Available online at www.sciencedirect.com

SCIENCE DIRECT®

Automatica 42 (2006) 535–547

automatica

www.elsevier.com/locate/automatica

Joint identification of plant rational models and noise distribution functions using binary-valued observations☆

Le Yi Wang[a,*], G. George Yin[b], Ji-Feng Zhang[c]

[a] *Department of Electrical and Computer Engineering, Wayne State University, Detroit, MI 48202, USA*
[b] *Department of Mathematics, Wayne State University, Detroit, MI 48202, USA*
[c] *LSC, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, China*

Received 12 April 2005; received in revised form 8 October 2005; accepted 5 December 2005

Abstract

System identification of plants with binary-valued output observations is of importance in understanding modeling capability and limitations for systems with limited sensor information, establishing relationships between communication resource limitations and identification complexity, and studying sensor networks. This paper resolves two issues arising in such system identification problems. First, regression structures for identifying a rational model contain non-smooth nonlinearities, leading to a difficult nonlinear filtering problem. By introducing a two-step identification procedure that employs periodic signals, empirical measures, and identifiability features, rational models can be identified without resorting to complicated nonlinear searching algorithms. Second, by formulating a joint identification problem, we are able to accommodate scenarios in which noise distribution functions are unknown. Convergence of parameter estimates is established. Recursive algorithms for joint identification and their key properties are further developed.

© 2006 Elsevier Ltd. All rights reserved.

Keywords: System identification; Estimation; Binary-valued observations; Identifiability; Parameter convergence; Recursive algorithms

1. Introduction

System identification of plants with binary-valued output observations is of importance in understanding modeling capability for systems with limited sensor information, establishing relationships between communication resource limitations and identification complexity, and studying sensor networks. The authors introduced in Wang, Zhang, and Yin (2003) a framework in which such identification problems can be rigorously pursued either in a stochastic setting or a worst-case scenario. This paper extends the results in two key directions: (a) systems in Wang et al. (2003) are finite impulse response models. Due to nonlinearity in output observations, switching (non-smooth) nonlinearity enters the regressor for rational models, leading to a difficult problem of nonlinear filtering. By introducing a two-step identification procedure that employs periodic signals, empirical measures, and identifiability features, rational models can be identified without resorting to complicated nonlinear searching algorithms. (b) Identification algorithms in Wang et al. (2003) assume knowledge of noise distribution functions. Unlike traditional identification problems where actual noise distribution functions are usually not used in the algorithms, the identification algorithms for binary-valued observations use explicitly the noise distribution functions. Consequently, they do not apply if noise distribution functions are unknown. Since in practice noise distribution functions are either unknown or only estimated with limited prior information, removing this condition is of essential importance. By formulating a joint identification problem, we are able to accommodate the situations in which noise distribution functions are unknown. Identification errors and input design are examined in a stochastic information framework. Convergence of parameter estimates is established. Recursive algorithms for joint identification and their basic properties are further derived.

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

* Corresponding author.

E-mail addresses: lywang@wayne.edu (L.Y. Wang), gyin@math.wayne.edu (G.G. Yin), jif@iss.ac.cn (J-F. Zhang).

0005-1098/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.automatica.2005.12.004

This work is based on the premise that the order of the system is finite and known. For infinite dimensional systems that are approximated by finite dimensional models, the problem of unmodeled dynamics and model complexity becomes an essential issue. This has been studied in Wang (1997); Wang and Yin (1999, 2000, 2002) for system identification with regular sensors, and in Wang et al. (2003) for system identification with binary-valued sensors. Estimating the order of the system is a worthwhile direction, but beyond the scope of this paper.

The paper is organized as follows. The main problem is formulated in Section 2. Our development starts in Section 3 with estimation of plant outputs when noise distribution functions are known. The main tool is empirical measures and their convergence. Section 4 establishes the main results on identifiability of plant parameters. A basic property of rational systems is established. It shows that if the input is periodic and full rank, system parameters are uniquely determined by its periodic outputs. Consequently, under such inputs, convergence of parameter estimates can be established when the convergence results of Section 3 are utilized. Section 5 is devoted to the general scenario where noise distribution functions are unknown and must be estimated. Identification of distribution functions and system parameters are intimately intertwined. Together, they form a nonlinear identification problem. Algorithms for identifying jointly plant parameters and distribution functions are introduced. It is shown that under some mild conditions, convergence of both estimates can be established when one uses signal scaling and threshold shifting to leverage on providing excitation for parameter estimation. A simple application example is given in Section 6 to summarize the main steps of identification experimental design, identification algorithms, and accuracy evaluation developed in this paper. For computational efficiency, recursive algorithms for joint identification are presented in Section 7. Some brief concluding remarks are made in Section 8.

For some related but different identification algorithms such as binary reinforcement and some applications, the reader is referred to Caianiello and de Luca (1966), Chen and Yin (2003), Elvitch, Sethares, Rey, and Johnson (1989), Eweda (1995), Gersho (1984), Pakdaman and Malta (1998), Yin, Krishnamurthy, and Ion (2003). The main tools for stochastic analysis and identification methodologies can be found in Billingsley (1968), Chen and Guo (1991), Feller (1968, 1971), Kushner and Yin (2003), Ljung (1987), Pollard (1984), Serfling (1980). This paper is a continuation of the authors' early work in Wang et al. (2003), Wang (1997); Wang and Yin (1999), Wang and Yin (2000, 2002).

2. Problem formulation

Consider the following system:

$$y_k = G(q)u_k + d_k = x_k + d_k, \tag{1}$$

which is in an *output error* form. Here, q is the one-step shift operator $qu_k = u_{k-1}$; $\{d_k\}$ is a sequence of random noise (sensor noise); $x_k = G(q)u_k$ is the noise-free output of the system; $G(q)$ is a stable rational function of q

$$G(q) = \frac{B(q)}{1-A(q)} = \frac{b_1 q + \cdots + b_n q^n}{1-(a_1 q + \cdots + a_n q^n)}.$$

The input $\{u_k\}$ is uniformly bounded by $|u_k| \leqslant u_{\max}$ and can be selected by the designer otherwise. The observation $\{y_k\}$ is measured by a binary-valued sensor of threshold $C > 0$,

$$s_k = I_{\{y_k \leqslant C\}} = \begin{cases} 1, & y_k \leqslant C, \\ 0, & y_k > C, \end{cases} \tag{2}$$

where I_A is the indicate function of the set A. The parameter $\theta = [a_1, \ldots, a_n, b_1, \ldots, b_n]^{\mathrm{T}}$ is to be identified.

For system identification, the system (1) is commonly expressed in its regression form

$$y_k = A(q)y_k + B(q)u_k + (1-A(q))d_k = \psi_k^{\mathrm{T}}\theta + \widetilde{d}_k, \tag{3}$$

where $\psi_k^{\mathrm{T}} = [y_{k-1}, \ldots, y_{k-n}, u_{k-1}, \ldots, u_{k-n}]$, and $\widetilde{d}_k = (1-A(q))d_k$. The *equation error* $\{\widetilde{d}_k\}$ may not be independent even if $\{d_k\}$ is.

Most identification algorithms, especially recursive ones, have been developed from the observation structure (3). Direct application of this structure in our problem encounters a daunting difficulty since y_k is not directly measured. Using s_k in this structure inevitably introduces nonlinearities that make it harder to design feasible algorithms and to establish their fundamental properties such as convergence, accuracy and robustness.

In this paper, we develop a new approach that involves two steps: (i) first, x_k is identified on the basis of s_k; (ii) θ is identified from the input u_k and estimated x_k, using the structure (3). The first step is accomplished by using periodic inputs and empirical measures. The second step is validated by using identifiability arguments and computed by recursive algorithms. Convergence of the algorithms will be derived. This approach will first be presented for the case of known noise distributions in Sections 3 and 4. It will then be extended to handle unknown noise distributions in Section 5.

3. Estimation of x_k: known noise distribution

To estimate x_k, select u_k to be $2n$-periodic. Then the noise-free output $x_k = G(q)u_k$ is also $2n$-periodic, after a short transient duration.[1] Hence, $x_{j+2ln} = x_j$, for any positive integer l. $\{x_k\}$ will be determined entirely by $2n$ unknown real numbers γ^j, $j = 1, \ldots, 2n$,

$$x_j = \gamma^j, \quad j = 1, \ldots, 2n. \tag{4}$$

$\Gamma = [\gamma^1, \ldots, \gamma^{2n}]^{\mathrm{T}}$ are to be estimated.

[1] Since the system is assumed to be stable, all transient modes decay exponentially, much faster than the convergence of the empirical measures. As a consequence, their impact is negligible and will be ignored in the analysis.

Consequently, for each $j=1,\ldots,2n$ the observations can be expressed as

$$y_{j+2ln}=x_{j+2ln}+d_{j+2ln}=\gamma^j+d_{j+2ln}, \quad l=0,1,\ldots. \tag{5}$$

When y_k in (5) is directly measured and $\{d_k\}$ is a sequence of independent and identically distributed (i.i.d.) random variables with zero mean, estimating γ^j is easily achieved by averaging. Complication arises when y_k is only measured by binary observations. This paper resorts to empirical measures for resolving this problem.

3.1. Empirical measures

Assumption A1. $\{d_k\}$ is a sequence of i.i.d. random variables whose distribution function $F(\cdot)$ and its inverse $F^{-1}(\cdot)$ are continuous and known. If the distribution $F(\cdot)$ has a density with a finite support, then $F(\cdot)$ and $F^{-1}(\cdot)$ are both continuous in the interior of this finite set. The moment generating function of d_1 exists.

