Nonparametric Instrumental Variables for Narmax System Identification

Grzegorz Mzyk

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

***Keywords*—System identification, instrumental variables, NARMAX system, nonparametric methods.**

I. Statement of the problem

In the paper we consider a scalar, discrete-time, asymptotically stable nonlinear dynamic system shown in Fig. 1 and described by the following equation (cf. [2], [13], [15], [14], [1]):

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

where

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

The structure is well known in the literature as the additive NARMAX model ([2]). The signals y_k, u_k are z_k are the output, the input and the noise, respectively. We take the following assumptions.

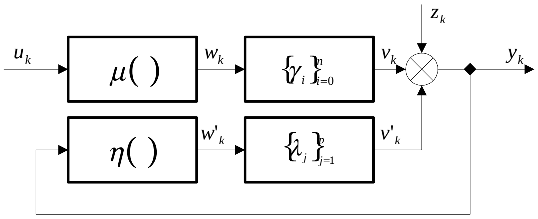

Figure 1. The additive NARMAX system

Assumption 1: The static nonlinear characteristics are of given parametric form

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

where $f_1()$,...,$f_m()$ and $g_1(),...,g_q()$ are a priori known linearly independent functions, such that

$$\begin{aligned} |f_t(u)| &\leqslant p_{\max}, \qquad (4) \\ |g_l(y)| &\leqslant p_{\max}, \end{aligned}$$

some constant $p_{\max}$.

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

$$\begin{aligned} v_k &= \sum_{i=0}^{n} \gamma_i w_{k-i} \qquad (5) \\ v'_k &= \sum_{j=1}^{p} \lambda_j w'_{k-j} \end{aligned}$$

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. it exists (unknown) $u_{\max}$, such that $|u_k| < u_{\max} < \infty$.

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

$$z_k = \sum_{i=0}^{\infty} \omega_i \varepsilon_{k-i}, \tag{6}$$

where $\{\varepsilon_k\}$ is some unknown zero-mean ($\mathbf{E}\varepsilon_k = 0$) 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 the unknown stable linear filter.

Assumption 5: The system as a whole is asymptotically stable.

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

Let

$$\begin{aligned} \mathbf{\Lambda} &= (\lambda_1, .., \lambda_p)^T \qquad (7) \\ \mathbf{\Gamma} &= (\gamma_0, ..., \gamma_n)^T \\ \mathbf{c} &= (c_1, ..., c_m)^T \\ \mathbf{d} &= (d_1, ..., d_q)^T \end{aligned}$$

denote true (unknown) parameters of the system. Let us notice that the input-output description of the system, given by (1)-(2) is not unique. For each pair of constants $\overline{\alpha}$ and $\overline{\beta}$, the systems with parameters $\mathbf{\Lambda}$, Γ, $\mathbf{c}$, $\mathbf{d}$ and $\overline{\beta}\mathbf{\Lambda}$, $\overline{\alpha}\Gamma$, $\mathbf{c}/\overline{\alpha}$, $\mathbf{d}/\overline{\beta}$ cannot be distinguished, i.e., are equivalent (see (1)-(2)). For the uniqueness of the solution we introduce the following technical assumption (see [1]):

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

(b) $||\mathbf{\Lambda}||_2 = 1$ and $||\Gamma||_2 = 1$, where $||.||_2$ is Euclidean vector norm;

(c) first non-zero elements of $\mathbf{\Lambda}$ and Γ are positive.

Let

$$\begin{aligned} \theta &= (\gamma_0 c_1, ..., \gamma_0 c_m, ..., \gamma_n c_1, ..., \gamma_n c_m, \qquad (8) \\ &\quad \lambda_1 d_1, ..., \lambda_1 d_q, ..., \lambda_p d_1, ..., \lambda_p d_q)^T \\ &= (\theta_1, ..., \theta_{(n+1)m}, \theta_{(n+1)m+1}, ..., \theta_{(n+1)m+pq})^T \end{aligned}$$

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

$$\begin{aligned} \phi_k &= (f_1(u_k), ..., f_1(u_{k-n}), ..., f_1(u_{k-1}), ..., f_m(u_{k-n}), \qquad (9) \\ &\quad g_1(y_{k-1}), ..., g_q(y_{k-1}), ..., g_1(y_{k-p}), ..., g_q(y_{k-p}))^T. \end{aligned}$$

Thanks to above notation, 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 parameters. For $k = 1, ..., N$ we obtain

$$\mathbf{Y}_N = \mathbf{\Phi}_N \theta + \mathbf{Z}_N \tag{10}$$

The author works for the Institute of Computer Engineering, Control and Robotics, Wrocław University of Technology, Janiszewskiego 11/17, 50-372 Wrocław, Poland, tel. +48 71 320 25 49, fax. +48 71 321 26 77, e-mail: Grzegorz.Mzyk@pwr.wroc.pl

where $\mathbf{Y}_N = (y_1, ..., y_N)^T$, $\mathbf{\Phi}_N = (\phi_1, ..., \phi_N)^T$, and $\mathbf{Z}_N = (z_1, ..., z_N)^T$.

