Int. J. Appl. Math. Comput. Sci., 2013, Vol. 23, No. 3, 521–537
DOI: 10.2478/amcs-2013-0040

NONPARAMETRIC INSTRUMENTAL VARIABLES FOR IDENTIFICATION OF BLOCK–ORIENTED SYSTEMS

GRZEGORZ MZYK

Institute of Computer Engineering, Control and Robotics
Wrocław University of Technology, Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland
e-mail: grzegorz.mzyk@pwr.wroc.pl

A combined, parametric-nonparametric identification algorithm for a special case of NARMAX systems is proposed. The parameters of individual blocks are aggregated in one matrix (including mixed products of parameters). The matrix is estimated by an instrumental variables technique with the instruments generated by a nonparametric kernel method. Finally, the result is decomposed to obtain parameters of the system elements. The consistency of the proposed estimate is proved and the rate of convergence is analyzed. Also, the form of optimal instrumental variables is established and the method of their approximate generation is proposed. The idea of nonparametric generation of instrumental variables guarantees that the I.V. estimate is well defined, improves the behaviour of the least-squares method and allows reducing the estimation error. The method is simple in implementation and robust to the correlated noise.

Keywords: system identification, instrumental variables, NARMAX system, Hammerstein system, Wiener system, Lur'e system, nonparametric methods.

1. Problem statement

1.1. System. The paper considers the problem of identification of a scalar, discrete-time, asymptotically stable nonlinear dynamic system shown in Fig. 1, and described by the following equation (cf. Bai, 1998):

$$y_k = \sum_{j=1}^{p} \lambda_j \eta(y_{k-j}) + \sum_{i=0}^{n} \gamma_i \mu(u_{k-i}) + z_k, \qquad (1)$$

where

$$\begin{aligned} \mu(u) &= \sum_{t=1}^{m} c_t f_t(u), \\ \eta(y) &= \sum_{l=1}^{q} d_l g_l(y). \end{aligned} \qquad (2)$$

The structure is well known in the literature (see, e.g., Giri and Bai, 2010), and can be treated as a special case of the additive NARMAX model (Chen and Billings, 1989).

The signals y_k, u_k and z_k are the output, the input and the noise, respectively. The system in Fig. 1 is more general than the Hammerstein system often met in the literature. The Hammerstein system is obtained when the function $\eta(\cdot)$ is linear (see Appendix A). Also, it is not equivalent to the Wiener–Hammerstein (sandwich) system widely considered in the literature, where two linear dynamic blocks surround one static nonlinearity. In spite of many possibilities of applications in various domains (Haber and Keviczky, 1999; Bai, 1998; Zhang *et al*., 1996; Suykens *et al*., 1998; Sastry, 1999; Lu and Hill, 2007), relatively little attention has been paid to this structure in the literature.

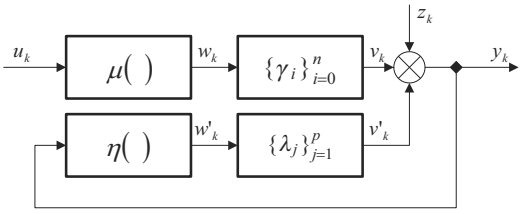

Fig. 1. Additive NARMAX system.

1.2. Assumptions. The following assumptions are made.

Assumption 1. The static nonlinear characteristics are of

a given parametric form,

$$\begin{aligned}\mu(u) &= \sum_{t=1}^{m} c_t f_t(u),\\ \eta(y) &= \sum_{l=1}^{q} d_l g_l(y),\end{aligned} \tag{3}$$

where $f_1(\cdot),\ldots,f_m(\cdot)$ and $g_1(\cdot),\ldots,g_q(\cdot)$ are *a priori* known linearly independent basis functions such that

$$|f_t(u)| \le p_{\max}, \tag{4}$$

$$|g_l(y)| \le p_{\max}, \tag{5}$$

for some constant $p_{\max}$.

Assumption 2. The linear dynamic blocks have finite impulse responses, i.e.,

$$v_k = \sum_{i=0}^{n} \gamma_i w_{k-i}, \tag{6}$$

$$v'_k = \sum_{j=1}^{p} \lambda_j w'_{k-j}, \tag{7}$$

with known orders n and p.

Assumption 3. The input process $\{u_k\}$ is a sequence of i.i.d. bounded random variables, i.e., there exists (unknown) $u_{\max}$, such that $|u_k| < u_{\max} < \infty$.

Assumption 4. The output noise $\{z_k\}$ is a correlated linear process. It can be written as

$$z_k = \sum_{i=0}^{\infty} \omega_i \varepsilon_{k-i}, \tag{8}$$

where $\{\varepsilon_k\}$ is some unknown zero-mean ($E\varepsilon_k = 0$) and bounded ($|\varepsilon_k| < \varepsilon_{\max} < \infty$) i.i.d. process, independent of the input $\{u_k\}$, and $\{\omega_i\}_{i=0}^{\infty}$ ($\sum_{i=0}^{\infty} |\omega_i| < \infty$) is an unknown stable linear filter.

Assumption 5. The overall system is asymptotically stable.

Assumption 6. Only the input $\{u_k\}$ and the output of the whole system $\{y_k\}$ are accessible for measurements.

Let

$$\begin{aligned}\Lambda &= (\lambda_1,\ldots,\lambda_p)^T,\\ \Gamma &= (\gamma_0,\ldots,\gamma_n)^T,\\ c &= (c_1,\ldots,c_m)^T,\\ d &= (d_1,\ldots,d_q)^T,\end{aligned} \tag{9}$$

denote true (unknown) parameters of the system. Obviously, the input-output description of the system, given by (1) and (2) is not unique. For each pair of constants $\overline{\alpha}$ and $\overline{\beta}$, the systems with parameters Λ, Γ, c, d and $\overline{\beta}\Lambda$, $\overline{\alpha}\Gamma$, $c/\overline{\alpha}$, $d/\overline{\beta}$ cannot be distinguished, i.e., they are equivalent (see (1)-(2)). For the uniqueness of the solution, the following technical assumptions are introduced (see Bai, 1998):

(a) the matrices $\Theta_{\Lambda d} = \Lambda d^T$ and $\Theta_{\Gamma c} = \Gamma c^T$ are not both zero;

(b) $||\Lambda||_2 = 1$ and $||\Gamma||_2 = 1$, where $||\cdot||_2$ is the Euclidean vector norm;

(c) first non-zero elements of Λ and Γ are positive.

Let

$$\begin{aligned}\theta &= (\gamma_0 c_1,\ldots,\gamma_0 c_m,\ldots,\gamma_n c_1,\ldots,\gamma_n c_m,\\ &\quad \lambda_1 d_1,\ldots,\lambda_1 d_q,\ldots,\lambda_p d_1,\ldots,\lambda_p d_q)^T\\ &= (\theta_1,\ldots,\theta_{(n+1)m},\theta_{(n+1)m+1},\ldots,\theta_{(n+1)m+pq})^T\end{aligned} \tag{10}$$

be the vector of aggregated parameters (1) obtained by inserting (2) to (1), and let ϕ_k be the respective generalized input vector

$$\phi_k \tag{11}$$

$$\begin{aligned}&= (f_1(u_k),\ldots,f_1(u_{k-n}),\ldots,f_1(u_{k-1}),\ldots,\\ &\quad f_m(u_{k-n}),g_1(y_{k-1}),\ldots,g_q(y_{k-1}),\ldots,g_1(y_{k-p}),\\ &\quad \ldots,g_q(y_{k-p}))^T.\end{aligned} \tag{12}$$

Thanks to above notation, the description (1)–(2) can be simplified to the form $y_k = \phi_k^T\theta + z_k$, which means that the system remains linear with respect to the parameters. For $k = 1,\ldots,N$, we obtain

$$Y_N = \Phi_N\theta + Z_N, \tag{13}$$

where $Y_N = (y_1,\ldots,y_N)^T$, $\Phi_N = (\phi_1,\ldots,\phi_N)^T$, and $Z_N = (z_1,\ldots,z_N)^T$.

The purpose of identification is to recover the parameters in Λ, Γ, c and d (given by (9)), using the input-output measurements (u_k, y_k) $(k = 1, ..., N)$ of the whole system.

1.3. Comments on the assumptions. The representation (1) belongs to the class of the so-called "equation-error" models, while in practical situations a more complicated case of "output-error" models is often met, i.e.,

$$\begin{cases}\overline{y}_k = \sum_{j=1}^{p} \lambda_j \eta(\overline{y}_{k-j}) + \sum_{i=0}^{n} \gamma_i \mu(u_{k-i}),\\ y_k = \overline{y}_k + \delta_k,\end{cases}$$

with zero-mean disturbance δ_k. Since the resulting noise z_k in (1) results from nonlinear filtering of δ_k, it can be of a relatively high order and may have a non-zero mean. The first problem is omitted by making Assumption 7 in

Section 5. The second one can be simply solved when the constant function is appended to the basis $f_1(\cdot),\ldots,f_m(\cdot)$.

To simplify the presentation, it was assumed that the input process, the nonlinear characteristics and the noise are bounded. In fact, since further analysis assumes only finite fourth-order moments of all signals, the approach can be simply generalized for Lipschitz nonlinearities and most of popular finite-variance distributions of excitations.

As regards the i.i.d. restriction imposed on the input process, it can be weakened, for invertible processes, by e.g., data pre-filtering and the use of specially designed instrumental variables in parameter identification (see Mzyk, 2013).

1.4. Organization of the paper. In Section 2, the least squares based identification algorithm (see Bai, 1998) is presented for white disturbances. Then, the reason of its asymptotic bias is shown for correlated noise. Next, in Section 3, an asymptotically unbiased, instrumental variables based estimate is proposed. The idea originates from linear system theory (see, e.g., Wong and Polak, 1967; Söderström and Stoica, 1983; Sagara and Zhao, 1990; Zhao *et al*., 1991), where the instrumental variables technique is used for identification of simple one-element linear dynamic plants. The proposed method is then compared with the least squares. In particular, the consistency of the proposed estimate is shown, in Section 4, even for correlated disturbances. The form of the optimal instrumental variables is established in Section 5, and the method of their approximate generation is described in Section 6. Also, the asymptotic rate of convergence of the estimate is analyzed.

2. Least squares and SVD approach

For comparison purposes with the instrumental variables method proposed further, let us start from the presentation of a two-stage algorithm based on the least-squares estimation of the aggregated parameter vector and decomposition of the obtained result with the use of the SVD algorithm (see Bai, 1998; Kincaid and Cheney, 2002). The algorithm has the following steps.

The fundamental meaning for the algorithm has the form of SVD representations of the theoretical matrices $\Theta_{\Gamma c}=\Gamma c^T$ and $\Theta_{\Lambda d}=\Lambda d^T$. Each matrix being the product of two vectors has the rank equal to 1, and only one singular value is non-zero, i.e.,

$$\Theta_{\Gamma c}=\sum_{i=1}^{\min(n,m)}\sigma_i\mu_i\nu_i^T$$

and

$$\sigma_1\neq 0,\quad \sigma_2=\cdots=\sigma_{\min(n,m)}=0.$$

Algorithm 1. LS-SVD method.

Step 1. Compute the LS estimate

$$\widehat{\theta}_N^{(LS)}=(\Phi_N^T\Phi_N)^{-1}\Phi_N^TY_N \tag{14}$$

of the aggregated parameter vector θ (see (10) and (13)), and next construct (by the plug-in method) evaluations $\widehat{\Theta}_{\Lambda d}^{(LS)}$ and $\widehat{\Theta}_{\Gamma c}^{(LS)}$ of the matrices $\Theta_{\Lambda d}=\Lambda d^T$ and $\Theta_{\Gamma c}=\Gamma c^T$, respectively (see the condition (a) above).