Note that in Assumption A1, we have assumed that the noise has a continuous distribution function, the distribution function is invertible, and the inverse is also continuous. A typical example of the noise that satisfies Assumption A1 is a sequence of Gaussian random variables. The second part of the assumption deals with random variables whose distribution function is only invertible in a finite interval $[-\delta,\delta]$ (e.g., uniform distribution). In this case, we require both the distribution and its inverse be continuous on $(-\delta,\delta)$. The continuity assumption on the distribution function of the noise is not a restriction. When one deals with discrete random variables, suitable scaling and the well-known central limit theorem lead to normal approximation.

Relationship (5) indicates that for a fixed j, γ^j is an unknown constant, and empirical measures can be calculated with respect to index l. Let the observation length be $N=2nm$ for some positive integer m. For a given $j=1,\ldots,2n$, define

$$\xi_m^j=\frac{1}{m}\sum_{l=0}^{m-1}s_{j+2ln}. \tag{6}$$

Note that the event $\{s_{j+2ln}=1\}=\{y_{j+2ln}\leqslant C\}$ is the same as the event $\{d_{j+2ln}\leqslant C-\gamma^j\}$. Then ξ_m^j is precisely the value of the m-sample empirical distribution $\widehat{F}_m(z)$ of the noise d at $z=C-\gamma^j$.

The well-known Glivenko–Cantelli Theorem (Billingsley, 1968, p. 103), guarantees convergence of ξ_m^j. These results are listed in Lemma 1 without proof.

Lemma 1 (*Billingsley, 1968; Pollard, 1984*). *Under Assumption* A1, (a) *for any compact subset* $S\subset\mathbb{R}$,

$$\lim_{m\to\infty}\sup_{z\in S}|\widehat{F}_m(z)-F(z)|\to 0,\ w.p.1;$$

(b) *let* $\widehat{K}_m(z)=\sqrt{m}(\widehat{F}_m(z)-F(z))$, *for each* $z\in S$. *Then* $\widehat{K}_m(\cdot)$ *converges weakly to* $K(\cdot)$, *a stretched Brownian bridge process such that the covariance of* $K(\cdot)$ *is given by* $EK(z_1)K(z_2)=\min\{F(z_1),F(z_2)\}-F(z_1)F(z_2)$ *for* z_1 *and* $z_2\in S$.

The uniform convergence in Lemma 1 is stronger than what is needed in this paper. Point-wise convergence of $\widehat{F}_m(z)$ at $z=C-\gamma^j$, $j=1,\ldots,2n$ will suffice.

Note that a Brownian bridge is a function of a Brownian motion defined on [0, 1]. Loosely speaking, it is a Brownian motion tied down at both end points of the interval [0, 1]. Between the two end points, the process evolves just as a Brownian motion. Now in the current case, since γ^j can take real values outside [0, 1], the Brownian bridge becomes a stretched one. The terminology "stretched Brownian bridge" follows from that of Pollard (1984).

Example 1. To illustrate the convergence of empirical measures, consider a uniformly distributed noise on $[-1.2, 1.2]$. The actual distribution function is $F(z)=(z+1.2)/2.4$. For different values of z, Fig. 1 shows convergence of the empirical measures at several points in $[-1.2, 1.2]$ when the sample size is increased gradually from $N=20$ to $N=1000$.

3.2. Estimation of γ^j

To proceed, we first construct an estimate of γ^j, which will then be used to identify the system parameter θ. Since $F(\cdot)$ is invertible and known, we define

$$\widehat{\gamma}_m^j=C-F^{-1}(\xi_m^j). \tag{7}$$

Theorem 1. *Under Assumption* A1,

$$\widehat{\gamma}_m^j\to\gamma^j, \quad w.p.1 \text{ as } m\to\infty.$$

Proof. By Lemma 1, as $m\to\infty$,

$$\xi_m^j\to F(C-\gamma^j), \quad \text{w.p.1.}$$

Hence, continuity of $F^{-1}(\cdot)$ implies that $F^{-1}(\widehat{F}_m(C-\gamma^j))\to C-\gamma^j$ w.p.1. Therefore,

$$F^{-1}(\xi_m^j)\to C-\gamma^j, \quad \text{w.p.1,}$$

or equivalently $C-F^{-1}(\xi_m^j)\to\gamma^j$ w.p.1. □

Example 2. Consider the case $\gamma^j=2.1$: $y_k=2.1+d_k$ and the sensor threshold $C=3.5$. The disturbance is uniformly distributed in $[-2, 2]$. Fig. 2 shows estimates of γ^j as a function of sample sizes.

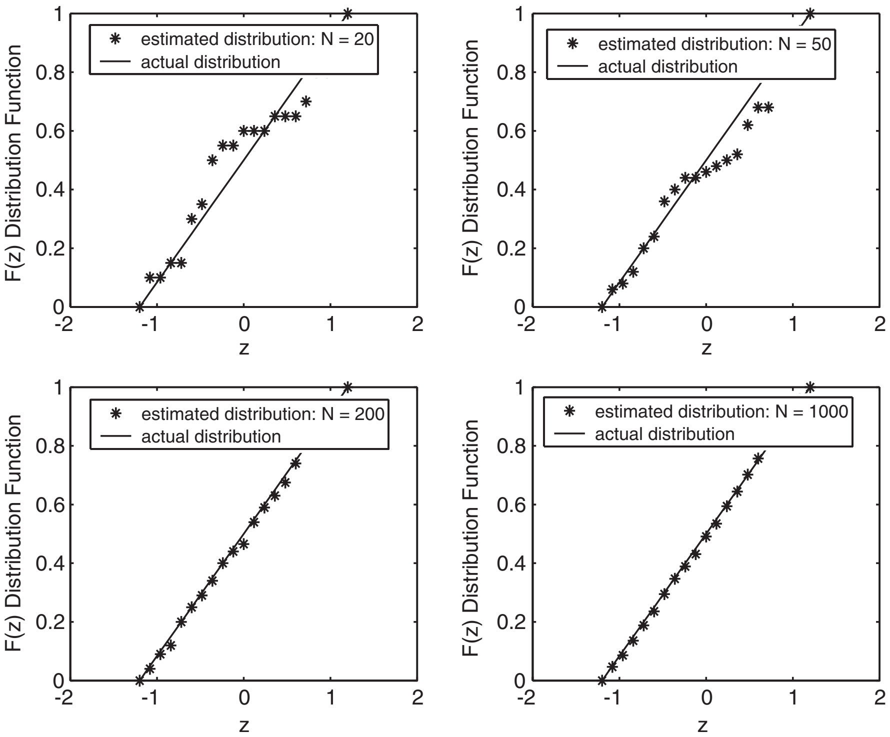

Fig. 1. Convergence of empirical measures on different setpoints.

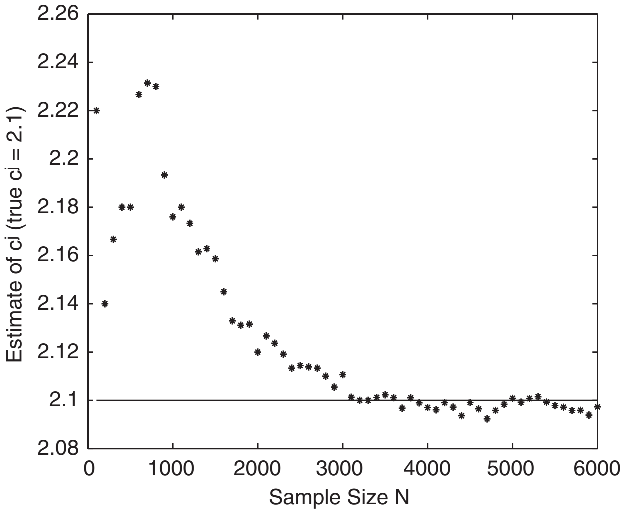

Fig. 2. Convergence of estimates of γ^j.

4. Estimation of parameter θ

Under a periodic input u, the one-to-one mapping between θ and the periodic output x of the system G will be first established. This relationship will be used to derive an estimate of θ from that of x.

4.1. Parameter identifiability

Recall that

$$x_k = G(q)u_k = \frac{b_1 q + \cdots + b_n q^n}{1-(a_1 q + \cdots + a_n q^n)} u_k$$

or in a regression form

$$x_k = \phi_k^{\mathrm{T}} \theta, \tag{8}$$

where $\phi_k^{\mathrm{T}} = [x_{k-1}, \ldots, x_{k-n}, u_{k-1}, \ldots, u_{k-n}]$, and $\theta = [a_1, \ldots, a_n, b_1, \ldots, b_n]^{\mathrm{T}}$. Then under a $2n$-periodic input, the noise-free output x and system parameters θ are related by

$$X = \Phi\theta$$

with

$$X = [x_{k_0}, \ldots, x_{k_0+2n-1}]^{\mathrm{T}},$$
$$\Phi = [\phi_{k_0}, \ldots, \phi_{k_0+2n-1}]^{\mathrm{T}}. \tag{9}$$

Apparently, if Φ is full rank, then there is a one-to-one correspondence between x and θ.

Since Φ contains both input u_k and output x_k, in general, the invertibility of Φ depends on both u_k and x_k, hence on the true (but unknown) plant $G(q)$. Furthermore, the invertibility may also vary with the starting time k_0. However, it will be shown that such complications dissipate when u_k is $2n$-periodic.

Definition 1. A periodic signal v_t of period l is said to be full rank if its discrete Fourier transform $V(\omega_k) = \sum_{t=1}^{l} v_t \mathrm{e}^{-\mathrm{i}\omega_k t}/\sqrt{l}$ is nonzero at $\omega_k = 2\pi k/l$, $k = 1, \ldots, l$.

Theorem 2. *Suppose that the pair $D(q) = 1 - A(q)$ and $B(q)$ are coprime. If u_k is $2n$-periodic and full rank, then*

(a) *Φ given by* (9) *is invertible for all* k_0.
(b) $\|\Phi^{-1}\|$ *is independent of* k_0, *where* $\|\cdot\|$ *is the largest singular value. Hence*, $\mu = \|\Phi^{-1}\| < \infty$ *is a constant for all* k_0.

