the Department of Aerospace Engineering, Ryerson University, Canada, from January to October 2009.

He has been a professor at the School of Communication and Control Engineering, Jiangnan University since 2004. His research interest covers model identification and adaptive control. Corresponding author of this paper.
E-mail: fding@jiangnan.edu.cn



**LIU Xiao-Ping** Received his B. Sc. and M. Sc. degrees from Northern Jiaotong University, China in 1992 and 1995, respectively, and his Ph. D. degree from the University of Alberta, Canada in 2002. He has been with the Department of Systems and Computer Engineering, Carleton University, Canada since July 2002 and he is currently a research chair professor. His research interest covers interactive networked systems and teleoperation, haptics, robotics, intelligent systems, context-aware intelligent networks, and their applications to biomedical engineering.
E-mail: xpliu@sce.carleton.ca