Step 2. Perform the SVD (Singular Value Decomposition, see Appendix B) of the matrices $\widehat{\Theta}_{\Lambda d}^{(LS)}$ and $\widehat{\Theta}_{\Gamma c}^{(LS)}$:

$$\begin{aligned}\widehat{\Theta}_{\Lambda d}^{(LS)}&=\sum_{i=1}^{\min(p,q)}\delta_i\widehat{\xi}_i\widehat{\zeta}_i^T,\\ \widehat{\Theta}_{\Gamma c}^{(LS)}&=\sum_{i=1}^{\min(n,m)}\sigma_i\widehat{\mu}_i\widehat{\nu}_i^T,\end{aligned} \tag{15}$$

and next compute the estimates of parameters of particular blocks (see (9)),

$$\begin{aligned}\widehat{\Lambda}_N^{(LS)}&=\operatorname{sgn}(\widehat{\xi}_1[\kappa_{\xi_1}])\widehat{\xi}_1\\ \widehat{\Gamma}_N^{(LS)}&=\operatorname{sgn}(\widehat{\mu}_1[\kappa_{\mu_1}])\widehat{\mu}_1\\ \widehat{c}_N^{(LS)}&=\operatorname{sgn}(\widehat{\mu}_1[\kappa_{\mu_1}])\sigma_1\widehat{\nu}_1,\\ \widehat{d}_N^{(LS)}&=\operatorname{sgn}(\widehat{\xi}_1[\kappa_{\xi_1}])\delta_1\widehat{\zeta}_1,\end{aligned} \tag{16}$$

where $x[k]$ denotes the k-th element of the vector x and $\kappa_{\mathbf{x}}=\min\{k:x[k]\neq 0\}$.

Thus

$$\Theta_{\Gamma c}=\sigma_1\mu_1\nu_1^T, \tag{17}$$

where $\|\mu_1\|_2=\|\nu_1\|_2=1$. The representation of $\Theta_{\Gamma c}$ given by (17) is obviously unique. To obtain Γ, which fulfills the condition (b), one can take $\Gamma=\mu_1$ or $\Gamma=-\mu_1$. The condition (c) guarantees the uniqueness of Γ. The remaining part of decomposition allows computing c. The vectors Λ and d can be obtained from $\Theta_{\Lambda d}$ in a similar way.

The singular value decomposition allows splitting the aggregated matrices of parameters $\widehat{\Theta}_{\Gamma c}^{(LS)}$ and $\widehat{\Theta}_{\Lambda d}^{(LS)}$ into products of two vectors (see (15)) and estimating $\widehat{\Gamma}_N^{(LS)}\widehat{c}_N^{(LS)T}$ and $\widehat{\Lambda}_N^{(LS)}\widehat{d}_N^{(LS)T}$ according to (16). It was shown by Bai (1998) that

$$(\widehat{\mu}_1,\sigma_1\widehat{\nu}_1)=\arg\min_{c\in\mathbb{R}^m,\Gamma\in\mathbb{R}^n}\|\widehat{\Theta}_{\Gamma c}^{(LS)}-\Gamma c^T\|^2, \tag{18}$$

and for the noise-free case ($z_k\equiv 0$) the estimates (16)

equal the true system parameters, i.e.,

$$\begin{aligned}\widehat{\Lambda}_N^{(LS)} &= \Lambda,\\ \widehat{\Gamma}_N^{(LS)} &= \Gamma,\\ \widehat{c}_N^{(LS)} &= c,\\ \widehat{d}_N^{(LS)} &= d.\end{aligned} \tag{19}$$

Moreover, if the noise $\{z_k\}$ is an i.i.d. process, independent of the input $\{u_k\}$, then

$$\begin{aligned}\widehat{\Lambda}_N^{(LS)} &\to \Lambda,\\ \widehat{\Gamma}_N^{(LS)} &\to \Gamma,\\ \widehat{c}_N^{(LS)} &\to c,\\ \widehat{d}_N^{(LS)} &\to d,\end{aligned} \tag{20}$$

with probability 1, as $N \to \infty$.

Remark 1. For a less sophisticated linear ARMAX model $y_k = \sum_{j=1}^{p} d\lambda_j y_{k-j} + \sum_{i=0}^{n} c\gamma_i u_{k-i} + z_k$, where c and d are scalar constants, the vector (10) reduces to $\theta = \left(\Theta_{\Gamma c}^T, \Theta_{\Lambda d}^T\right)^T$, with single column matrices $\Theta_{\Gamma c}$ and $\Theta_{\Lambda d}$. Consequently, the estimate (14) plays the role of the standard least-squares method and the SVD decomposition in (15) guarantees normalization, i.e., $||\Lambda||_2 = 1$ and $||\Gamma||_2 = 1$.

By taking (13) and (14) into account, the estimation error of the vector θ by the least squares can be expressed as follows:

$$\begin{aligned}\Delta_N^{(LS)} &= \widehat{\theta}_N^{(LS)} - \theta\\ &= \left(\Phi_N^T\Phi_N\right)^{-1}\Phi_N^T Z_N\\ &= \left(\frac{1}{N}\sum_{k=1}^{N}\phi_k\phi_k^T\right)^{-1}\left(\frac{1}{N}\sum_{k=1}^{N}\phi_k z_k\right).\end{aligned} \tag{21}$$

If $\{z_k\}$ is a zero-mean white noise with finite variance, independent of $\{u_k\}$, then all elements of the vector Z_N are independent of the elements of the matrix Φ_N and from the ergodicity of the noise and the process $\{\phi_k\}$ get that $\Delta_N^{(LS)} \to 0$ with probability 1, as $N \to \infty$. Nevertheless, if $\{z_k\}$ is correlated, i.e., $Ez_kz_{k+i} \neq 0$ for some $i \neq 0$, then the LS estimate (14) of θ is not consistent because of the dependence between z_k and the values $g_l(y_{k-i})$ ($l = 1, \ldots, q$ and $i = 1, \ldots, p$) included in ϕ_k. Consequently, the estimates given by (16) are not consistent, either.

3. Instrumental variables approach

As was shown by Hasiewicz and Mzyk (2009), for any Hammerstein system, the bias can be reduced by the instrumental variables method, known from linear system theory. This result was generalized by Mzyk (2013) for a correlated input. In this paper, a similar approach is proposed for more general systems, including nonlinear feedback.

Let us assume that we have given, or we are able to generate, an additional matrix Ψ_N of instrumental variables, which fulfills (even for correlated z_k) the following conditions (see Wong and Polak, 1967; Finigan and Rowe, 1974; Ward, 1977; Hansen and Singleton, 1982; Söderström and Stoica, 1983; 2002; Kowalczuk and Kozłowski, 2000; Hasiewicz and Mzyk, 2009):

(C1) $\dim\Psi_N = \dim\Phi_N$, and the elements of $\Psi_N = (\psi_1, \psi_2, \ldots, \psi_N)^T$, where $\psi_k = (\psi_{k,1}, \psi_{k,2}, \ldots, \psi_{k,m(n+1)+pq})^T$, are jointly bounded, i.e., there exists $0 < \psi_{\max} < \infty$ such that $|\psi_{k,j}| \leq \psi_{\max}$ ($k = 1, \ldots, N$ and $j = 1, \ldots, m(n+1)+pq$) and $\psi_{k,j}$ are ergodic, not necessarily zero-mean, processes;

(C2) there exists $\text{Plim}(\frac{1}{N}\Psi_N^T\Phi_N) = E\psi_k\phi_k^T$ and the limit is not singular, i.e., $\det\{E\psi_k\phi_k^T\} \neq 0$;

(C3) $\text{Plim}(\frac{1}{N}\Psi_N^T Z_N) = E\psi_k z_k$ and $E\psi_k z_k = \text{cov}(\psi_k, z_k) = 0$ (see Assumption 4).

Lemma 1. *A necessary condition for the existence of the instrumental variables matrix Ψ_N, which fulfills (C2) is the asymptotic non-singularity of $\frac{1}{N}\Phi_N^T\Phi_N$.*

Proof. For the proof, see Appendix A. ■

Premultiplying (13) by Ψ_N^T, we get

$$\Psi_N^T Y_N = \Psi_N^T\Phi_N\theta + \Psi_N^T Z_N.$$

Taking into account the conditions (C1)–(C3), a natural idea is to replace the LS estimate, given by (14) and computed in Step 1 (see Section 2), with the instrumental variables estimate

$$\widehat{\theta}_N^{(IV)} = (\Psi_N^T\Phi_N)^{-1}\Psi_N^T Y_N. \tag{22}$$

Step 2 is analogous, i.e., the SVD decomposition is made for the estimates $\widehat{\Theta}_{\Lambda d}^{(IV)}$ and $\widehat{\Theta}_{\Gamma c}^{(IV)}$ of matrices $\Theta_{\Lambda d}$ and $\Theta_{\Gamma c}$, obtained on the basis of $\widehat{\theta}_N^{(IV)}$.

4. Limit properties

For the algorithm (22) the estimation error of the aggregated parameter vector θ has the form

$$\begin{aligned}\Delta_N^{(IV)} &= \widehat{\theta}_N^{(IV)} - \theta\\ &= \left(\Psi_N^T\Phi_N\right)^{-1}\Psi_N^T Z_N\\ &= \left(\frac{1}{N}\sum_{k=1}^{N}\psi_k\phi_k^T\right)^{-1}\left(\frac{1}{N}\sum_{k=1}^{N}\psi_k z_k\right).\end{aligned} \tag{23}$$

Theorem 1. *Under (C1)–(C3), the estimate (22) converges in probability to the true parameters of the system, independently of the autocorrelation of the noise, i.e.,*

$$\operatorname*{Plim}_{N\to\infty} \Delta_N^{(IV)} = 0. \tag{24}$$

Proof. For the proof, see Appendix A. ■

Theorem 2. *The estimation error $\Delta_N^{(IV)}$ converges to zero with the asymptotic rate $O(1/\sqrt{N})$ in probability, for each strategy of instrumental variable generation, which guarantees the fulfillment of (C1)–(C3).*

Proof. For the proof, see Appendix A. ■

5. Optimal instrumental variables

Theorem 2 gives a universal guaranteed asymptotic rate of convergence of the estimate (22). Nevertheless, for a moderate number of measurements, the error depends on particular instruments used in a given application. In this section, a optimal form of instruments is established for the special case of NARMAX systems, which fulfills the following assumption concerning $\eta()$ and $\{\lambda_j\}_{j=1}^p$.