Proof. (a) The proof will follow from some arguments of identifiability.

The true plant $G(q)$ is of order n with transfer function $G(q) = (b_1 q + \cdots + b_n q^n)/(1 - a_1 q - \cdots - a_n q^n) = (B(q))/(D(q))$, where $D(q)$ and $B(q)$ are coprime polynomials. The observation equation is $X = \Phi\theta$. Obviously, Φ is invertible if and only if θ can be uniquely determined from the observation equation. Assume that there exists another nth-order system $\widetilde{G}(q) = \widetilde{B}(q)/\widetilde{D}(q)$, with $\widetilde{D}$ and $\widetilde{B}$ coprime and $\widetilde{D}(0) = 1$, also satisfying the observation. In particular, $\widetilde{x}_k = (\widetilde{G}u)_k = x_k$, for $k = 1, \ldots, 2n$. Define

$$\Delta(q) = G(q) - \widetilde{G}(q) = \frac{B(q)\widetilde{D}(q) - \widetilde{B}(q)D(q)}{D(q)\widetilde{D}(q)} := \frac{qN(q)}{R(q)},$$

where $R(q)$ is a polynomial of order $2n$ and $N(q)$ a polynomial of order $2n - 1$.

For the given $2n$-periodic input u, by hypothesis we have $h_k = (\Delta u)_k = 0$, $k = 1, \ldots, 2n$. It follows that $\widetilde{H}(\omega) = (1/\sqrt{2n})\sum_{k=1}^{2n} h_k \mathrm{e}^{-\mathrm{i}\omega k} = 0$. On the other hand, by frequency-domain analysis of the system $\widetilde{H}(\omega) = \Delta(\mathrm{e}^{\mathrm{i}\omega})U(\omega) + R(\omega)$, where $U(\omega) = (1/\sqrt{2n})\sum_{k=1}^{2n} u_k \mathrm{e}^{-\mathrm{i}\omega k}$ and $R(\omega) = 0$, for $\omega = 2\pi j/(2n)$, $j = 1, \ldots, 2n$. By hypothesis, $U(\omega) \neq 0$, for $\omega = 2\pi j/(2n)$, $j = 1, \ldots, 2n$. Hence, $\Delta(\mathrm{e}^{\mathrm{i}\omega}) = 0$, for $\omega = 2\pi j/(2n)$, $j = 1, \ldots, 2n$. However, since $N(q)$ is of order $2n - 1$, if $\Delta \not\equiv 0$, $\Delta(\mathrm{e}^{\mathrm{i}\omega})$ can have maximum $2n - 1$ finite zeros. Consequently, $\Delta(q) \equiv 0$, i.e., $G(q) \equiv \widetilde{G}(q)$. Now, this equality, together with the coprimeness of $G(q)$ and $\widetilde{G}(q)$, implies that there exists a constant c for which $B(q) = c\widetilde{B}(q)$ and $D(q) = c\widetilde{D}(q)$. Finally, $D(0) = \widetilde{D}(0) = 1$ implies $c = 1$. Therefore, $B(q) = \widetilde{B}(q)$, $D(q) = \widetilde{D}(q)$. Namely, $B(q)$ and $D(q)$, or equivalently θ, are uniquely determined by the observation equation.

(b) For Φ given in (9), to emphasize on the dependence of Φ on k_0, we write it as $\Phi(k_0)$. To show that $\|\Phi^{-1}(k_0)\|$ is independent of k_0, we observe that since both u_k and x_k are $2n$-periodic, $\Phi(k_0 + 1) = J\Phi(k_0)$, where

$$J = \begin{bmatrix} 0 & I_{(2n-1)\times(2n-1)} \\ 1 & 0 \end{bmatrix}$$

is a $(2n) \times (2n)$ unitary matrix obtained by permuting the rows of the identity matrix. As a result, $\|\Phi^{-1}(k_0)\| = \|J\Phi^{-1}(k_0+1)\| = \|\Phi^{-1}(k_0+1)\|$ since the norm $\|\cdot\|$ is unitary invariant. □

Example 3. Suppose that the true system has the transfer function $G(p) = (q + 0.5q^2)/(1 - 0.5q + 0.2q^2)$. Hence, the true plant has the regression model

$$x_k = 0.5x_{k-1} - 0.2x_{k-2} + u_{k-1} + 0.5u_{k-2}.$$

Since the order of the system is $n = 2$, we select the input to be 4-periodic with $u_1 = 1$, $u_2 = -0.2$, $u_3 = 1.5$, $u_4 = -0.1$. For a selected $k_0 = 20$,

$$\Phi = \begin{bmatrix} -1.4884 & -0.6889 & -0.1000 & 1.5000 \\ -1.2564 & -1.4884 & 1.0000 & -0.1000 \\ -1.2805 & -1.2564 & -0.2000 & 1.0000 \\ -0.6890 & -1.2805 & 1.5000 & -0.2000 \end{bmatrix},$$

$$\Phi^{-1} = \begin{bmatrix} -0.8079 & -1.9384 & 1.3004 & 1.4118 \\ 1.0345 & 0.7749 & -1.6066 & -0.6619 \\ 0.5624 & -0.4417 & -0.7069 & 0.9044 \\ 0.3776 & -1.5969 & 0.5053 & 1.1572 \end{bmatrix},$$

and $\|\Phi^{-1}\| = 3.8708$. It can be verified that for different k_0, Φ will be different only by permutation of its rows. Consequently, $\|\Phi^{-1}\| = 3.8708$ for all k_0.

4.2. Identification algorithms for θ and convergence analysis

For each $j = 1, \ldots, 2n$, the estimate $\widehat{\gamma}_m^j$ of γ^j can be written as

$$\widehat{\gamma}_m^j = \gamma^j + e_m^j,$$

where, by Theorem 1, $e_m^j \to 0$, w.p.1 as $m \to \infty$.

Define an estimated $2n$-periodic output sequence of $G(q)$ by

$$\widehat{x}_{j+2ln} = \widehat{x}_j = \widehat{\gamma}_m^j, \tag{10}$$

for $j = 1, \ldots, 2n$, and $l = 1, ..., m - 1$. Then

$$\widehat{x}_{j+2ln} = x_{j+2ln} + e_m^j.$$

To estimate the parameter θ we use $\widehat{x}_k$ in place of x_k in (8)

$$\widehat{x}_k = \widehat{\phi}_k^{\mathrm{T}}\widehat{\theta}_m,$$

where $\widehat{\phi}_k^{\mathrm{T}} = [\widehat{x}_{k-1}, \ldots, \widehat{x}_{k-n}, u_{k-1}, \ldots, u_{k-n}]$. Then

$$\widehat{X} = \widehat{\Phi}\widehat{\theta}_m \tag{11}$$

for the estimated system, where $\widehat{X} = [\widehat{x}_{k_0}, \ldots, \widehat{x}_{k_0+2n-1}]^{\mathrm{T}}$, $\widehat{\Phi} = [\widehat{\phi}_{k_0}, \ldots, \widehat{\phi}_{k_0+2n-1}]^{\mathrm{T}}$. When $\widehat{\Phi}^{\mathrm{T}}\widehat{\Phi}$ is invertible (w.p.1), the estimate $\widehat{\theta}_m$ is calculated from[2]

$$\widehat{\theta}_m = (\widehat{\Phi}^{\mathrm{T}}\widehat{\Phi})^{-1}\widehat{\Phi}^{\mathrm{T}}\widehat{X}, \qquad \text{w.p.1.} \tag{12}$$

We proceed to establish the convergence of $\widehat{\theta}_m$ to θ.

Theorem 3. *Suppose that $D(q)$ and $B(q)$ are coprime. If $\{u_k\}$ is $2n$-periodic and full rank, then*

$$\widehat{\theta}_m \to \theta, \quad \text{w.p.1} \quad \textit{as } m \to \infty.$$

Proof. From $\widehat{x}_{j+2ln} = x_{j+2ln} + e_m^j$, (11) can be expressed as

$$X + E_m = (\Phi + \varsigma(E_m))\widehat{\theta}_m, \tag{13}$$

[2] Since $\widehat{\Phi}$ is a square matrix, one may also write $\widehat{\theta}_m = \widehat{\Phi}^{-1}\widehat{X}$. Eq. (12) is the standard least-squares expression.

where both E_m and $\varsigma(E_m)$ are perturbation terms, $E_m \to 0$ w.p.1 as $m \to \infty$, and $\varsigma(\cdot)$ is a continuous function of its argument satisfying $\varsigma(E) \to 0$ as $E \to 0$.

Since Φ has a uniformly bounded inverse and $\varsigma(E_m) \to 0$, w.p.1, $\Phi + \varsigma(E_m)$ is invertible w.p.1 for sufficiently large m. It follows that for sufficiently large m, by (13)

$$\Phi^{\mathrm{T}} X + \Phi^{\mathrm{T}} E_m = (\Phi^{\mathrm{T}}\Phi + \Phi^{\mathrm{T}}\varsigma(E_m))\widehat{\theta}_m.$$

This implies that

$$\begin{aligned}\widehat{\theta}_m &= (\Phi^{\mathrm{T}}\Phi + \Phi^{\mathrm{T}}\varsigma(E_m))^{-1}(\Phi^{\mathrm{T}} X + \Phi^{\mathrm{T}} E_m)\\ &\to (\Phi^{\mathrm{T}}\Phi)^{-1}\Phi^{\mathrm{T}} X = \theta\end{aligned}$$

w.p.1 as $m \to \infty$. □

5. Joint identification of distribution functions and system parameters

The developments above rely on the knowledge of the distribution function $F(\cdot)$ or its inverse. However, in most applications, the noise distributions are not known, or only limited information is available. On the other hand, input/output data from the system contain information about the noise distribution. By viewing unknown distributions and system parameters jointly as uncertainties, we develop a methodology of joint identification.