Assumption 7. The nonlinear characteristic $\eta()$ is a Lipschitz function, i.e.,

$$|\eta(y^{(1)}) - \eta(y^{(2)})| \le r|y^{(1)} - y^{(2)}|, \tag{25}$$

and

$$\eta(0) = 0. \tag{26}$$

Moreover, the constant $r > 0$ is such that

$$\alpha = r\sum_{j=1}^{p} |\lambda_j| < 1. \tag{27}$$

Let us consider the following conditional processes (cf. (2)):

$$G_{l,k} \triangleq E\{g_l(y_k) \mid \{u_{k-i}\}_{i=0}^{\infty}\}, \tag{28}$$

where $l = 1, 2, \ldots, q$, and write

$$\xi_l \triangleq g_l(y) - G_l.$$

We have

$$g_l(y_k) = G_{l,k} + \xi_{l,k},$$

and the signals

$$\xi_{l,k} = g_l(y_k) - G_{l,k}, \tag{29}$$

for $l = 1, 2, \ldots, q$ and $k = 1, 2, \ldots, N$, will be interpreted as the "noise". Equation (1) can now be presented as follows:

$$\begin{aligned} y_k &= \sum_{j=1}^{p} \lambda_j \eta(y_{k-j}) + \sum_{i=0}^{n} \gamma_i \mu(u_{k-i}) + z_k \\ &= A_k\left(\{y_{k-j}\}_{j=1}^{p}\right) + B_k\left(\{u_{k-i}\}_{i=1}^{n}\right) \\ &\quad + C_k\left(u_k\right) + z_k, \end{aligned} \tag{30}$$

where

$$\begin{aligned} A_k\left(\{y_{k-j}\}_{j=1}^{p}\right) &= \sum_{j=1}^{p} \lambda_j \eta(y_{k-j}), \\ B_k\left(\{u_{k-i}\}_{i=1}^{n}\right) &= \sum_{i=1}^{n} \gamma_i \mu(u_{k-i}), \\ C_k\left(u_k\right) &= \gamma_0 \mu(u_k). \end{aligned}$$

The random variables A_k, B_k and z_k are independent of the input u_k (see Assumptions 1–6). For a fixed $u_k = u$, we get $C_k\left(u\right) = \gamma_0 \mu(u)$. The expectation in (28) has the following interpretation:

$$\begin{aligned} G_{l,k} = E\Big\{ & g_l(C_k\left(u_k\right) + A_k\left(\{y_{k-j}\}_{j=1}^{p}\right) \\ & + B_k\left(\{u_{k-i}\}_{i=1}^{n}\right) + z_k) \mid \{u_i\}_{i=-\infty}^{k}\Big\}, \end{aligned} \tag{31}$$

and cannot be computed explicitly. However, as will be shown further, the relation between $G_{l,k}$ and the characteristics $\mu(\cdot)$, $\eta(\cdot)$ is not needed. The most significant are the properties below.

(P1) The "disturbances" $\{\xi_{l,k}\}_{k=1}^{N}$ given by (29) are independent of the input process $\{u_k\}$ and are all ergodic.

The mutual independence of $\{\xi_{l,k}\}_{k=1}^{N}$ and $\{u_k\}_{k=-\infty}^{\infty}$ is a direct consequence of the definition (28). On the basis of Assumptions 3–5, the output $\{y_k\}_{k=1}^{N}$ of the system is bounded and ergodic. Owing to Assumption 1, concerning the nonlinear characteristics, the processes $\{g_l\left(y_k\right)\}_{k=1}^{N}$ and $\{G_{l,k}\}_{k=1}^{N}$ $(l = 1, 2, \ldots, q)$ are also bounded and ergodic. Consequently, the "noises" $\{\xi_{l,k}\}_{k=1}^{N}$ $(l = 1, 2, \ldots, q)$, as the sums of ergodic processes, are ergodic too (see (29)).

(P2) The processes $\{\xi_{l,k}\}$ are zero-mean.

By the definition (29) of $\xi_{l,k}$, we have

$$\begin{aligned} E\xi_{l,k} &= Eg_l(y_k) - EG_{l,k} \\ &= E_{\{u\}_{j=-\infty}^{k}} E\left\{g_l(y_k) \mid \{u\}_{i=-\infty}^{k}\right\} \\ &\quad - E_{\{u\}_{j=-\infty}^{k}} E\left\{g_l(y_k) \mid \{u\}_{i=-\infty}^{k}\right\} = 0. \end{aligned}$$

(P3) If the instrumental variables $\psi_{k,j}$ are generated by the nonlinear filtering

$$\psi_{k,j} = H_j(\{u_i\}_{i=-\infty}^{k}), \tag{32}$$

where the transformations $H_j(\cdot)$ $(j = 1, 2, \ldots, m(n+1) + pq)$ guarantee the ergodicity of $\{\psi_{k,j}\}$, then all products $\psi_{k_1,j}\xi_{l,k_2}$ $(j = 1, 2, \ldots, m(n+1) + pq,\ l = 1, 2, \ldots, q)$ are zero-mean, i.e., $E\psi_{k_1,j}\xi_{l,k_2} = 0$.

Owing to *(P1)* and *(P2)*, we have

$$\begin{aligned} E\left[\psi_{k_1,j}\xi_{l,k_2}\right] &= E\left[H_j(\{u_i\}_{i=-\infty}^{k_1})\xi_{l,k_2}\right] \\ &= EH_j(\{u_i\}_{i=-\infty}^{k_1})E\xi_{l,k_2} = 0. \end{aligned}$$

(P4) If the measurement noise z_k and the instrumental variables $\psi_{k,j}$ are bounded (i.e., Assumption 4 and the condition (C1) are fulfilled), i.e., $|z_k| < z_{\max} < \infty$ and $|\psi_{k,j}| = |H_j(u_k)| < \psi_{\max} < \infty$ (see 3), then

$$\frac{1}{N}\sum_{k=1}^{N}\psi_k z_k \to E\psi_k z_k, \tag{33}$$

with probability 1, as $N \to \infty$ (cf. the condition (C3)).

The product $s_{k,j} = \psi_{k,j} z_k$ of stationary and bounded signals $\psi_{k,j}$ and z_k is also stationary, with finite variance. To prove (33) making use of Lemma B.1 by Söderström and Stoica (1989), it must be shown that $r_{s_{k,j}}(\tau) \to 0$, as $|\tau| \to \infty$. Let us notice that the autocovariance function of z_k ($Ez_k = 0$),

$$r_z(\tau) = E[(z_k - Ez)(z_{k+\tau} - Ez)] = Ez_k z_{k+\tau}, \tag{34}$$

as the output of linear filter excited by a white noise has the property that

$$r_z(\tau) \to 0, \tag{35}$$

as $|\tau| \to \infty$. Hence, the processes $\psi_{k,j} = H_j\left(\{u_i\}_{i=-\infty}^{k}\right)$ are ergodic (see (P3)), and independent of z_k (see Assumption 4). Thus

$$\begin{aligned} r_{s_{k,j}}(\tau) &= E[(s_{k,j} - Es_{k,j})(s_{k+\tau,j} - Es_{k,j})] \\ &= E[\psi_{k,j}\psi_{k+\tau,j} z_k z_{k+\tau}] = c r_z(\tau), \end{aligned} \tag{36}$$

where $c = (E\psi_{k,j})^2$ is a finite constant, $0 \le c < \infty$. Consequently,

$$r_{s_{k,j}}(\tau) \to 0, \tag{37}$$

as $|\tau| \to \infty$, and

$$\frac{1}{N}\sum_{k=1}^{N} s_{k,j} \to Es_{k,j}, \tag{38}$$

with probability 1, as $N \to \infty$.

(P5a) For the NARMAX system with the characteristic $\eta()$ as in Assumption 7 and the order of autoregression $p = 1$ (see Eqn. (1)), it holds that

$$\frac{1}{N}\sum_{k=1}^{N}\psi_k\phi_k^T \to E\psi_k\phi_k^T, \tag{39}$$

with probability 1 as $N \to \infty$, where ψ_k is given by (32); compare the condition (C2).

For $p = 1$ (for clarity of presentation, let also $\lambda_1 = 1$) the system is described by

$$y_k = \eta(y_{k-1}) + \sum_{i=0}^{n}\gamma_i\mu(u_{k-i}) + z_k, \tag{40}$$

and the nonlinearity $\eta()$, according to Assumption 7, fulfills the condition

$$|\eta(y)| \le a|y|, \tag{41}$$

where $0 < a < 1$. Introducing the symbol

$$\delta_k = \sum_{i=0}^{n}\gamma_i\mu(u_{k-i}) + z_k, \tag{42}$$

we get

$$y_k = \eta(y_{k-1}) + \delta_k. \tag{43}$$

Since the input $\{u_k\}$ is an i.i.d. sequence, independent of $\{z_k\}$, and the noise $\{z_k\}$ has the property that $r_z(\tau) \to 0$, as $|\tau| \to \infty$ (see (35)). There holds $r_\delta(\tau) \to 0$ as $|\tau| \to \infty$. Equation (43) can be written in the following form:

$$y_k = \delta_k + \eta\{\delta_{k-1} + \eta[\delta_{k-2} + \eta(\delta_{k-3} + \dots)]\}. \tag{44}$$

Let us introduce the coefficients c_k defined, for $k = 1, 2, \dots, N$, as

$$c_k = \frac{\eta(y_k)}{y_k}, \tag{45}$$

with $0/0$ treated as 0. Owing to (41), we have

$$|c_k| \le a < 1, \tag{46}$$

and using c_k Eqn. (44) can be rewritten as follows:

$$\begin{aligned} y_k = \delta_k + c_{k-1}\Big(\delta_{k-1} + c_{k-2}\big(\delta_{k-2} \\ + c_{k-3}(\delta_{k-3} + \dots)\big)\Big), \end{aligned}$$

i.e.,

$$y_k = \sum_{i=0}^{\infty} c_{k,i}\delta_{k-i},$$

where $c_{k,0} \triangleq 1$, and $c_{k,i} = c_{k-1}c_{k-2}\dots,c_{k-i}$. From (46) we conclude that

$$|c_{k,i}| < a^i. \tag{47}$$

Since for $0 < a < 1$ the sum $\sum_{i=0}^{\infty} a^i$ is finite, from (47) we get $\sum_{i=0}^{\infty}|c_{k,i}| < \infty$, and from (42) we simply conclude that for $|\tau| \to \infty$ we have $r_y(\tau) \to 0$ and $r_{g_l(y_k)}(\tau) \to 0$, where the processes $g_l(y_k)$ ($l = 1, \dots, q$) are elements of the vector ϕ_k. Thus, for the system with the nonlinearity $\eta(\cdot)$ as in (41), the processes $\{y_k\}$ and $\{g_l(y_k)\}$ ($l = 1, \dots, q$) fulfill the assumption of the ergodic law of large numbers, and the property (39) holds.

(P5b) Under Assumption 7, the convergence (39) takes place also for the system (1) with $p \ge 1$.

For any number sequence $\{x_k\}$, let us define the norm

$$\|\{x_k\}\| = \lim_{K\to\infty}\sup_{k>K}|x_k|, \tag{48}$$

and let us present Eqn. (1) in the form

$$y_k = \sum_{j=1}^{p}\lambda_j\eta(y_{k-j}) + \delta_k, \tag{49}$$

where δ_k is given by (42). The proof of the property (P5b) (for $p > 1$) is based of the following theorem (see Kudrewicz, 1976, p. 53).

Theorem 3. *Let $\{y_k^{(1)}\}$ and $\{y_k^{(2)}\}$ be two different output sequences of the system (1) (see also (49)), and $\{\delta_k^{(1)}\}$, $\{\delta_k^{(2)}\}$ be respective aggregated inputs (see (42)). If (25), (26) and (27) are fulfilled, then*

$$\frac{\|\{\delta_k^{(1)} - \delta_k^{(2)}\}\|}{1+\alpha} \leq \|\{y_k^{(1)} - y_k^{(2)}\}\| \leq \frac{\|\{\delta_k^{(1)} - \delta_k^{(2)}\}\|}{1-\alpha}, \tag{50}$$

where the norm $\|\cdot\|$ is defined in (48).

From (50) and under the conditions (25)–(27), the steady state of the system (1) depends only on the steady state of the input $\{\delta_k\}$. The special case of (50) is $\delta_k^{(2)} \equiv 0$, in which $\lim_{K\to\infty} \sup_{k>K} \left|y_k^{(2)}\right| = 0$, and

$$\frac{1}{1+\alpha}\|\{\delta_k^{(1)}\}\| \leq \|\{y_k^{(1)}\}\| \leq \frac{1}{1-\alpha}\|\{\delta_k^{(1)}\}\|.$$

The impulse response of the system tends to zero, as $k \to \infty$, and for an i.i.d. input the autocorrelation function of the output $\{y_k\}$ is such that

$$r_y(\tau) \to 0 \quad \text{as } |\tau| \to \infty.$$

Moreover, on the basis of (1)–(4), since the process $\{y_k\}$ is bounded, it has finite moments of any orders and the ergodic theorems hold (see (Söderström and Stoica, 1989) Definition B.2, Lemma B.1, B.2). In consequence, the convergence (39) holds.

The properties (P5a) and (P5b) (see (39), (12) and (32)) can be rewritten for particular elements of ψ_k and ϕ_k in the following way:

$$\frac{1}{N}\sum_{k=1}^{N} \psi_{k_1,j} g_l(y_{k_2}) \to E\psi_{k_1,j} g_l(y_{k_2}),$$

with probability 1 as $N \to \infty$.

Under the property that $E\left[\psi_{k_1,j}\xi_{l,k_2}\right] = 0$ (see (P3)), for instrumental variables generated according to (32), there obviously holds that

$$E\left[\psi_{k_1,j} g_l(y_{k_2})\right] = E\left[\psi_{k_1,j} G_{l,k_2}\right].$$

Writing (cf. (12))

$$\begin{aligned} \Phi_N^{\#} &= (\phi_1^{\#}, \phi_2^{\#}, \ldots, \phi_N^{\#})^T, \\ \phi_k^{\#} &\triangleq (f_1(u_k), \ldots, f_m(u_k), \ldots \\ &\quad f_1(u_{k-n}), \ldots, f_m(u_{k-n}), \\ &\quad G_{1,k-1}, \ldots, G_{q,k-1}, \ldots, \\ &\quad G_{1,k-p}, \ldots, G_{q,k-p})^T, \end{aligned} \tag{51}$$

where $G_{l,k} \triangleq E\{g_l(y_k) \mid \{u_i\}_{i=-\infty}^{k}\}$ (see (28)), and making use of the ergodicity of the processes $\{\psi_{k,j}\}$ $(j = 1, \ldots, m(n+1)+pq)$, $\{f_t(u_k)\}$ $(t = 1, \ldots, m)$ and $\{G_{l,k}\}$ $(l = 1, \ldots, q)$ (see (32) and Assumption 3), we get

$$\frac{1}{N}\Psi_N^T\Phi_N^{\#} = \frac{1}{N}\sum_{k=1}^{N} \psi_k \phi_k^{\#T} \to E\psi_k\phi_k^{\#T} \text{ with p. 1,}$$

and, using (39), we get

$$\frac{1}{N}\Psi_N^T\Phi_N = \frac{1}{N}\sum_{k=1}^{N} \psi_k \phi_k^{T} \to E\psi_k\phi_k^{T} \text{ with p. 1,}$$

for the instruments as in (32).

Directly from the definitions (28) and (51), we conclude that $E\left[\psi_{k_1,j} g_l(y_{k_2})\right] = E\left[\psi_{k_1,j} G_{l,k_2}\right]$ and

$$E\psi_k\phi_k^{\#T} = E\psi_k\phi_k^{T}.$$

Thus, for any choice of instrumental variables matrix Ψ_N, which fulfills the property (P3) (see (32)), the following equivalence takes place asymptotically with probability 1, as $N \to \infty$:

$$\frac{1}{N}\Psi_N^T\Phi_N^{\#} = \frac{1}{N}\Psi_N^T\Phi_N. \tag{52}$$

The estimation error (i.e., the difference between the estimate and the true value of parameters) has the form

$$\Delta_N^{(IV)} = \widehat{\theta}_N^{(IV)} - \theta = \left(\frac{1}{N}\Psi_N^T\Phi_N\right)^{-1}\left(\frac{1}{N}\Psi_N^T Z_N\right).$$

Introducing

$$\begin{aligned} \Gamma_N &\triangleq \left(\frac{1}{N}\Psi_N^T\Phi_N\right)^{-1}\frac{1}{\sqrt{N}}\Psi_N^T, \\ Z_N^* &\triangleq \frac{\frac{1}{\sqrt{N}}Z_N}{z_{\max}}, \end{aligned}$$

where $z_{\max}$ is an upper bound of the absolute value of the noise (see Assumption 4), we obtain that

$$\Delta_N^{(IV)} = z_{\max}\Gamma_N Z_N^*, \tag{53}$$

with the Euclidean norm of Z_N^*,

$$\|Z_N^*\| = \sqrt{\sum_{k=1}^{N}\left(\frac{\frac{1}{\sqrt{N}}z_k}{z_{\max}}\right)^2} = \sqrt{\frac{1}{N}\sum_{k=1}^{N}\left(\frac{z_k}{z_{\max}}\right)^2} \leq 1.$$

Let the quality of the instrumental variables be evaluated on the basis of the following criterion (see, e.g., Wong and Polak, 1967)

$$Q\left(\Psi_N\right) = \max_{\|Z_N^*\|\leq 1}\left\|\Delta_N^{(IV)}\left(\Psi_N\right)\right\|^2, \tag{54}$$

where $\|\cdot\|$ denotes the Euclidean norm, and $\Delta_N^{(IV)}\left(\Psi_N\right)$ is the estimation error obtained for the instrumental variables Ψ_N.

Theorem 4. *If Assumptions 1–7 and the condition (32) hold, then the criterion $Q(\Psi_N)$ given by (54) attains a minimum for the choice*

$$\Psi_N^{\#} = \Phi_N^{\#}, \tag{55}$$

i.e., for each Ψ_N,

$$\lim_{N\to\infty} Q(\Psi_N^{\#}) \le \lim_{N\to\infty} Q(\Psi_N) \text{ with prob. } 1.$$

Proof. For the proof, see Appendix A. ■

Obviously, instrumental variables given by (55) fulfill the postulates (C1)–(C3).

6. Nonparametric generation of instrumental variables

The optimal matrix of instruments $\Psi_N^{\#}$ cannot be computed analytically, because of the lack of prior knowledge of the system (the probability density functions of excitations and the values of parameters are unknown). Estimation of $\Psi_N^{\#}$ is also difficult, because the elements $G_{l,k}$ depend on an infinite number of measurements of the input process. Therefore, the only choice is the following FIR approximation:

$$\begin{aligned}
\Psi_N^{(r)\#} &= (\psi_1^{(r)\#}, \psi_2^{(r)\#}, \dots, \psi_N^{(r)\#})^T,\\
\psi_k^{(r)\#} &\triangleq (f_1(u_k), \dots, f_m(u_k), \dots,\\
&\quad f_1(u_{k-n}), \dots, f_m(u_{k-n}),\\
&\quad G_{1,k-1}^{(r)}, \dots, G_{q,k-1}^{(r)}, \dots,\\
&\quad G_{1,k-p}^{(r)}, \dots, G_{q,k-p}^{(r)})^T,
\end{aligned}$$

where r is a cut-off level in (28), i.e.,

$$G_{l,k}^{(r)} = E\{g_l(y_k) \mid \{u_{k-i}\}_{i=0}^{r}\}. \tag{56}$$

It is based on the intuition that the approximate value $\Psi_N^{(r)\#}$ becomes better, i.e.,

$$\Psi_N^{(r)\#} \cong \Psi_N^{\#}$$

when r is increasing (this question is treated as open). The simplest realization of the algorithm (i.e., for $r = 0$) has the form

$$\begin{aligned}
\Psi_N &= \Psi_N^{(0)\#},\\
\psi_k^{(0)\#} &\triangleq (f_1(u_k), \dots, f_m(u_k), \dots,\\
&\quad f_1(u_{k-n}), \dots, f_m(u_{k-n}),\\
&\quad R_1(u_{k-1}), \dots, R_q(u_{k-1}), \dots,\\
&\quad R_1(u_{k-p}), \dots, R_q(u_{k-p}))^T,
\end{aligned}$$

where

$$R_l(u) = G_l^{(0)}(u) = E\{g_l(y_k)\} \mid u_k = u\}. \tag{57}$$

All elements of $\psi_k^{(0)\#}$ (white noises) fulfill (P3). After introducing

$$x_{l,k} = g_l(y_k),$$

the regression functions in (57) can be written as

$$R_l(u) = E\{x_{l,k} \mid u_k = u\}.$$

Both u_k and y_k can be measured, and $x_{l,k} = g_l(y_k)$ can be computed, because the functions $g_l()$ are known a priori. Thus the most natural method for generation of $\Psi_N^{(r)\#}$ is the kernel method. A traditional estimate of the regression function $R_l(u)$ computed on the basis of M pairs $\{(u_i, x_{l,i})\}_{i=1}^{M}$ has the form (see, e.g., Greblicki and Pawlak, 2008)

$$\widehat{R}_{l,M}(u) = \frac{\frac{1}{M}\sum_{i=1}^{M} x_{l,i} K\left(\frac{u-u_i}{h(M)}\right)}{\frac{1}{M}\sum_{i=1}^{M} K\left(\frac{u-u_i}{h(M)}\right)}, \tag{58}$$

where K is a kernel function, and h is the bandwidth parameter.

Further deliberations will be based on the following two theorems (see Greblicki and Pawlak, 2008).

Theorem 5. *If $h(M) \to 0$ and $Mh(M) \to \infty$ as $M \to \infty$, and $K(v)$ is one of $\exp(-|v|)$, $\exp(-v^2)$, or $(1+|v|^{1+\delta})^{-1}$, then*

$$\frac{\frac{1}{M}\sum_{i=1}^{M} y_i K\left(\frac{u-u_i}{h(M)}\right)}{\frac{1}{M}\sum_{i=1}^{M} K\left(\frac{u-u_i}{h(M)}\right)} \to E\{y_i \mid u_i = u\} \tag{59}$$

in probability as $M \to \infty$, provided that $\{(u_i, y_i)\}_{i=1}^{M}$ is an i.i.d. sequence.

Theorem 6. *If both the regression $E\{y_i \mid u_i = u\}$ and the input probability density function $\vartheta(u)$ have finite second order derivatives, then for $h(M) = O(M^{-\frac{1}{5}})$ the asymptotic rate of convergence is $O(M^{-\frac{2}{5}})$ in probability.*

To apply the above theorems, let us additionally make the following assumption.

Assumption 8. The functions $g_1(y),\dots,g_q(y)$, $f_1(u),\dots,f_m(u)$ and the input probability density $\vartheta(u)$ have finite second order derivatives for each $u \in (-u_{\max}, u_{\max})$ and each $y \in (-y_{\max}, y_{\max})$.

In our problem, the process $\{x_{l,i}\}$ appearing in the numerator of (58) is correlated. Let us decompose the sums in the numerator and denominator in (58) for $r = \left\lfloor M^{\frac{1}{\chi(M)}} \right\rfloor$ partial sums, where $\chi(M)$ is such that $\chi(M) \to \infty$ and $r \to \infty$, as $M \to \infty$ (e.g., $\chi(M) =$

$\sqrt{\log M}$), i.e.,

$$L(\{(u_i,x_{l,i})\}_{i=1}^M) \triangleq \frac{1}{M}\sum_{i=1}^M x_{l,i}K\left(\frac{u-u_i}{h(M)}\right)=\frac{1}{r}\sum_{t=1}^r s_t, \tag{60}$$

$$W(\{u_i\}_{i=1}^M) \triangleq \frac{1}{M}\sum_{i=1}^M K\left(\frac{u-u_i}{h(M)}\right)=\frac{1}{r}\sum_{t=1}^r w_t,$$

with

$$s_t=\frac{1}{M/r}\sum_{\{i:0<ir+t\le M\}} x_{l,ir+t}K\left(\frac{u-u_{ir+t}}{h(M)}\right), \tag{61}$$

$$w_t=\frac{1}{M/r}\sum_{\{i:0<ir+t\le M\}} K\left(\frac{u-u_{ir+t}}{h(M)}\right).$$

The components of the sum (60) have the time distance r and become uncorrelated as $r\to\infty$. This fact is a simple consequence of the property that $r_x(\tau)\to 0$, as $|\tau|\to\infty$. Moreover, the components in (61) are i.i.d. Each of the partial sums $\{s_t\}$ has the same probability density, but uses a different subset of measurements. All of them include $\overline{M}=M/r$ data.