To estimate the distribution function $\xi = F(\lambda)$, one needs interpolation data in the form of $\xi_i = F(\lambda_i)$, $i = 1, 2, \ldots, T$. When $F(\cdot)$ is not parameterized, estimation of F can become sufficiently accurate only if the data points $\{\lambda_i\}$ are sufficiently dense, rendering an estimation problem of high complexity. Consequently, we adopt a parametrization approach for $F(\cdot)$.

Our approach involves three key ideas: (a) $F(\cdot)$ is approximately parameterized by a model with unknown parameter α. (b) We have shown that the empirical measure ξ_m^j is an approximate of $F(C - \gamma^j)$, where γ^j is to be estimated as well. Since the underlying system is linear, when the input u_k is scaled to ρu_k and the threshold C is shifted to C_i, we shift the data point from $F(C - \gamma^j)$ to $F(C_i - \rho_i \gamma^j)$. This allows us to generate more data points for estimation of F. (c) Since γ^j is also unknown, we estimate jointly γ^j and α. As a result, we can simultaneously estimate γ^j for system identification and α for distribution functions.

5.1. Parameterized distribution functions

Suppose that the unknown noise distribution function is $F(\cdot)$ that is approximated by a parameterized model $F(z, \alpha)$, where $\alpha = [\alpha_1, \ldots, \alpha_L]^{\mathrm{T}}$ is the unknown model parameter. For a given class $\mathbb{F}$ of possible distribution functions, the representation error of $F(z) \in \mathbb{F}$ by $F(z, \alpha)$ is

$$\varepsilon = \sup_{F \in \mathbb{F}} \inf_{\alpha} \sup_{z \in \mathscr{S}} |F(z) - F(z, \alpha)|, \tag{14}$$

where $\mathscr{S}$ is the union of supports of $F \in \mathbb{F}$.

For a given $F \in \mathbb{F}$, if the corresponding minimizer of (14) is α, then

$$F(z) = F(z, \alpha) + \Delta(z) \tag{15}$$

with $|\Delta(z)| \leqslant \varepsilon$, $\forall z \in \mathscr{S}$. When α is estimated from the data, its estimate $\widehat{\alpha}$ induces an estimated distribution function $F(z, \widehat{\alpha})$. The overall representation error becomes

$$F(z) - F(z, \widehat{\alpha}) = F(z, \alpha) - F(z, \widehat{\alpha}) + \Delta(z).$$

When a class $\mathbb{F}$ of distribution functions is given, explicit structure of the parametrization may become apparent. As an explanation, we note the following two cases.

(1) If $\mathbb{F}$ is the class of normal distributions with unknown mean μ and variance σ^2, then $F(z) = F_0((z - \mu)/\sigma)$, where $F_0(z)$ is the standard normal distribution of $\mu = 0$ and $\sigma^2 = 1$. In this case, $\mathbb{F}$ can be parameterized by $\alpha = [\mu, \sigma^2]^{\mathrm{T}}$ with $\varepsilon = 0$.

(2) Suppose F is a uniform distribution of a fixed but unknown interval $\mathscr{S} = [a, b]$. For $z \in \mathscr{S}$, $F(z)$ is completely parameterized by $F(z) = \alpha_1 + \alpha_2 z$ with $\varepsilon = 0$. On the other hand, if the uniform distribution is known to have zero mean, then $\mathscr{S} = [-\delta, \delta]$ and

$$F(z) = \frac{z}{2\delta} + \frac{1}{2}, \quad z \in \mathscr{S}.$$

In these examples, the parametrization $F(z, \alpha)$ comes naturally and represents $F(z)$ precisely for all $z \in \mathscr{S}$. However, in general one may need to use more generic structures of parametrization. For example, for computational convenience, it is common to use a set of L base functions $p^j(\cdot)$, $j = 1, \ldots, L$ to represent $F(\cdot)$. Then

$$F(z, \alpha) = \sum_{j=1}^{L} \alpha_j p^j(z) = p^{\mathrm{T}}(z)\alpha, \tag{16}$$

where $p(z) = [p^1(z), \ldots, p^{\rho}(z)]^{\mathrm{T}}$.

It is noted that some routine modifications to (16) may be needed. For example, suppose polynomials of z are used as base functions. If $F(z)$ is a normal distribution, then for any finite L, $F(z)$ cannot be well approximated by $F(z, \alpha)$ over all $z \in \mathbb{R}$. In this case, one may limit (16) to a finite interval $[a, b]$ and modify $F(z, \alpha)$ for $z \notin [a, b]$ so that $F(z, \alpha)$ decreases towards 0 for $z \to -\infty$ and $F(z, \alpha)$ increases towards 1 when $z \to \infty$. Since these techniques are standard in function approximations, they will not be discussed further.

5.2. Joint identification problems

The main idea of our approach is to explore input scaling, possibly together with threshold shifting, to provide joint information on the unknown distribution function and system parameters. Due to parametrization of the uncertainty set $\mathbb{F}$, identification of F is reduced to parameter estimation of α.

To be more specific, the $2n$-periodic full-rank input u employed in the previous section can be expanded by scaling: let

ρ_i, $i=1,\ldots,\kappa$ be κ nonzero scaling factors. Define $u^i=\rho_i u$. Note that by linearity of the system, when the input is u^i, for a given $j=1,\ldots,2n$ the corresponding output from (5) becomes

$$y^i_{2nl+j}=\rho_i\gamma^j+d_{j+2ln},\quad l=0,1,\ldots,m-1.$$

In addition, the threshold C may also be shifted to C_i.

Under the periodic signal u, scaling factors ρ_i, and thresholds C_i, let the corresponding sequences of the sensor output be $\{s^i_k\}$. Now, the empirical measures

$$\xi^j_m(i)=\frac{1}{m}\sum_{l=0}^{m-1}s^i_{2nl+j}\to\xi^j(i),\quad \text{w.p.1.} \tag{17}$$

The limit of the empirical measures satisfies, for a given $j=1,\ldots,2n$,

$$\xi^j(i)=F(C_i-\rho_i\gamma^j,\alpha),\quad i=1,\ldots,\kappa, \tag{18}$$

which will be used to calculate $\Gamma=[\gamma^1,\ldots,\gamma^{2n}]^{\mathrm{T}}$ and α.

By Theorem 3, when u_k is $2n$-periodic and full rank, θ can be identified from γ. As a result, joint identification of α and θ is reduced to joint identification of α and Γ.

5.3. Richness conditions for joint identification

An essential property for identifying α and Γ is that the Eqs. (18) have a unique solution. It is noted that for a given j, the κ equations

$$\xi^j(i)=F(C_i-\rho_i\gamma^j,\alpha),\quad i=1,\ldots,\kappa,$$

contain $L+1$ unknowns: γ^j and α. Hence, we should take $\kappa\geqslant L+1$. Since this applies to each γ^j, we will concentrate only on a generic expression

$$\xi(i)=F(C_i-\rho_i\gamma,\alpha),\quad i=1,\ldots,\kappa. \tag{19}$$

If C_i and ρ_i are selected such that (19) has a unique solution α and γ, then by repeating the procedure for $\gamma=\gamma^j$, $j=1,\ldots,2n$, (18) will have a unique solution α and Γ. Denote $\Lambda=\{(C_i,\rho_i),i=1,\ldots,\kappa\}$.

Suppose the prior information on α and γ is that $[\alpha^{\mathrm{T}},\gamma]^{\mathrm{T}}\in\Omega\subseteq\mathbb{R}^{L+1}$.

Definition 2. Given a parametrization $F(z,\alpha)$, a set of pairs $\Lambda=\{(C_i,\rho_i),i=1,\ldots,\kappa\}$ is said to be *sufficiently rich* for joint identification of α and Γ if under Λ, (19) has a unique solution α and γ in Ω.

Remark 1. A sufficient condition for Λ to be sufficiently rich is that the $\kappa\times(L+1)$ Jacobian matrix

$$J=\begin{bmatrix}\dfrac{\partial F(C_1-\rho_1\gamma,\alpha)}{\partial\alpha} & -\rho_1\dfrac{\partial F(C_1-\rho_1\gamma,\alpha)}{\partial(C_1-\rho_1\gamma)}\\ \vdots & \vdots\\ \dfrac{\partial F(C_\kappa-\rho_\kappa\gamma,\alpha)}{\partial\alpha} & -\rho_\kappa\dfrac{\partial F(C_\kappa-\rho_\kappa\gamma,\alpha)}{\partial(C_\kappa-\rho_\kappa\gamma)}\end{bmatrix}$$

is full rank for all $[\alpha^{\mathrm{T}},\gamma]^{\mathrm{T}}\in\Omega$.

Example 4. Suppose F is a normal distribution function with unknown μ and σ, $F(z)=F_0((z-\mu)/\sigma)$, where F_0 is the normal distribution of $\mu=0$ and $\sigma=1$. Then (19) becomes

$$\xi(i)=F_0((C_i-\rho_i\gamma-\mu)/\sigma),\quad i=1,2,3.$$

Define $x_i=F_0^{-1}(\xi(i))$, $i=1,2,3$. We have

$$x_i=\frac{C_i-\rho_i\gamma-\mu}{\sigma},\quad i=1,2,3$$

or

$$\begin{bmatrix}C_1 & -\rho_1 & -1\\ C_2 & -\rho_2 & -1\\ C_3 & -\rho_3 & -1\end{bmatrix}\begin{bmatrix}1/\sigma\\ \gamma/\sigma\\ \mu/\sigma\end{bmatrix}=\begin{bmatrix}x_1\\ x_2\\ x_3\end{bmatrix}.$$

In this case, (19) has a unique solution if the matrix

$$M=\begin{bmatrix}C_1 & -\rho_1 & -1\\ C_2 & -\rho_2 & -1\\ C_3 & -\rho_3 & -1\end{bmatrix}$$

is full rank. For example, if $C_1=1$; $C_2=2$; $C_3=4$; $\rho_1=1$; $\rho_2=3$; $\rho_3=5$, then it can be calculated that

$$M=\begin{bmatrix}1 & -1 & -1\\ 2 & -3 & -1\\ 4 & -5 & -1\end{bmatrix},$$

which is full rank for any μ,σ,γ.