For simplicity, let us write

$$s_t=\frac{1}{\overline{M}}\sum_{\{i:0<ir+t\le M\}} x_{l,ir+t}K\left(\frac{u-u_{ir+t}}{H(\overline{M})}\right), \tag{62}$$

$$w_t=\frac{1}{\overline{M}}\sum_{\{i:0<ir+t\le M\}} K\left(\frac{u-u_{ir+t}}{H(\overline{M})}\right),$$

where $H(\overline{M})\triangleq h(M)$. Let $h(M)=cM^\alpha$, where $-1<\alpha<0$. Then

$$H(\overline{M})=cM^\alpha \quad =c\left(\overline{M}^\alpha\right)^{\frac{1}{1-\frac{1}{\chi(M)}}}=O(\overline{M}^\alpha), \tag{63}$$

and as $\overline{M}\to\infty$, we get

$$H(\overline{M})\to 0 \quad \text{and} \quad \overline{M}H(\overline{M})\to\infty. \tag{64}$$

From (62)–(64) and Theorem 5, as $r\to\infty$, we get

$$\operatorname*{Plim}_{M\to\infty}\left(\frac{s_t}{w_t}\right)=\frac{\operatorname{Plim}_{\overline{M}\to\infty}(s_t)}{\operatorname{Plim}_{\overline{M}\to\infty}(w_t)}=\frac{a(u)}{b(u)}=R_l(u),$$

for each $t=1,2,\dots,r$, and since

$$\widehat{R}_{l,M}(u)=\frac{L(\{(u_i,x_{l,i})\}_{i=1}^M}{W(\{u_i\}_{i=1}^M)}=\frac{\frac{1}{r}\sum_{t=1}^r s_t}{\frac{1}{r}\sum_{t=1}^r w_t}$$

we obtain

$$\operatorname*{Plim}_{M\to\infty}\left(\widehat{R}_{l,M}(u)\right)=R_l(u). \tag{65}$$

Under Assumption 8, from the property (63) and Theorem 6 we conclude that for $h(M)=cM^{-\frac{1}{5}}$ the rate of convergence of (58) is $O(M^{-\frac{2}{5}})$ in probability.

7. Three-stage identification

Taking into account the conclusions from Section 6, in particular the form of optimal instruments Ψ_N^*, the following combined parametric-nonparametric identification procedure is proposed in the paper (see Mzyk, 2007; 2009).

Stage 1. (Nonparametric) Using $M+\max(n,p)$ measurements $\{(u_i,y_i)\}_{i=1-\max(n,p)}^M$, generate the empirical matrix of instruments

$$\widehat{\Psi}_{N,M}^*=(\widehat{\psi}_{1,M}^*,\widehat{\psi}_{2,M}^*,\dots,\widehat{\psi}_{N,M}^*)^T,$$

where

$$\begin{aligned}\widehat{\psi}_{k,M}^*=(&f_1(u_k),\dots,f_m(u_k),\dots,\\ &f_1(u_{k-n}),\dots,\dots,f_m(u_{k-n}),\\ &\widehat{R}_{1,M}(u_{k-1}),\dots,\widehat{R}_{q,M}(u_{k-1}),\dots,\\ &\widehat{R}_{1,M}(u_{k-p}),\dots,\widehat{R}_{q,M}(u_{k-p}))^T,\end{aligned} \tag{66}$$

and

$$\widehat{R}_{l,M}(u)=\sum_{i=1}^M g_l(y_i)K(\frac{u-u_i}{h(M)})/\sum_{i=1}^M K(\frac{u-u_i}{h(M)}).$$

Stage 2. (Parametric) Estimate the aggregated parameter vector (10)

$$\begin{aligned}\theta=(&\gamma_0c_1,\dots,\gamma_oc_m,\dots,\gamma_nc_1,\dots,\gamma_nc_m,\\ &\lambda_1d_1,\dots,\lambda_1d_q,\dots,\lambda_pd_1,\dots,\lambda_pd_q)^T\end{aligned}$$

by the instrumental variables method

$$\widehat{\theta}_{N,M}^{*(IV)}=\left(\widehat{\Psi}_{N,M}^{*T}\Phi_N\right)^{-1}\widehat{\Psi}_{N,M}^{*T}Y_N, \tag{67}$$

where

$$Y_N=(y_1,y_2,\dots,y_N)^T,$$

$$\Phi_N=(\phi_1,\phi_2,\dots,\phi_N)^T,$$

$$\begin{aligned}\phi_k=(&f_1(u_k),\dots,f_m(u_k),\dots,\\ &f_1(u_{k-n}),\dots,f_m(u_{k-n}),\\ &g_1(y_{k-1}),\dots,g_q(y_{k-1}),\dots,\\ &g_1(y_{k-p}),\dots,g_q(y_{k-p}))^T,\end{aligned}$$

(see (12)), and next, using $\widehat{\theta}_{N,M}^{*(IV)}$, construct the estimates $\widehat{\Theta}_{\lambda d}^{(IV)}$ and $\widehat{\Theta}_{\gamma c}^{(IV)}$ of the matrices $\Theta_{\lambda d}=\Lambda d^T$ and $\Theta_{\gamma c}=\Gamma c^T$.

Stage 3. (Decomposition) Compute the SVD of the matrices $\widehat{\Theta}_{\lambda d}^{(IV)}$ and $\widehat{\Theta}_{\gamma c}^{(IV)}$, i.e., $\widehat{\Theta}_{\gamma c}^{(IV)}=\sum_{i=1}^{\min(n,m)}\sigma_i\widehat{\mu}_i\widehat{\nu}_i^T$, $\widehat{\Theta}_{\lambda d}^{(IV)}=\sum_{i=1}^{\min(p,q)}\delta_i\widehat{\xi}_i\widehat{\zeta}_i^T$ to obtain the estimates of the parameters (elements of the impulse

responses of the linear dynamic blocks and the parameters of static nonlinear characteristics)

$$\widehat{\Lambda}_N = \operatorname{sgn}(\widehat{\xi}_1[\kappa_{\xi_1}])\widehat{\xi}_1, \qquad \widehat{\Gamma}_N = \operatorname{sgn}(\widehat{\mu}_1[\kappa_{\mu_1}])\widehat{\mu}_1, \tag{68}$$

$$\widehat{c}_N = \operatorname{sgn}(\widehat{\mu}_1[\kappa_{\mu_1}])\sigma_1\widehat{\nu}_1, \qquad \widehat{d}_N = \operatorname{sgn}(\widehat{\xi}_1[\kappa_{\xi_1}])\delta_1\widehat{\zeta}_1,$$

where $x[k]$ denotes the k-th element of the vector x, and $\kappa_{\mathbf{x}} = \min\{k : x[k] \neq 0\}$.

Under the condition (65), the following theorem holds.

Theorem 7. *For the NARMAX system with the characteristic $\eta(y)$ as in Assumption 7 we have*

$$\widehat{\theta}^{*(IV)}_{N,M} \to \theta \quad \textit{in probability}$$

as $M \to \infty$ and $N \to \infty$, provided that $h(M)$ fulfills the assumptions of Theorem 5.

Proof. For the proof, see Appendix A. ■

8. Example

8.1. Simulation. The simulated system was a special case of the model (1), commonly known in the literature as a Lur'e system (see Fig. 2), and often met in applications (see Hill and Chong, 1989; Hill and Mareels, 1990; Suykens *et al.*, 1998; Lu and Hill, 2007).

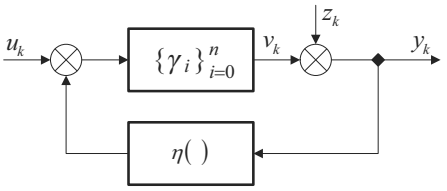

Fig. 2. Lur'e system.

In this case, the static block $\mu(\cdot)$ is linear, i.e., $\mu(u) = u$, and both linear dynamic blocks $\{\gamma_i\}$ and $\{\lambda_j\}$ have the same impulse responses. Thus, in the computer experiment we set

$$\begin{aligned} &n = 3, \\ &\gamma_0 = 0, \qquad \gamma_1 = 1, \qquad \gamma_2 = 1, \\ &p = 2, \\ &\lambda_j = \gamma_j, \qquad j = 1, 2, \end{aligned}$$

and the nonlinear feedback

$$\eta(y) = \frac{1}{4}\,|y|$$

was applied. Since, for the case considered,

$$r = \frac{1}{4}, \qquad \sum_{j=1}^{p} \lambda_j = 2,$$

$$\alpha = r\sum_{j=1}^{p} \lambda_j = \frac{1}{2} < 1,$$

the simulated system is stable (see (27)) and can be described by the following nonlinear difference equation:

$$y_k = u_{k-1} + u_{k-2} + \frac{1}{4}\,|y_{k-1}| + \frac{1}{4}\,|y_{k-2}| + z_k.$$

The system was excited by a uniformly distributed random sequence

$$u_k \sim U\,[-1, 1]\,,$$

and disturbed by the colored noise

$$z_k = \frac{1}{2}z_{k-1} + \varepsilon_k,$$

where $\varepsilon_k \sim U\,[-1, 1]$.

8.2. Identification. The linear model of $\mu(\cdot)$ was assumed,

$$\mu(u) = c_1 u + c_2,$$

i.e.,

$$f_1(u) = u,\ f_2(u) = 1,\ m = 2,$$

and a two-segment piecewise linear model of $\eta(\cdot)$,

$$\eta(y) = d_1 y \cdot 1(y) + d_2 y \cdot 1(-y),$$

i.e.,

$$g_1(y) = y \cdot 1(y),\ g_2(y) = y \cdot 1(-y),\ q = 2,$$

where

$$1(x) = \begin{cases} 1, & \text{if } x \geq 0, \\ 0, & \text{otherwise.} \end{cases}$$

The system with the true vectors of parameters

$$\begin{aligned} \Lambda_{\text{true}} &= (1, 1)^T, \\ \Gamma_{\text{true}} &= (0, 1, 1)^T, \\ c_{\text{true}} &= (1, 0)^T, \\ d_{\text{true}} &= \left(\frac{1}{4}, -\frac{1}{4}\right)^T, \end{aligned}$$

was normalized to the following equivalent version (see

the condition (b) in Section 1.2):

$$\Lambda = \left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right)^T,$$
$$\Gamma = \left(0, \frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right)^T,$$
$$c = (\sqrt{2}, 0)^T,$$
$$d = \left(\frac{\sqrt{2}}{4}, -\frac{\sqrt{2}}{4}\right)^T.$$

The aggregated vector of mixed products of parameters θ and identified matrices $\Theta_{\Lambda d}$ and $\Theta_{\Gamma c}$ are as follows:

$$\theta = \left(0, 0, 1, 0, 1, 0, \frac{1}{4}, -\frac{1}{4}, \frac{1}{4}, -\frac{1}{4}\right)^T,$$

$$\Theta_{\Lambda d} = \begin{bmatrix} \frac{1}{4} & -\frac{1}{4} \\ \frac{1}{4} & -\frac{1}{4} \end{bmatrix},$$

$$\Theta_{\Gamma c} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 1 & 0 \end{bmatrix}.$$

The estimates (14) and (22) were compared, with

$$\phi_k = (u_k, 1, u_{k-1}, 1, u_{k-2}, 1, y_{k-1}1(y_{k-1}), y_{k-1}1(-y_{k-1}), y_{k-2}1(y_{k-2}), y_{k-2}1(-y_{k-2}))^T,$$
$$\widehat{\psi}^*_{k,M} = \left(u_k, 1, u_{k-1}, 1, u_{k-2}, 1, \widehat{R}_{1,M}(u_{k-2}), \widehat{R}_{2,M}(u_{k-2}), \widehat{R}_{1,M}(u_{k-3}), \widehat{R}_{2,M}(u_{k-3})\right)^T,$$

where

$$\widehat{R}_{l,M}(u) = \frac{\frac{1}{M}\sum_{i=1}^{M} g_l(y_{i+1}) K\left(\frac{u-u_i}{h(M)}\right)}{\frac{1}{M}\sum_{i=1}^{M} K\left(\frac{u-u_i}{h(M)}\right)}.$$

The mean normalized errors of both subsystems,

$$MNE_\Gamma = \frac{\|\widehat{\Gamma}_N - \Gamma\|_2}{\|\Gamma\|_2},$$
$$MNE_d = \frac{\|\widehat{d}_N - d\|_2}{\|d\|_2},$$

were computed and averaged over ten re-runs, and for various numbers of measurements. Figures 3 and 4 show that, contrary to the least-squares method, the algorithm is free of an asymptotic bias (i.e., as $N \to \infty$) and converges to true system parameters. The experiment was also repeated for various variances of the noise ε_k. The results for $N = 100$, shown in Fig. 5, confirm a linear increase of the estimation errors, which is typical in 'linear in the parameters' system identification. The results confirm the usability of the proposed scheme.