In this example, it is easy to verify that shifting the threshold is necessary for M to be full rank. Indeed, if $C_1=C_2=C_3=C$, then M is not full rank. In fact the expression

$$\frac{C_i-\rho_i\gamma-\mu}{\sigma}=\frac{C-\mu}{\sigma}-\rho_i\frac{\gamma}{\sigma}$$

cannot be used to determine three parameters μ,σ,γ.

On the other hand, if it is known that the noise is zero mean, namely, $\mu=0$, then one may use a fixed threshold. In this case we have

$$x_1=\frac{C}{\sigma}-\rho_1\frac{\gamma}{\sigma},\quad x_2=\frac{C}{\sigma}-\rho_2\frac{\gamma}{\sigma}.$$

γ and σ can be solved uniquely if $\rho_1\neq\rho_2$.

5.4. Identification algorithms of system parameters and distribution functions

Note that the event $\{y^i_{j+2ln}\leqslant C_i\}$ is the same as $\{d_{j+2ln}\leqslant C_i-\rho_i\gamma^j\}$. Then $\xi^j_m(i)$ is precisely the value of the m-sample empirical distribution $\widehat{F}_m(x)$ of the noise d at $x=C_i-\rho_i\gamma^j$: $\xi^j_m(i)=\widehat{F}_m(C_i-\rho_i\gamma^j)$. Consequently, in consideration of parameterized models of F, over κ input sequences we obtain the following κ sample values of $F(z,\alpha)$ at

$$\xi^j_m(i)=F(C_i-\rho_i\gamma^j,\alpha)+e^j_m(i)+\Delta,\quad i=1,\ldots,\kappa,$$

where $\mathrm{e}^j_m(i)$ is the identification error and Δ is the representation error in (15).

For notational simplicity, we shall use generic symbols $\gamma=\gamma^j$, $\xi_m(i)=\xi_m^j(i)$, and $e_m(i)=e_m^j(i)$ for algorithm derivations. Hence, consider

$$\xi_m(i)=F(C_i-\rho_i\gamma,\alpha)+e_m(i)+\Delta,\quad i=1,\ldots,\kappa.$$

In general, parametrization $F(z,\alpha)$ is a nonlinear mapping with respect to α. Consequently, nonlinear equations $\xi_m(i)= =F(C_i-\rho_i\gamma,\widehat{\alpha})$ will be used to derive an estimate $\widehat{\alpha}$. For simplicity of discussions, we will present our algorithms for linear parametrization since it renders a simpler sequential procedure.

By the linear representation of $F(z)$ in (16), we have that for $i=1,\ldots,\kappa$,

$$\xi_m(i)=p^{\mathrm{T}}(C_i-\rho_i\gamma)\alpha+\Delta(C_i-\rho_i\gamma)+e_m(i).$$

By defining $\Xi_m=[\xi_m(1),\ldots,\xi_m(\kappa)]^{\mathrm{T}}$ and $P(\gamma)=[p(C_1-\rho_1\gamma),\ldots,p(C_\kappa-\rho_\kappa\gamma)]^{\mathrm{T}}$, $\Delta=[\Delta(C_1-\rho_1\gamma),\ldots,\Delta(C_\kappa-\rho_\kappa\gamma)]^{\mathrm{T}}$, $\widetilde{E}_m=[e_m(1),\ldots,e_m(\kappa)]^{\mathrm{T}}$, we obtain the relationship

$$\Xi_m=P(\gamma)\alpha+\Delta+\widetilde{E}_m. \tag{20}$$

The goal here is to select α and γ to minimize $\|\Xi_m(\gamma,\alpha)-P(\gamma)\alpha\|_2^2$, where $\|\cdot\|_2$ is the Euclidean norm. The following joint identification algorithm is introduced.

We shall write (20) as

$$\Xi=P(\gamma)\alpha+\Delta+\widetilde{E}.$$

For any given γ, if the corresponding $P(\gamma)$ is full rank, the optimal least-squares estimation error for α is

$$V(\gamma)=\|(I-P(\gamma)(P^{\mathrm{T}}(\gamma)P(\gamma))^{-1}P^{\mathrm{T}}(\gamma))\Xi\|_2^2.$$

Then the following optimal line search optimization is conducted:

$$\min_{\gamma} V(\gamma). \tag{21}$$

Denote the optimal solution by $\widehat{\gamma}$. Then

$$\widehat{\alpha}=(P^{\mathrm{T}}(\widehat{\gamma})P(\widehat{\gamma}))^{-1}P^{\mathrm{T}}(\widehat{\gamma})\Xi. \tag{22}$$

This algorithm is based on the consecutive-marginal optimization

$$\inf_{\gamma}\left(\inf_{\alpha}\|\Xi-P(\gamma)\alpha\|_2^2\right)=\inf_{\gamma} V(\gamma). \tag{23}$$

Observe that in general, the joint identification

$$\inf_{\alpha,\gamma}\|\Xi-P(\gamma)\alpha\|_2^2 \tag{24}$$

is a nonlinear optimization problem, which bears higher computational complexity. Although for a finite observation, the consecutive optimization (23) may not be equivalent to the estimates from the joint optimization (24), convergence results on $\widehat{\gamma}$ and $\widehat{\alpha}$ can be established.

Note that for algorithm execution, (23) will be repeated for $\gamma=\gamma^j$, $j=1,\ldots,2n$. This understanding will be assumed for the rest of the paper and will not be reiterated.

5.5. Convergence analysis

We now derive convergence properties of $\widehat{\gamma}$ and $\widehat{\alpha}$.

Assumption A2. $\{d_k\}$ is a sequence of independent and identically distributed (i.i.d.) random variables whose distribution function $F(\cdot)$ together with its inverse $F^{-1}(\cdot)$ is differentiable. $F(\cdot)$ is unknown but belongs to a class $\mathbb{F}$.

Theorem 4. *Suppose that $\Lambda=\{(C_i,\rho_i),i=1,\ldots,\kappa\}$ is sufficiently rich. Under Assumption A2 and representation error bound* (14), *for any compact subset $S\subset\mathbb{R}$,*

$$\lim_{m\to\infty}\sup_{z\in S}|F(z,\widehat{\alpha}_m)-F(z)|\to W,\quad w.p.1,$$

where $|W|\leqslant\beta\varepsilon$ for some constant $\beta>0$.

Proof. By virtue of the well-known Glivenko–Cantelli Theorem (Billingsley, 1968, p. 103), $|\widehat{F}_m(z)-F(z)|\to 0$ w.p.1, and the convergence is uniform on any compact subset. Since both $F(\cdot)$ and F^{-1} are continuous, and

$$F(z)=p^{\mathrm{T}}(z)\alpha+\Delta(z)$$

w.p.1, when $m\to\infty$, for $i=1,\ldots,\kappa$,

$$F(C_i-\rho_i\gamma)=p^{\mathrm{T}}(C_i-\rho_i\gamma)\alpha+\Delta(C_i-\rho_i\gamma).$$

Or to put it into a vector form

$$\Xi=P\alpha+\Delta.$$

Since Λ is sufficiently rich, P^{-1} exists. Hence, by the least-squares method, we obtain

$$\widehat{\alpha}_m=(P^{\mathrm{T}}P)^{-1}P^{\mathrm{T}}\Xi$$

and

$$\widehat{\alpha}_m-\alpha=(P^{\mathrm{T}}P)^{-1}P^{\mathrm{T}}\Delta,$$

which implies

$$\|\widehat{\alpha}_m-\alpha\|_2\leqslant\beta_1\varepsilon,$$

for some constant $\beta_1>0$. Consequently, for some $\beta_2>0$,

$$\begin{aligned}|F(z,\widehat{\alpha}_m)-F(z)|&\leqslant|p^{\mathrm{T}}(z)(\alpha-\widehat{\alpha}_m)+\Delta(z)|\\&\leqslant\beta_2\beta_1\varepsilon+\varepsilon=\beta\varepsilon.\end{aligned}\qquad\square$$

Theorem 4 implies a bound on estimation errors of γ.

Theorem 5. *Under Assumption A2 and the representation of error bound* (14),

$$\lim_{m\to\infty}\sup|\widehat{\gamma}_m-\gamma|\to\beta_0\varepsilon\quad w.p.1,\quad j=1,\ldots,2n,$$

for some constant $\beta_0>0$.

Proof. We can write

$$F(C_i-\rho_i\gamma)=F(C_i-\rho_i\gamma,\widehat{\alpha}_m)+\Delta+\widetilde{e}_m,$$

where $\{\widetilde{e}_m\}$ is a sequence of random errors satisfying $\widetilde{e}_m \to 0$, w.p.1. Since $F^{-1}(\cdot)$ is differentiable, we conclude that

$$|(C_i - \rho_i\widehat{\gamma}) - (C_i - \rho_i\gamma)| \leqslant \beta_1 |F^{-1}(\widetilde{\Delta} + e_m)|$$

for some constant β_1. The continuity of $F^{-1}(\cdot)$ implies that $F^{-1}(\Delta + \widetilde{e}_m) \to F^{-1}(\Delta)$ w.p.1, as $m \to \infty$. As a result,

$$\limsup_{m\to\infty} |\widehat{\gamma} - \gamma| \leqslant \beta_0 \varepsilon,$$

as stated. □

In particular, if F is well represented by the parameterized model, i.e., $\varepsilon \to 0$, then $\widehat{\gamma}_m \to \gamma$ w.p.1 as $m \to \infty$.

6. Algorithm flowcharts and an illustrative example

Our algorithms for joint identification of system parameters and noise distributions are summarized in Fig. 3.

We now use an example to demonstrate the identification algorithms presented so far.

Suppose that the true plant is a first-order system

$$x_k = -a_0 x_{k-1} + b_0 u_{k-1}, \qquad y_k = x_k + d_k,$$

where $a_0 = 0.4$; $b_0 = 1.6$. $\{d_k\}$ is an i.i.d. sequence, uniformly distributed on $[-1.2, 1.2]$. Hence, the true distribution function is $\xi = F(z) = (1/2.4)z + 0.5$ for $z \in [-1.2, 1.2]$. The true system parameters and the distribution function interval are unknown.

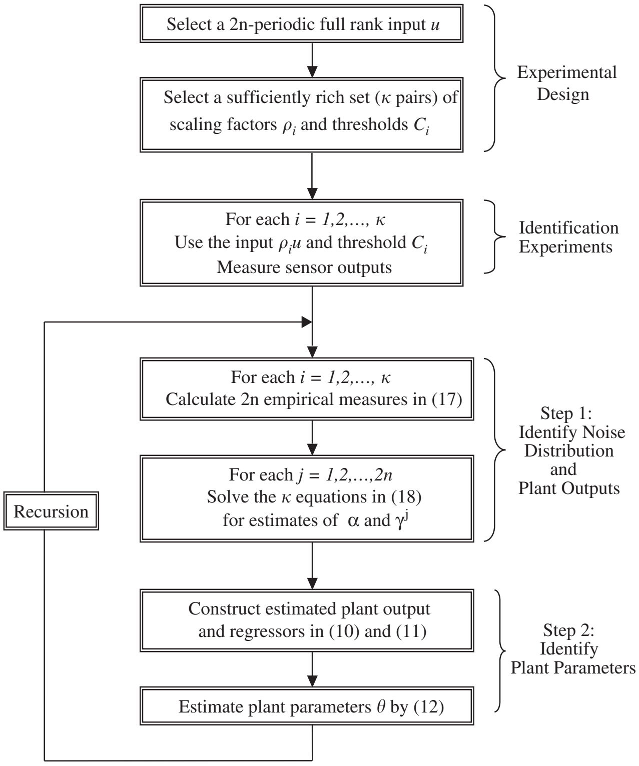
Select a 2n-periodic full rank input u
Select a sufficiently rich set (κ pairs) of scaling factors ρ_i and thresholds C_i
Experimental Design
For each $i = 1,2,\ldots,\kappa$
Use the input $\rho_i u$ and threshold C_i
Measure sensor outputs
Identification Experiments
Recursion
For each $i = 1,2,\ldots,\kappa$
Calculate 2n empirical measures in (17)
For each $j = 1,2,\ldots,2n$
Solve the κ equations in (18)
for estimates of α and γ^j
Step 1: Identify Noise Distribution and Plant Outputs
Construct estimated plant output and regressors in (10) and (11)
Estimate plant parameters θ by (12)
Step 2: Identify Plant Parameters

Fig. 3. An algorithm flowchart.

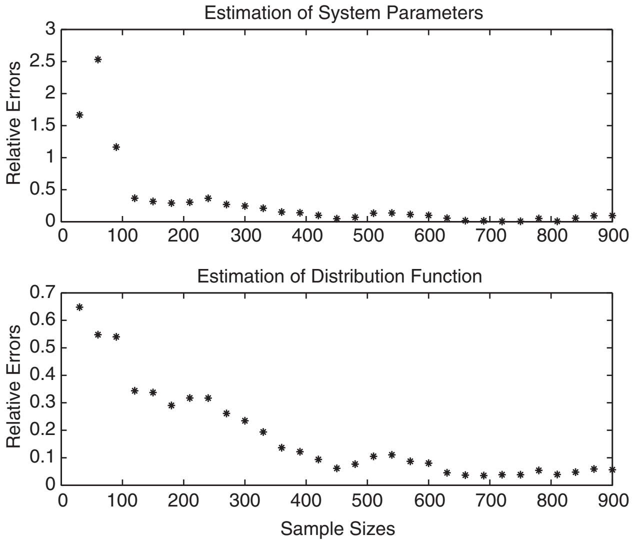

Fig. 4. Joint Identification of distribution functions and plant parameters.

(1) *Experimental design*: Here we need to select parametrization of the unknown distribution function, input signal, scaling, and threshold selections.
   (a) For this example, we assume the linear function parametrization of F: $\xi = F(z, \alpha) = \alpha_1 + \alpha_2 z$. Since this is a correct parametrization, the function representation error $\varepsilon = 0$.
   (b) To identify the two system parameters a_0 and b_0, the base input is 2-periodic with $u_1 = 0.7$; $u_2 = 0.2$, which is full rank.
   (c) Signal scaling factors ρ_i and thresholds C_i are to be selected such that (19) can be solved uniquely for α and γ. In this application, we have

$$\xi(1) = \alpha_1 + \alpha_2(C_1 - \rho_1\gamma),$$
$$\xi(2) = \alpha_1 + \alpha_2(C_2 - \rho_2\gamma),$$
$$\xi(3) = \alpha_1 + \alpha_2(C_3 - \rho_3\gamma).$$

This system has a unique solution if

$$M = \begin{bmatrix} 1 & C_1 & \rho_1 \\ 1 & C_2 & \rho_2 \\ 1 & C_3 & \rho_3 \end{bmatrix}$$

is full rank. For example, we use the following three sets of values: $\rho_1 = 0.3$, $C_1 = -0.4$; $\rho_2 = 0.5$, $C_2 = 0.4$; $\rho_3 = 0.8$, $C_3 = 0.8$. This leads to

$$M = \begin{bmatrix} 1 & -0.4 & 0.3 \\ 1 & 0.4 & 0.5 \\ 1 & 0.8 & 0.8 \end{bmatrix},$$

which is full rank.

(2) *Identification*: Identify α and θ.
   (a) the system output y_k is simulated by

$$y_k = -a_0 x_{k-1} + b_0 u_{k-1} + d_k$$

for a total of 900 sample steps.

(b) The sensor outputs are observed and empirical measures are calculated.

(c) The recursive identification algorithms (21) and (22) are applied to identify the plant parameters and distribution function simultaneously.

(3) *Evaluation*: The plant parameter estimates are compared to the true values $a_0 = 0.4$, $b_0 = 1.6$; and distribution function parameters $[\alpha_1, \alpha_2]$ are compared to their true values $[0.5, 1/2.4]$. The results are shown in Fig. 4, where relative errors are plotted as a function of sample sizes.

7. Recursive algorithms

In this section, we develop a class of recursive algorithms for estimating α and γ. In lieu of the line search (21) and least-squares procedure (22), the estimate $\widehat{\alpha}_m$ will be constructed via an adaptive filtering algorithm to reduce the computational complexity, and estimate $\widehat{\gamma}_m$ will be recursified. This section is divided into three parts: first, we present the algorithms. Then, we establish the convergence of the schemes. Finally, we make some additional remarks on alternatives.

The identification problem involves several indices which can be confusing in our recursive algorithms: (a) the time index k; (b) the time-block index m. Iteration from m to $m+1$ represents an acquisition of $2n$ observation points on s_k; (c) the cyclic index $j = 1, \ldots, 2n$. This index indicates rotation of parameters γ^j in identification, in other words, indicating one of the sequential optimization problems; (d) the index i in ρ_i, $i = 1, \ldots, \kappa$. This represents the ith scaling factor ρ_i is applied at input.

For example, due to the cyclic nature, $\widehat{\gamma}^j$ can only be updated once every $2n$ data points. As a result, it is indexed as $\widehat{\gamma}^j_m$. On the other hand, all data points contain information on α. Hence, it can be indexed as $\widehat{\alpha}_k$. In case that we choose to update $\widehat{\alpha}$ at the same time of updating γ^j, we shall use $\widehat{\alpha}_m$ instead.

7.1. Recursive schemes

The following two typical classes of recursive algorithms will be considered.

(A) *Adaptive filtering algorithms*: For each $i = 1, \ldots, \kappa$ of scaling values at the input, and $j = 1, \ldots, 2n$

$$\begin{cases} \xi^j_{m+1}(i) = \xi^j_m(i) - \frac{1}{m+1}[\xi^j_m(i) - s_{j+2(m+1)n}], \\ \widehat{\alpha}^j_{m+1}(i) = \widehat{\alpha}^j_m(i) + \frac{p_m[\xi^j_m(i) - p^{\mathrm{T}}_m\widehat{\alpha}^j_m(i)]}{m+1}, \\ \widehat{\gamma}^j_{m+1}(i) = \widehat{\gamma}^j_m(i) + \frac{1}{m+1}\left[\widehat{\gamma}^j_m(i) - \frac{C_i - \widetilde{F}^{-1}(\xi^j_m(i), \widehat{\alpha}_m(i))}{\rho_i}\right], \end{cases} \tag{25}$$

where

$$p_m = p(C_i - \rho_i\widehat{\gamma}^j_m),$$

with $p(\cdot)$ given in (16), and $F^{-1}(z, \widehat{\alpha})$ denotes the inverse of $F(z, \widehat{\alpha})$ when $\widehat{\alpha}$ is used. Note that in fact, p_m is j-dependent, so it should have been written as p^j_m. We have suppressed j-dependence for notational simplicity.