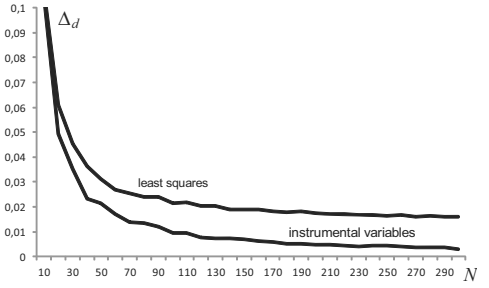

Fig. 3. Estimation error of the nonlinear static block.

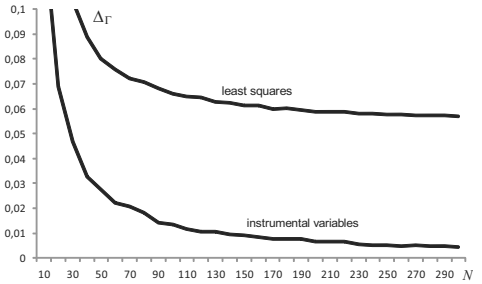

Fig. 4. Estimation error of the linear dynamic block.

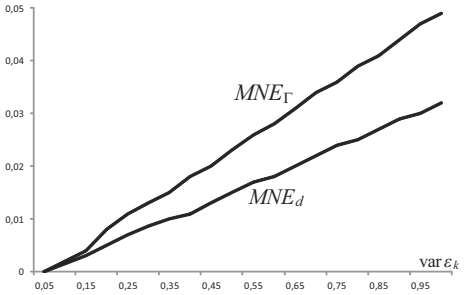

Fig. 5. Estimation error vs. variance of the noise, for $N = 100$ (instrumental variables method).

9. Summary

The advantages of the approach and the contribution of the paper can be summarized as follows. The structure of the system considered is more general than Hammerstein systems and Lur'e systems. Moreover, nonlinear characteristics of static blocks do not have to be of a polynomial form, which is commonly assumed in the literature. Also excitations can have arbitrary correlation

properties. The method, as a whole, is computationally simple, and standard numerical LS/IV procedures (e.g., LU and Cholesky decompositions) can be applied at the main stage of the routine. The consistency of the proposed estimate is proved, even for correlated noise and with correlation between the noise and the input caused by structural feedback. Full versions of the proofs of theorems are included. Good cooperation between parametric and nonparametric methods is shown. The problem of suboptimal generation of instrumental variables is solved by application of nonparametric (kernel) methods. Also the scope of applicability of the instrumental variables technique is extended for nonlinear systems with feedback.

Obviously, the algorithm proposed in the paper has some drawbacks. The most significant is the fact that the class is limited to the 'linear in the parameters' additive NARMAX models, and neither input cross-terms nor lagged noise terms are admitted in the difference equation describing the system. The consistency of the estimate with intuitive approximation $\Psi_N^{(r)}$ of Ψ_N was not proved formally. This issue is treated as open. Moreover, for technical reasons (SVD method), only FIR linear blocks are acceptable. It was also assumed that the input is an i.i.d. sequence. Nevertheless, recent results (see, e.g., Mzyk, 2013) show that the instrumental variables approach can be useful for reducing the bias in the correlated input case.

The presented method can help in identification of more complicated, large-scale interconnected systems (see Fig. 6), and to design the decomposition/coordination algorithms (see, e.g., Findeisen *et al.*, 1980), for nonlinear dynamic models, consisting of n blocks described by

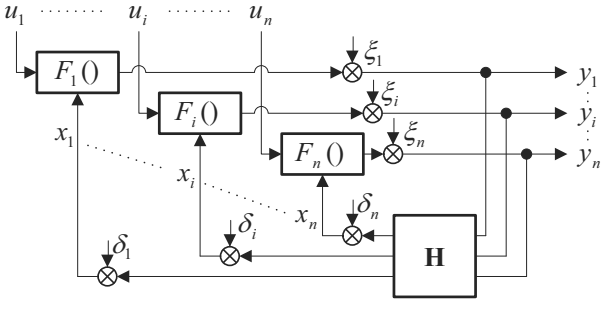

Fig. 6. System with an arbitrary structure.

unknown functionals:

$$F_i\left(\{u_i\},\{x_i\}\right), \qquad i=1,2,\ldots,n,$$

where only external inputs u_i and outputs y_i of the system can be measured. The interactions x_i are hidden, but the structure of connections is known and coded in the zero-one matrix H, i.e.,

$$x_i = H_i \cdot (y_1 y_2, \ldots, y_n)^T + \delta_i,$$

where H_i denotes the i-th row of H and δ_i is a random disturbance. In the simplest case of static linear system (see Hasiewicz, 1989), the single block is described as follows:

$$y_i = (a_i, b_i)\,(x_i, u_i)^T + \xi_i \qquad (i=1,2,\ldots,n),$$

where a_i and b_i are unknown parameters and ξ_i is a random output noise. In a more general case of nonlinear and dynamic system, the single block $F_i()$ can be represented (approximated) by, e.g., two channels of Hammerstein models (see Fig. 7), resembling the Narmax/Lur'e system, considered in this paper.

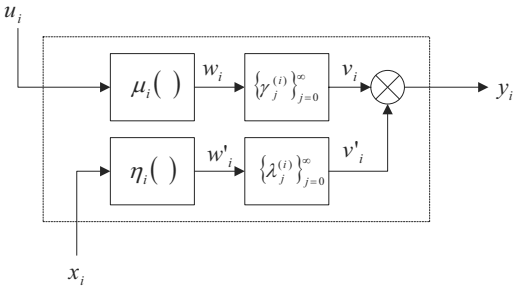

Fig. 7. Model of a single element in a complex system.

Acknowledgment

The author would like to thank the reviewers for their numerous valuable comments and suggestions. This research was supported by the Polish National Science Centre, Grant No. N N514 700640.

References

Bai, E. (1998). An optimal two-stage identification algorithm for Hammerstein–Wiener nonlinear systems, *Automatica* **34**(3): 333–338.

Chen, S. and Billings, S. (1989). Representations of non-linear systems: The NARMAX model, *International Journal of Control* **49**(3): 1013–1032.

Chow, Y. and Teicher, H. (2003). *Probability Theory: Independence, Interchangeability, Martingales*, Springer-Verlag, New York, NY.

Findeisen, W., Bailey, F., Brdyś, M., Malinowski, K., Tatjewski, P. and Woźniak, A. (1980). *Control and Coordination in Hierarchical Systems*, J. Wiley, Chichester/New York, NY.

Finigan, B. and Rowe, I. (1974). Strongly consistent parameter estimation by the introduction of strong instrumental variables, *IEEE Transactions on Automatic Control* **19**(6): 825–830.

Giri, F. and Bai, E.W. (2010). *Block-Oriented Nonlinear System Identification*, Lecture Notes in Control and Information Sciences, Vol. 404, Springer, Berlin.

Greblicki, W. and Pawlak, M. (2008). *Nonparametric System Identification*, Cambridge University Press, New York, NY.

Haber, R. and Keviczky, L. (1999). *Nonlinear System Identification: Input-Output Modeling Approach*, Kluwer Academic Publishers, Dordrecht.

Hannan, E. and Deistler, M. (1988). *The Statistical Theory of Linear Systems*, John Wiley and Sons, New York, NY.

Hansen, L. and Singleton, K. (1982). Generalized instrumental variables estimation of nonlinear rational expectations models, *Econometrica: Journal of the Econometric Society* **50**(5): 1269–1286.

Hasiewicz, Z. (1989). Applicability of least-squares to the parameter estimation of large-scale no-memory linear composite systems, *International Journal of Systems Science* **20**(12): 2427–2449.

Hasiewicz, Z. and Mzyk, G. (2009). Hammerstein system identification by non-parametric instrumental variables, *International Journal of Control* **82**(3): 440–455.

Hill, D. and Chong, C. (1989). Lyapunov functions of Lur'e–Postnikov form for structure preserving models of power systems, *Automatica* **25**(3): 453–460.

Hill, D. and Mareels, I. (1990). Stability theory for differential/algebraic systems with application to power systems, *IEEE Transactions on Circuits and Systems* **37**(11): 1416–1423.

Kincaid, D. and Cheney, E. (2002). *Numerical Analysis: Mathematics of Scientific Computing*, Vol. 2, American Mathematical Society, Pacific Grove, CA.

Kowalczuk, Z. and Kozłowski, J. (2000). Continuous-time approaches to identification of continuous-time systems, *Automatica* **36**(8): 1229–1236.

Kudrewicz, J. (1976). *Functional Analysis for Control and Electronics Engineers*, PWN, Warsaw, (in Polish).

Lu, J. and Hill, D. (2007). Impulsive synchronization of chaotic Lur'e systems by linear static measurement feedback: An LMI approach, *IEEE Transactions on Circuits and Systems II: Express Briefs* **54**(8): 710–714.

Mzyk, G. (2007). Generalized kernel regression estimate for the identification of Hammerstein systems, *International Journal of Applied Mathematics and Computer Science* **17**(2): 189–197, DOI: 10.2478/v10006-007-0018-z.

Mzyk, G. (2009). Nonlinearity recovering in Hammerstein system from short measurement sequence, *IEEE Signal Processing Letters* **16**(9): 762–765.

Mzyk, G. (2013). Instrumental variables for nonlinearity recovering in block-oriented systems driven by correlated signal, *International Journal of Systems Science*, DOI: 10.1080/00207721.2013.775682.

Rao, C. (1973). *Linear Statistical Inference and Its Applications*, Wiley, New York, NY.

Sagara, S. and Zhao, Z.-Y. (1990). Numerical integration approach to on-line identification of continuous-time systems, *Automatica* **26**(1): 63–74.

Sastry, S. (1999). *Nonlinear Systems: Analysis, Stability, and Control*, Interdisciplinary Applied Mathematics, Vol. 10, Springer, New York, NY.

Söderström, T. and Stoica, P. (1983). *Instrumental Variable Methods for System Identification*, Vol. 161, Springer-Verlag, Berlin.

Söderström, T. and Stoica, P. (1989). *System Identification*, Prentice Hall, Englewood Cliffs, NJ.

Söderström, T. and Stoica, P. (2002). Instrumental variable methods for system identification, *Circuits, Systems, and Signal Processing* **21**(1): 1–9.

Stoica, P. and Söderström, T. (1982). Instrumental-variable methods for identification of Hammerstein systems, *International Journal of Control* **35**(3): 459–476.

Suykens, J., Yang, T. and Chua, L. (1998). Impulsive synchronization of chaotic Lur'e systems by measurement feedback, *International Journal of Bifurcation and Chaos* **8**(06): 1371–1381.

Ward, R. (1977). Notes on the instrumental variable method, *IEEE Transactions on Automatic Control* **22**(3): 482–484.

Wong, K. and Polak, E. (1967). Identification of linear discrete time systems using the instrumental variable method, *IEEE Transactions on Automatic Control* **12**(6): 707–718.

Zhang, Y., Bai, E., Libra, R., Rowden, R. and Liu, H. (1996). Simulation of spring discharge from a limestone aquifer in Iowa, USA, *Hydrogeology Journal* **4**(4): 41–54.

Zhao, Z.-Y., Sagara, S. and Wada, K. (1991). Bias-compensating least squares method for identification of continuous-time systems from sampled data, *International Journal of Control* **53**(2): 445–461.

Grzegorz Mzyk was born in 1973 in Poland. He received his M.Sc. and Ph.D. degrees from the Wrocław University of Technology, Wrocław, Poland, in 1998, and 2002, respectively. He is now an assistant professor at that university and teaches courses in control theory and system identification. His current research interests are in mixed parametric non-parametric identification of Hammerstein and Wiener systems.

Appendix A
Proofs of theorems and lemmas

A1. Hammerstein system as a special case of a NARMAX system

Lemma A1. *The additive NARMAX system with the linear function $\eta(y_k)$, i.e., of the form $\eta(y_k) = dy_k$, is equivalent to the Hammerstein system.*

Proof. The NARMAX system description

$$y_k = \sum_{j=1}^{p} a_j \eta(y_{k-j}) + \sum_{i=0}^{n} b_i \mu(u_{k-i}) + v_k,$$

for $\eta(y_k) = dy_k$ and the 'input'

$$x_k \triangleq \sum_{i=0}^{n} b_i \mu(u_{k-i}) + v_k, \tag{A1}$$

resembles the difference equation of the AR linear model

$$y_k = \sum_{j=1}^{p} a_j dy_{k-j} + x_k,$$

which can be presented equivalently as (Hannan and Deistler, 1988)

$$y_k = \sum_{l=0}^{\infty} r_l x_{k-l}. \tag{A2}$$

Inserting (A1) to (A2) leads to

$$y_k = \sum_{l=0}^{\infty} r_l \left(\sum_{i=0}^{n} b_i \mu(u_{k-i-l}) + v_{k-l} \right),$$

and further

$$y_k = \sum_{q=0}^{\infty} \gamma_q \mu(u_{k-q}) + z_k, \tag{A3}$$

where $z_k = \sum_{l=0}^{\infty} r_l v_{k-l}$, $\gamma_q = \sum_{l=0}^{\infty} \sum_{i=0}^{n} r_l b_i \delta(l+i-q)$, and $\delta()$ is a discrete impulse. Equation (A3) represents a Hammerstein system with an infinite impulse response. ■

A2. Necessary condition for the 3-stage algorithm

Lemma A2. *If* $\det(B^T A) \neq 0$ *for given matrices* $A, B \in \mathbb{R}^{\alpha \times \beta}$ *with finite elements, then* $\det(A^T A) \neq 0$.

Proof. Let $\det(A^T A) = 0$, i.e., $\operatorname{rank}(A^T A) < \beta$. From the obvious property that

$$\operatorname{rank}(A^T A) = \operatorname{rank}(A)$$

one can conclude that there exists the non-zero vector $\xi \in \mathbb{R}^{\beta}$, such that $A\xi = 0$. Premultiplying this equation by B^T we get $B^T A \xi = 0$, and hence $\det(B^T A) = 0$. Thus, for $A = \frac{1}{\sqrt{N}} \Phi_N$ and $B = \frac{1}{\sqrt{N}} \Psi_N$, a necessary condition for $\frac{1}{N} \Psi_N^T \Phi_N$ to be of full rank is $\det(\frac{1}{N} \Phi_N^T \Phi_N) \neq 0$, i.e., a persistent excitation of $\{\phi_k\}$. ■

A3. Proof of Theorem 1

Proof. From the Slutzky theorem (cf. the work of Chow and Teicher (2003) and Appendix B) we have

$$\operatorname*{Plim}_{N\to\infty} (\Delta_N^{(IV)})$$

$$= \left(\operatorname*{Plim}_{N\to\infty} \left(\frac{1}{N} \Psi_N^T \Phi_N \right) \right)^{-1} \operatorname*{Plim}_{N\to\infty} \left(\frac{1}{N} \Psi_N^T Z_N \right),$$

and directly from the conditions (C2) and (C3), we get,

$$\operatorname*{Plim}_{N\to\infty} \left(\Delta_N^{(IV)} \right) = 0. \tag{A4}$$

■

A4. Proof of Theorem 2

Proof. Let us define the scalar random variable

$$\xi_N = \|\Delta_N^{(IV)}\| = \|\widehat{\theta}_N^{(IV)} - \theta\|,$$

where $\|\cdot\|$ denotes any vector norm. It must be shown that

$$P\left\{ r_N \frac{\xi_N}{a_N} > \varepsilon \right\} \to 0 \quad \text{as } N \to \infty,$$

for each $\varepsilon > 0$, each $r_N \to 0$ and $a_N = 1/\sqrt{N}$. To prove that $\xi_N = O(1/\sqrt{N})$ in probability, it suffices to show that $\xi_N = O(1/N)$ in the mean square sense. Introducing

$$A_N = \frac{1}{N} \Psi_N^T \Phi_N = \frac{1}{N} \sum_{k=1}^{N} \psi_k \phi_k^T,$$

$$B_N = \frac{1}{N} \Psi_N^T Z_N = \frac{1}{N} \sum_{k=1}^{N} \psi_k z_k,$$

we obtain that

$$\Delta_N^{(IV)} = A_N^{-1} B_N. \tag{A5}$$

Therefore, under Assumptions 1–6, the system output y_k is bounded, i.e., $|y_k| < y_{\max} < \infty$. Moreover, under the condition (C1), we have

$$\left| A_N^{i,j} \right| \leq \psi_{\max} p_{\max} < \infty,$$

for $j = 1, 2, \ldots, m(n+1)$, and

$$\left| A_N^{i,j} \right| \leq \psi_{\max} p_{\max} < \infty,$$

for $j = m(n+1)+1, \ldots, m(n+1)+pq$, so each element of A_N is bounded.

Similarly, one can show the boundedness of the elements of the vector B_N. The norm of the error error $\Delta_N^{(IV)}$ given by (A5) can be evaluated as follows:

$$\begin{aligned} \xi_N = \|\Delta_N^{(IV)}\| &= \| \left(\frac{1}{N} \Psi_N^T \Phi_N \right)^{-1} \left(\frac{1}{N} \Psi_N^T Z_N \right) \| \\ &\leq \| \left(\frac{1}{N} \Psi_N^T \Phi_N \right)^{-1} \| \| \frac{1}{N} \Psi_N^T Z_N \| \\ &\leq c \| \frac{1}{N} \Psi_N^T Z_N \| = c \| \frac{1}{N} \sum_{k=1}^{N} \psi_k z_k \|, \end{aligned}$$

where c is some positive constant. Obviously, one can find $\alpha \geq 0$ such that

$$c \| \frac{1}{N} \sum_{k=1}^{N} \psi_k z_k \| \leq \alpha c \sum_{i=1}^{\dim \psi_k} \left(\frac{1}{N} | \sum_{k=1}^{N} \psi_{k,i} z_k | \right),$$

and hence

$$\begin{aligned}\xi_N^2 &= \|\Delta_N^{(IV)}\|^2 \\ &\le \alpha^2 c^2 \left[\sum_{i=1}^{\dim\psi_k}\left(\frac{1}{N}|\sum_{k=1}^{N}\psi_{k,i}z_k|\right)\right]^2 \\ &\le \alpha^2 c^2 \dim\psi_k \sum_{i=1}^{\dim\psi_k}\left(\frac{1}{N}|\sum_{k=1}^{N}\psi_{k,i}z_k|\right)^2 \\ &= \alpha^2 c^2 \dim\psi_k \sum_{i=1}^{\dim\psi_k}\frac{1}{N^2}\left(\sum_{k=1}^{N}\psi_{k,i}z_k\right)^2.\end{aligned}$$

Moreover, for uncorrelated processes$\{\psi_k\}$ and $\{z_k\}$ (see the condition (C3)) we have that

$$\begin{aligned}&E\xi_N^2 \\ &\le \alpha^2 c^2 \dim\psi_k \sum_{i=1}^{\dim\psi_k}\frac{1}{N^2}E\left(\sum_{k=1}^{N}\psi_{k,i}z_k\right)^2 \\ &= \alpha^2 c^2 \dim\psi_k \sum_{i=1}^{\dim\psi_k}\frac{1}{N^2}E\left[\sum_{k_1=1}^{N}\sum_{k_2=1}^{N}\psi_{k_1,i}\psi_{k_2,i}z_{k_1}z_{k_2}\right] \\ &\le \alpha^2 c^2 \dim\psi_k \sum_{i=1}^{\dim\psi_k}\frac{1}{N^2} \\ &\quad\times\sum_{k_1=1}^{N}\sum_{k_2=1}^{N}\left|E\left[\psi_{k_1,i}\psi_{k_2,i}\right]\right|\left|E\left[z_{k_1}z_{k_2}\right]\right| \\ &\le \alpha^2 c^2 \left(\dim\psi_k\right)^2\frac{\psi_{\max}^2}{N}\left[|r_z(0)|\right. \\ &\quad\left.+2\sum_{\tau=1}^{N}\left(1-\frac{\tau}{N}\right)|r_z(\tau)|\right] \\ &\le \frac{C}{N}\sum_{\tau=0}^{\infty}|r_z(\tau)|,\end{aligned}$$

where

$$\begin{aligned}r_z(\tau) &= \operatorname{var}\varepsilon\sum_{i=0}^{\infty}\omega_i\omega_{i+\tau}, \\ C &= 2\alpha^2c^2\left(\dim\psi_k\right)^2\psi_{\max}^2.\end{aligned}$$

Since

$$\begin{aligned}&\left|\operatorname{var}\varepsilon\sum_{\tau=0}^{\infty}\sum_{i=0}^{\infty}\omega_i\omega_{i+\tau}\right| \\ &\quad\le \operatorname{var}\varepsilon\sum_{\tau=0}^{\infty}\sum_{i=0}^{\infty}|\omega_i|\,|\omega_{i+\tau}| \\ &\quad\le \operatorname{var}\varepsilon\sum_{i=0}^{\infty}|\omega_i|\sum_{i=0}^{\infty}|\omega_{i+\tau}| < \infty,\end{aligned}$$

we have

$$E\xi_N^2 \le D\frac{1}{N},$$

where

$$D = C\operatorname{var}\varepsilon\left|\sum_{\tau=0}^{\infty}\sum_{i=0}^{\infty}\omega_i\omega_{i+\tau}\right|.$$