(B) *Combined adaptive filtering and least-squares algorithm*: For each $i = 1, \ldots, \kappa$ of scaling values at the input, $j = 1, \ldots, 2n$

$$\begin{cases} \xi^j_{m+1}(i) = \xi^j_m(i) - \frac{1}{m+1}[\xi^j_m(i) - s_{j+2(m+1)n}], \\ \widehat{\alpha}^j_{m+1}(i) = \widehat{\alpha}^j_m(i) + a_m\Psi_m p_m[\xi^j_m(i) - p^{\mathrm{T}}_m\widehat{\alpha}^j_m(i)], \\ \Psi_{m+1} = \Psi_m - a_m\Psi_m p_m p^{\mathrm{T}}_m\Psi_m, \\ a_m = (1 + p^{\mathrm{T}}_m\Psi_m p_m)^{-1}, \\ \widehat{\gamma}^j_{m+1}(i) = \widehat{\gamma}^j_m(i) + \frac{1}{m+1}\left[\widehat{\gamma}^j_m(i) - \frac{C_i - F^{-1}(\xi^j_m(i), \widehat{\alpha}^j_m(i))}{\rho_i}\right]. \end{cases} \tag{26}$$

7.2. Asymptotic properties of recursive algorithm (25)

In what follows, we present asymptotic properties of the algorithms given in (25) and (26). To proceed, we need some conditions, which are listed below.

Assumption A3. The following system of differential equations

$$\begin{cases} \frac{\mathrm{d}}{\mathrm{d}t}\alpha^{ji}(t) = p(C_i - \gamma^{ji}(t))F(C_i - \gamma^j) \\ \qquad\qquad -p(C_i - \gamma^{ji}(t))p^{\mathrm{T}}(C_i - \gamma^{ji}(t))\alpha^{ji}(t) \\ \frac{\mathrm{d}}{\mathrm{d}t}\gamma^{ji}(t) = \gamma^{ji}(t) - \left[\frac{C_i - F^{-1}(F(C_i - \gamma^j\rho_i), \alpha^{ji}(t))}{\rho_i}\right] \end{cases} \tag{27}$$

has a unique solution for each initial condition. In addition, (27) has a unique asymptotically stable point $(\alpha^{j,0}, \gamma^{j,0})$ in the sense of Lyapunov.

Assumption A4. The following conditions hold:

- The sequences $\{\widehat{\gamma}^j_m\}$ for $j = 1, \ldots, 2n$ are bounded w.p.1.
- Denoting $A^j = p(C_i - \rho_i\gamma^{j,0})p^{\mathrm{T}}(C_i - \rho_i\gamma^{j,0})$, A^j is symmetric and positive definite.
- Both $p(\cdot)$ and $F^{-1}(\cdot, \cdot)$ are continuous.

Remark 2. To ensure the boundedness, we can use a projection algorithm

$$\widehat{\gamma}^j_{m+1}(i) = \Pi_G\left[\widehat{\gamma}^j_m(i) + \frac{\widehat{\gamma}^j_m - \frac{[C_i - F^{-1}(\xi^j_m(i), \widehat{\alpha}_m)]}{\rho_i}}{m+1}\right],$$

where Π_G is the projection operator onto the bounded set G (see (Kushner & Yin, 2003) for more details). Owing to the use of $\{s_m\}$, $\{\xi^j_m(i)\}$ is bounded. Note that we can choose G to be as simple as a box, and choose it to be large enough so that it contains the true parameter γ^j. However, for simplifying notation and for convenience, we have assumed that the boundedness of $\{\widehat{\gamma}^j_m\}$ in (A4). The continuity of $p(\cdot)$ implies that $F(\cdot, \cdot)$ is also continuous since it is linear in α. We require the matrix A^j be positive definite, which is essentially a solvability or identifiability condition.

We claim that $\{\widehat{\alpha}^j_m(i)\}$ is bounded w.p.1 uniformly in m. To see this, write

$$\widehat{\alpha}^j_{m+1}(i) = A_{m,0}\widehat{\alpha}^j_0(i) + \sum_{l=0}^{m} \frac{1}{l+1} A_{m,l}[A^j - p_l p_l^{\mathrm{T}}]\widehat{\alpha}^j_l(i) + \sum_{l=0}^{m} \frac{1}{l+1} A_{m,l} p_l \xi^j_l, \tag{28}$$

where

$$A_{m,l} = \begin{cases} \prod_{i=l+1}^{m} \left(I - \frac{1}{i+1} A^j\right), & l < m, \\ I, & l = m. \end{cases}$$

Note that following the convention for p_m, we suppressed the j-dependence in the notation $A_{m,l}$. Thus, taking norm in (28), an application of the Gronwall's inequality yields that

$$|\widehat{\alpha}^j_{m+1}(i)| \leqslant K_{1,m} \exp(K_{2,m}), \tag{29}$$

where

$$K_{1,m} = |A_{m,0}\widehat{\alpha}^j_0(i)| + \sum_{l=0}^{m} \frac{1}{l+1}|A_{m,l}||p_l \xi^j_l|,$$

$$K_{2,m} = \sum_{l=0}^{m} \frac{1}{l+1}|A_{m,l}||A^j - p_l p_l^{\mathrm{T}}|.$$

With

$$a_{m,l} = \begin{cases} \prod_{i=l+1}^{m} \left(1 - \frac{1}{i+1} \lambda\right), & l < m, \\ 1, & l = m, \end{cases}$$

where λ is the minimal eigenvalues of A^j,

$$\sum_{l=0}^{m} \frac{1}{l+1}|A_{m,l}| \leqslant \sum_{l=0}^{m} \frac{1}{l+1} a_{m,l} = \frac{K_0}{\lambda} \sum_{l=0}^{m} [a_{m,l+1} - a_{m,l}] = K_0(1 - a_{m,0}) < \infty. \tag{30}$$

Note that in the above, we used K_0 as a generic positive constant whose value may change for different appearances. The bound in (30) together with (28), the boundedness of $\{\xi^j_m\}$ and $\{\widehat{\gamma}^j_m\}$, implies that $K_{1,m}$ is bounded w.p.1 uniformly in m, so is $K_{2,m}$. The w.p.1 boundedness (uniform in m) of $\{\widehat{\alpha}^j_m(i)\}$ then follows from (29).

Next, consider the joint process $z^{ji}_m = (\widehat{\alpha}^j_m(i), \widehat{\gamma}^j_m(i))^{\mathrm{T}}$. Set

$$t_m = \sum_{l=0}^{m-1} \frac{1}{l+1}, \quad \text{and} \quad m(t) = \max\{m : t_m \leqslant t\}.$$

Define the piecewise constant interpolation

$$z^0(t) = z_m \text{ for } t \in [t_m, t_{m+1}) \quad \text{and} \quad z^m(t) = z^0(t + t_m).$$

Note that $z^m(\cdot)$ is a shifted sequence for bringing the asymptotic properties of the sequence to the foreground. We also define the component of the interpolation $z^{m,ji}(\cdot)$ as $\alpha^{m,i}(\cdot)$ and $\gamma^{m,ji}(\cdot)$. The boundedness on $\{\xi^j_m(i), \widehat{\alpha}_m(i), \widehat{\gamma}^j_m(i)\}$ yields that $z^{m,ji}(\cdot)$ is uniformly bounded. The continuity condition in Assumption A4 and the continuity of the distribution function and its inverse imply $z^{m,ji}(\cdot)$ is equicontinuous in the extended sense as defined in Kushner and Yin (2003, p. 102). By the Arzela–Ascoli theorem (Kushner & Yin, 2003, p. 102) applied to a sequence of equicontinuous functions (in the extended sense), we can extract a convergent subsequence $z^{m',ji}(\cdot)$ such that $z^{m',ji}(\cdot)$ converges to $z(\cdot)$ w.p.1 and the convergence is uniform on any bounded interval. For convenience, in what follows, we simply write m' as m.

Using the usual ODE approach (see (Kushner & Yin, 2003)), we can show $(\alpha^{m,ji}(\cdot), \gamma^{m,ji}(\cdot)) \to (\alpha^{ji}(\cdot), \gamma^{ji}(\cdot))$. Considering the pair $(\widehat{\alpha}_m, \widehat{\gamma}^j_m)$ jointly, $\alpha^{ji}(\cdot)$ and $\gamma^{ji}(\cdot)$ as components of the pair satisfy the differential equation in (27).

Next, let $\{\tau_m\}$ be a sequence of positive real numbers satisfying $\tau_m \to \infty$ as $m \to \infty$. Then it can be shown (see (Kushner & Yin, 2003) for more details) that $(\alpha^{m,ji}(\cdot + \tau_m), \gamma^{m,ji}(\cdot + \tau_m)) \to (\alpha^{0,ji}, \gamma^{ji,0})$ as $m \to \infty$. Thus, $(\widehat{\alpha}^j_m(i), \widehat{\gamma}^j_m(i)) \to (\alpha^{j,0}, \gamma^{j,0})$ w.p.1 as $m \to \infty$.

Note that the stationary point $\gamma^{0,ji}$ is given by

$$\gamma^{0,ji} = [C_i - F^{-1}(F(C_i - \gamma^{0,ji}\rho_i), \alpha^{0,i})]/\rho_i.$$

As in the previous section, it can be shown that for some $\beta_0 > 0$,

$$\limsup_{m \to \infty} |C_i - F^{-1}(\xi^j_m) - \gamma^{j,0}| \leqslant \beta_0 \varepsilon.$$

Summarizing what has been proved, we obtain the following theorem.

Theorem 6. *Assume* A1–A4. $\{\xi^j_m, \widehat{\alpha}_m, \widehat{\gamma}^j_m\}$ *converges w.p.*1. *Moreover, we have the following upper bound on the deviation* $\widehat{\gamma}^j_m - \gamma^j$:

$$\limsup_{m \to \infty} |\widehat{\gamma}^j_m - \gamma^j| \leqslant \beta_0 \varepsilon, \quad w.p.1 \textit{ for some } \beta_0 > 0.$$

7.3. Remarks

The entire procedure consists of an inner loop and an outer loop. The purpose of the inner loop is to obtain the empirical distribution and to estimate α. We first update the empirical process recursively. Then we construct a sequence of estimates of α by using a recursive least-squares-type method. After carrying out a number of iterations in the inner loop, we update the estimate of γ^j one step in the outer loop via function evaluation and the utility of $F(\cdot, \alpha)$. To some extent, the approach can be considered as a two-time-scale and two-stage approximation, in which the inner loop is updated more frequently than that of the outer one.