■

A5. Proof of Theorem 4

Proof. To simplify the presentation, let $z_{\max}=1$. From (53) we get

$$\begin{aligned}\|\Delta_N^{(IV)}(\Psi_N)\|^2 &= \Delta_N^{(IV)T}(\Psi_N)\Delta_N^{(IV)}(\Psi_N) \\ &= Z_N^{*T}\Gamma_N^T\Gamma_N Z_N^*,\end{aligned}$$

and the maximum value of the cumulated error is

$$\begin{aligned}Q(\Psi_N) &= \max_{\|\mathbf{z}_N^*\|\le 1}\left(\Delta_N^{(IV)T}(\Psi_N)\Delta_N^{(IV)}(\Psi_N)\right) \\ &= \max_{\|Z_N^*\|\le 1}\left\langle Z_N^*,\Gamma_N^T\Gamma_N Z_N^*\right\rangle \\ &= \|\Gamma_N\|^2 = \lambda_{\max}\left(\Gamma_N^T\Gamma_N\right),\end{aligned}$$

where $\|\cdot\|$ is the spectral matrix norm induced by the Euclidean vector norm, and $\lambda_{\max}(\cdot)$ denotes the largest eigenvalue of a matrix. Since (see Wong and Polak, 1967; Rao, 1973)

$$\lambda_{\max}\left(\Gamma_N^T\Gamma_N\right) = \lambda_{\max}\left(\Gamma_N\Gamma_N^T\right),$$

from the definition of Γ_N we obtain that

$$\begin{aligned}&\max_{\|Z_N^*\|\le 1}\left(\Delta_N^{(IV)T}(\Psi_N)\Delta_N^{(IV)}(\Psi_N)\right) \\ &\quad= \max_{\|\zeta\|\le 1}\left\langle\zeta,\Gamma_N\Gamma_N^T\zeta\right\rangle \\ &\quad= \max_{\|\zeta\|\le 1}\left\langle\zeta,\left(\frac{1}{N}\Psi_N^T\Phi_N\right)^{-1}\right. \\ &\qquad\left.\times\frac{1}{N}\Psi_N^T\Psi_N\left(\frac{1}{N}\Phi_N^T\Psi_N\right)^{-1}\zeta\right\rangle.\end{aligned}$$

On the basis of (52), we get

$$\begin{aligned}&\max_{\|Z_N^*\|\le 1}\left(\Delta_N^{(IV)T}(\Psi_N)\Delta_N^{(IV)}(\Psi_N)\right) \\ &\quad= \max_{\|\zeta\|\le 1}\left\langle\zeta,\left(\frac{1}{N}\Psi_N^T\Phi_N^{\#}\right)^{-1}\frac{1}{N}\Psi_N^T\Psi_N\right. \\ &\qquad\left.\left(\frac{1}{N}\Phi_N^{\#T}\Psi_N\right)^{-1}\zeta\right\rangle,\end{aligned}$$

with probability 1, as $N\to\infty$, where Φ_N and $\Phi_N^{\#}$ are given by (12) and (51), respectively. Using Lemma B1 for $M_1=\frac{1}{\sqrt{N}}\Phi_N^{\#}$ and $M_2=\frac{1}{\sqrt{N}}\Psi_N$, we get

$$\zeta^T\Gamma_N\Gamma_N^T\zeta \ge \zeta^T\left(\frac{1}{N}\Phi_N^{\#T}\Phi_N^{\#}\right)^{-1}\zeta,$$

for each vector ζ, and consequently

$$Q(\Psi_N) = \max_{\|\zeta\|\le 1} \left(\zeta^T \Gamma_N \Gamma_N^T \zeta\right) \geq \max_{\|\zeta\|\le 1} \left(\zeta^T \left(\frac{1}{N}\Phi_N^{\#T}\Phi_N^{\#}\right)^{-1}\zeta\right).$$

For $\Psi_N = \Phi_N^{\#}$, we have

$$\max_{\|\zeta\|\le 1} \left(\zeta^T \Gamma_N \Gamma_N^T \zeta\right) = \max_{\|\zeta\|\le 1} \left(\zeta^T \left(\frac{1}{N}\Phi_N^{\#T}\Phi_N^{\#}\right)^{-1}\zeta\right),$$

and the criterion $Q(\Psi_N)$ attains a minimum. The choice $\Psi_N = \Phi_N^{\#}$ is thus asymptotically optimal. ■

A6. Proof of Theorem 7

Proof. The estimation error (67) can be decomposed as follows

$$\Delta_{N,M}^{(IV)} = \widehat{\theta}_{N,M}^{*(IV)} - \theta = \widehat{\theta}_{N,M}^{*(IV)} - \widehat{\theta}_N^{*(IV)} + \widehat{\theta}_N^{*(IV)} - \theta,$$

where $\widehat{\theta}_N^{*(IV)} = \left(\Psi_N^{*T}\Phi_N\right)^{-1}\Psi_N^{*T}Y_N$, and Ψ_N^* is defined by (55) and (51). From the triangle inequality, for each norm $\|\cdot\|$ we have

$$\|\Delta_{N,M}^{(IV)}\| \le \|\widehat{\theta}_{N,M}^{*(IV)} - \widehat{\theta}_N^{*(IV)}\| + \|\widehat{\theta}_N^{*(IV)} - \theta\|. \quad \text{(A6)}$$

On the basis of Theorem 1,

$$\|\widehat{\theta}_N^{*(IV)} - \theta\| \to 0 \text{ in probability,}$$

as $N \to \infty$. To prove 7, let us analyze the component $\|\widehat{\theta}_{N,M}^{*(IV)} - \widehat{\theta}_N^{*(IV)}\|$ in (A6) to show that, for fixed N, it tends to zero in probability as $M \to \infty$.

Write

$$\varepsilon_N \triangleq \frac{1}{\left\|\frac{1}{N}\Psi_N^{*T}\Phi_N\right\|} \qquad (N \text{ – fixed}).$$

From (65) we have that

$$\left\|\left(\frac{1}{N}\widehat{\Psi}_{N,M}^{*T}\Phi_N\right) - \left(\frac{1}{N}\Psi_N^{*T}\Phi_N\right)\right\| \to 0$$

in probability as $M \to \infty$, and particularly

$$\lim_{M\to\infty} P\left\{\left\|\frac{1}{N}\widehat{\Psi}_{N,M}^{*T}\Phi_N - \frac{1}{N}\Psi_N^{*T}\Phi_N\right\| < \varepsilon_N\right\} = 1.$$

Introducing

$$r_M \triangleq \frac{\left\|\left(\frac{1}{N}\widehat{\Psi}_{N,M}^{*T}\Phi_N\right) - \left(\frac{1}{N}\Psi_N^{*T}\Phi_N\right)\right\|}{\varepsilon_N\left(\varepsilon_N - \left\|\left(\frac{1}{N}\widehat{\Psi}_{N,M}^{*T}\Phi_N\right) - \left(\frac{1}{N}\Psi_N^{*T}\Phi_N\right)\right\|\right)}$$

and using the Banach theorem (see Kudrewicz, 1976, Theorem 5.8.), we get

$$\lim_{M\to\infty} P\left\{\left\|\left(\frac{\widehat{\Psi}_{N,M}^{*T}\Phi_N}{N}\right)^{-1} - \left(\frac{\Psi_N^{*T}\Phi_N}{N}\right)^{-1}\right\| \le r_M\right\} = 1.$$

Since $r_M \to 0$ in probability as $M \to \infty$, there holds

$$\left\|\widehat{\theta}_{N,M}^{*(IV)} - \widehat{\theta}_N^{*(IV)}\right\| \to 0 \quad \text{in probability,}$$

as $M \to \infty$, for each N. ■

Appendix B
Technical lemmas, theorems and definitions

B1. SVD decomposition

Theorem B1. (Kincaid and Cheney, 2002) *For each $A \in \mathbb{R}^{m,n}$ there exist the unitary matrices $U \in \mathbb{R}^{m,m}$ and $V \in \mathbb{R}^{n,n}$, such that*

$$U^T A V = \Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_l), \qquad \text{(B1)}$$

where $l = \min(m,n)$, and

$$\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_r > 0,$$
$$\sigma_{r+1} = \cdots = \sigma_l = 0,$$

where $r = \operatorname{rank}(A)$.

The numbers $\sigma_1, \ldots, \sigma_l$ are called the singular values of the matrix A. Solving (B1) with respect to A, we obtain

$$A = U\Sigma V^T = \sum_{i=1}^{r} u_i \sigma_i v_i^T = \sum_{i=1}^{r} \sigma_i u_i v_i^T, \qquad \text{(B2)}$$

where u_i and v_i denote the i-th columns of U and V, respectively.

B2. Factorization theorem

Theorem B2. (Rao, 1973) *Each positive definite matrix M can be shown in the form $M = PP^T$, where P (a root of M) is nonsingular.*

B3. Technical lemma

Lemma B1. (Wong and Polak, 1967) *Let M_1 and M_2 be two matrices with the same dimensions. If $\left(M_1^T M_1\right)^{-1}$, $\left(M_1^T M_2\right)^{-1}$ and $\left(M_2^T M_1\right)^{-1}$ exist, then*

$$D_N = \left(M_2^T M_1\right)^{-1} M_2^T M_2 \left(M_1^T M_2\right)^{-1} - \left(M_1^T M_1\right)^{-1}$$

is nonnegative definite, i.e., for each ζ

$$\zeta^T D_N \zeta \ge 0.$$

B4. Slutzky theorem

Theorem B3. (Roe, 1973) *If* $Plim_{k\to\infty}\varkappa_k = \varkappa^{\#}$ *and the function* $g(\cdot)$ *is continuous, then* $Plim_{k\to\infty}g(\varkappa_k) = g(\varkappa^{\#})$.

B5. Chebyshev's inequality

Lemma B2. (Chow and Teicher, 2003, p. 106) *For each constant c, each random variable* X *and each* $\varepsilon > 0$*, there holds* $P\{|X-c| > \varepsilon\} \le \frac{1}{\varepsilon^2}E(X-c)^2$. *In particular, for* $c = EX$, $P\{|X-EX| > \varepsilon\} \le \frac{1}{\varepsilon^2}\mathrm{var}\,X$.

B6. Persistent excitation

Definition B1. A stationary random process $\{\alpha_k\}$ is strongly persistently exciting of orders $n \times m$ (denote $SPE(n,m)$) if the matrix

$$R_{\varkappa}(n,m) = E\left[\begin{array}{c}\varkappa_k\\ :\\ \varkappa_{k-n+1}\end{array}\right]\left[\begin{array}{c}\varkappa_k\\ :\\ \varkappa_{k-n+1}\end{array}\right]^T,$$

where $\varkappa_k = \left[\begin{array}{cccc}\alpha_k & \alpha_k^2 & \dots & \alpha_k^m\end{array}\right]^T$, is of full rank.

Lemma B3. (Stoica and Söderström, 1982) *The i.i.d. process* $\{\alpha_k\}$ *is* $SPE(n,m)$ *for each* n *and* m.

Lemma B4. (Stoica and Söderström, 1982) *Let* $x_k = H(q^{-1})u_k$, $H(q^{-1})$ *be an asymptotically stable linear filter, and* $\{u_k\}$ *be a random sequence with finite variance. If the frequency function of* $\{u_k\}$ *is strictly positive in at least* $m+1$ *distinct points, then* $\{x_k\}$ *is* $SPE(n,m)$ *for each* n.

B7. Modified triangle inequality

Lemma B5. (Chow and Teicher, 2003) *If* X *and* Y *are* k*-dimensional random vectors, then* $P[\|X+Y\| \ge \varepsilon] \le P[\|X\| \ge \varepsilon/2] + P[\|Y\| \ge \varepsilon/2]$ *for each vector norm* $\|\cdot\|$ *and each* $\varepsilon > 0$.

Received: 6 December 2012
Revised: 29 April 2013