As a convention, we use ℓ and m to denote the indices in what follows. For a sequence of scalar or vector $\{z_k\}$, by the notation $z^{\ell m}_k$, we mean that $z^{\ell m}_0 = z_{\ell m}$ and $z^{\ell m}_k = z_{\ell m + k}$. Algorithm (25) can be modified as follows. For each ℓ and m, let

$$\begin{cases} \xi^{j,\ell m}_{k+1} = \xi^{j,\ell m}_{k} - \frac{1}{k+1}\xi^{j,\ell m}_{k} + \frac{1}{k+1}s_{j+2(\ell m+k+1)n}, \\ \quad 0\leqslant k<m, \\ \widehat{\alpha}^{\ell m}_{k+1} = \widehat{\alpha}^{\ell m}_{k} + \frac{1}{(k+1)}p_{\ell m}[\xi^{j,\ell m}_{k} - p^{\mathrm{T}}_{\ell m}\widehat{\alpha}^{\ell m}_{k}],\ 0\leqslant k<m, \\ \widehat{\gamma}^{j}_{(\ell+1)m} = \widehat{\gamma}^{j}_{\ell m} + \frac{[\widehat{\gamma}^{j}_{\ell m} - \frac{[C_i - F^{-1}(C_i - \widehat{\gamma}^{j}_{\ell m}, \widehat{\alpha}^{\ell m}_{0})]}{\rho_i}]}{\ell m}. \end{cases}$$

Moreover, for tracking slightly parameter variation, we can use a constant step size $\eta > 0$. For each ℓ and m, let

$$\begin{cases} \xi^{j,\ell m}_{k+1} = \xi^{j,\ell m}_{k} - \frac{1}{k+1}\xi^{j,\ell m}_{k} + \frac{1}{k+1}s_{j+2(\ell m+k+1)n}, \\ \quad 0\leqslant k<m, \\ \widehat{\alpha}^{\ell m}_{k+1} = \widehat{\alpha}^{\ell m}_{k} + \eta p_{\ell m}[\xi^{j,\ell m}_{k} - p^{\mathrm{T}}_{\ell m}\widehat{\alpha}^{\ell m}_{k}], \quad 0\leqslant k<m, \\ \widehat{\gamma}^{j}_{(\ell+1)m} = \widehat{\gamma}^{j}_{\ell m} + \eta\left[\widehat{\gamma}^{j}_{\ell m} - \frac{[C_i - F^{-1}(C_i - \widehat{\gamma}^{j}_{\ell m}, \widehat{\alpha}^{\ell m}_{0})]}{\rho_i}\right]. \end{cases}$$

8. Conclusions

When sensors are nonlinear and non-smooth such as the switching sensors investigated in this paper, system identification for plants in ARMA structures usually becomes difficult, due to lack of constructive and convergent identification algorithms. This paper introduces a two-step approach to resolve this complicated problem. This approach is further extended to accommodate the common scenarios in which noise distribution functions are unknown. Convergence properties of all the algorithms are established.

The main results of this paper can be extended in several directions. Practical systems are infinite dimensional. Introduction of unmodeled dynamics and system order estimation will accommodate model uncertainties for practical systems. Similar results can also be derived for identification of nonlinear systems of simple structures.

Acknowledgements

The research of L.Y. Wang was supported in part by the National Science Foundation under ECS-0329597, and in part by Wayne State University Research Enhancement Program. The research of G. Yin was supported in part by the National Science Foundation under DMS-0304928, and in part by Wayne State University Research Enhancement Program. The research of J.F. Zhang was supported by the National Natural Science Foundation of China.

References

Billingsley, P. (1968). *Convergence of probability measures*. New York: Wiley.

Caianiello, E. R., & de Luca, A. (1966). Decision equation for binary systems: application to neural behavior. *Kybernetik*, *3*, 33–40.

Chen, H. F., & Guo, L. (1991). *Identification and stochastic adaptive control*. Boston: Birkhäuser.

Chen, H. F., & Yin, G. (2003). Asymptotic properties of sign algorithms for adaptive filtering. *IEEE Transactions on Automatic Control*, *48*, 1545–1556.

Elvitch, C. R., Sethares, W. A., Rey, G. J., & Johnson, C. R., Jr. (1989). Quiver diagrams and signed adaptive fiters. *IEEE Transactions on Acoustics Speech and Signal Processing*, *30*, 227–236.

Eweda, E. (1995). Convergence analysis of an adaptive filter equipped with the sign-sign algorithm. *IEEE Transactions on Automatic Control*, *40*, 1807–1811.

Feller, W. (1968). *An introduction to probability theory and its applications*, (3rd ed.) vol. I, New York: Wiley.

Feller, W. (1971). *An introduction to probability theory and its applications*, (2nd ed.) vol. II, New York: Wiley.

Gersho, A. (1984). Adaptive filtering with binary reinforcement. *IEEE Transactions on Information Theory*, *30*, 191–199.

Kushner, H. J., & Yin, G. (2003). *Stochastic approximation and recursive algorithms and applications*. 2nd ed., New York: Springer.

Ljung, L. (1987). *System identification: theory for the user*. Englewood Cliffs, NJ: Prentice-Hall.

Pakdaman, K., & Malta, C. P. (1998). A note on convergence under dynamical thresholds with delays. *IEEE Transactions on Neural Networks*, *9*(1), 231–233.

Pollard, D. (1984). *Convergence of stochastic processes*. New York: Springer.

Serfling, R. J. (1980). *Approximation theorems of mathematical statistics*. New York: Wiley.

Wang, L. Y. (1997). Persistent identification of time varying systems. *IEEE Transactions on Automatic Control*, *42*, 66–82.

Wang, L. Y., & Yin, G. (1999). Towards a harmonic blending of deterministic and stochastic frameworks in information processing. In: A. Garulli, A. Tesi, & A. Vicino (Eds.), *Robustness in Identification and Control* (pp. 102–116) Lecture Notes in Computer Science, vol. 245. Berlin: Springer.

Wang, L. Y., & Yin, G. (2000). Persistent identification of systems with unmodeled dynamics and exogenous disturbances. *IEEE Transactions on Automatic Control*, *45*(7), 1246–1256.

Wang, L. Y., & Yin, G. (2002). Closed-loop persistent identification of linear systems with unmodeled dynamics and stochastic disturbances. *Automatica*, *38*, 1463–1474.

Wang, L. Y., Zhang, J. F., & Yin, G. (2003). System identification using binary sensors. *IEEE Transactions on Automatic Control*, *48*, 1892–1907.

Yin, G., Krishnamurthy, V., & Ion, C. (2003). Iterate-averaging sign algorithms for adaptive filtering with applications to blind multiuser detection. *IEEE Transactions on Information Theory*, *49*, 657–671.

Le Yi Wang received the Ph.D. degree in electrical engineering from McGill University, Montreal, Canada, in 1990. Since 1990, he has been with Wayne State University, Detroit, MI, where he is currently a Professor in the Department of Electrical and Computer Engineering. His research interests are in the areas of H-infinity optimization, complexity and information, robust control, system identification, adaptive systems, hybrid and nonlinear systems, information processing and learning, as well as automotive, computer and medical applications of control methodologies.

Dr. Wang was awarded the Research Initiation Award in 1992 from the National Science Foundation. He also received the Faculty Research Award from Wayne State University, in 1992, and the College Outstanding Teaching Award from the College of Engineering, Wayne State University, in 1995. He was a keynote speaker in two international conferences. He serves on the IFAC Technical Committee on Modeling, Identification and Signal Processing. He served as an Associate Editor of the IEEE Transactions on Automatic Control, and currently is an Editor of the Journal of System Sciences and Complexity, an Associate Editor of Journal of Control Theory and Applications and International Journal of Control and Intelligent Systems.

G. George Yin received his B.S. in Mathematics from the University of Delaware in 1983, M.S. in Electrical Engineering and Ph.D. in Applied Mathematics from Brown University in 1987. Subsequently, he joined the Department of Mathematics, Wayne State University, and became a professor in 1996. He is a fellow of IEEE. He served on the Mathematical Review Date Base Committee, IFAC Technical Committee on Modeling, Identification and Signal Processing, and various conference program committees; he was the editor of SIAM Activity Group on Control and Systems Theory Newsletters, the SIAM Representative to the 34th CDC, Co-Chair of 1996 AMS–SIAM Summer Seminar in Applied Mathematics, and Co-Chair of 2003 AMS–IMS–SIAM Summer Research Conference: Mathematics of Finance, Co-organizer of 2005 IMA Workshop on Wireless Communications. He is an Associate Editor of Automatica and SIAM Journal on Control and Optimization, and was also an Associate Editor of IEEE Transactions on Automatic Control from 1994 to 1998, and is (or was) on the editorial board of five other journals.

Ji-Feng Zhang received his B.S. degree in Mathematics from Shandong University in 1985, and the Ph.D. degree from the Institute of Systems Science (ISS), Chinese Academy of Sciences (CAS) in 1991. Since 1985 he has been with the ISS, CAS, where he is now a Professor of the Academy of Mathematics and Systems Science, the Vice-Director of the ISS. His current research interests are system modeling and identification, adaptive control, stochastic systems, and descriptor systems. He received the Science Fund for Distinguished Young Scholars from NSFC in 1997, the First Prize of the Young Scientist Award of CAS in 1995, and now is a Vice-General Secretary of the Chinese Association of Automation (CAA), Vice-Director of the Control Theory Committee of CAA, Deputy Editor-in-Chief of the journals “Acta Automatica Sinica” and “Journal of Systems Science and Mathematical Sciences”.