The system in Fig. 1 is more general then often met in the literature classic Hammerstein system. The Hammerstein system is its special case, when the function $\eta()$ is linear (see Appendix VIII-A). The additive NARMAX system is also not equivalent to widely considered in the literature Wiener-Hammerstein (sandwich) system, where two linear dynamic blocks surround one static nonlinearity. In spite of many possibilities of applications in various domains ([7], [1], [22]), relatively small attention was paid to this structure in the literature.

The purpose of identification is to recover parameters in $\mathbf{\Lambda}$, $\mathbf{\Gamma}$, $\mathbf{c}$ and $\mathbf{d}$ (given by (7)), using the input-output measurements (u_k, y_k) $(k = 1, ..., N)$ of the whole system.

In the next Section the least squares based identification algorithm (see [1]) will be presented for white disturbances. Then, the reason of its asymptotic bias will be shown for correlated noise. Next, the new, asymptotically unbiased, instrumental variables based estimate is proposed. The idea originates from the linear system theory (see e.g. [18] and [21]), where the instrumental variables technique is used for identification of simple one-element linear dynamic objects. The proposed method is then compared with the least squares. In particular, the consistency of the proposed estimate was shown even for correlated disturbances. The form of the optimal instrumental variables was established and the method of their appropriate generation was described. Also, the rate of convergence is analyzed and the results of experiments are included.

II. Least squares and SVD approach

For comparison studies with the proposed further instrumental variables method we start from presentation of the two-stage algorithm based on least squares estimation of the aggregated parameter vector and decomposition of the obtained result with the use of SVD algorithm (see [1], [10], [11]).

Stage 1. Compute the *LS* estimate

$$\widehat{\theta}_N^{(LS)} = (\mathbf{\Phi}_N^T \mathbf{\Phi}_N)^{-1} \mathbf{\Phi}_N^T \mathbf{Y}_N \tag{11}$$

of the aggregated parameter vector θ (see (8) and (10)), and next construct (by the *plug-in* method) evaluations $\widehat{\mathbf{\Theta}}_{\Lambda d}^{(LS)}$ and $\widehat{\mathbf{\Theta}}_{\Gamma c}^{(LS)}$ of the matrices $\mathbf{\Theta}_{\Lambda d} = \mathbf{\Lambda} \mathbf{d}^T$ and $\mathbf{\Theta}_{\Gamma c} = \mathbf{\Gamma} \mathbf{c}^T$, respectively (see condition (a) above).

Stage 2. Perform the SVD (singular value decomposition – see Appendix IX-A) of the matrices $\widehat{\mathbf{\Theta}}_{\Lambda d}^{(LS)}$ and $\widehat{\mathbf{\Theta}}_{\Gamma c}^{(LS)}$:

$$\begin{aligned} \widehat{\mathbf{\Theta}}_{\Lambda d}^{(LS)} &= \sum_{i=1}^{\min(p,q)} \delta_i \widehat{\xi}_i \widehat{\zeta}_i^T \qquad (12) \\ \widehat{\mathbf{\Theta}}_{\Gamma c}^{(LS)} &= \sum_{i=1}^{\min(n,m)} \sigma_i \widehat{\mu}_i \widehat{\nu}_i^T \end{aligned}$$

and next compute the estimates of parameters of particular blocks (see (7))

$$\begin{aligned} \widehat{\mathbf{\Lambda}}_N^{(LS)} &= sgn(\widehat{\xi}_1[\kappa_{\xi_1}]) \widehat{\xi}_1 \qquad (13) \\ \widehat{\mathbf{\Gamma}}_N^{(LS)} &= sgn(\widehat{\mu}_1[\kappa_{\mu_1}]) \widehat{\mu}_1 \\ \widehat{\mathbf{c}}_N^{(LS)} &= sgn(\widehat{\mu}_1[\kappa_{\mu_1}]) \sigma_1 \widehat{\nu}_1 \\ \widehat{\mathbf{d}}_N^{(LS)} &= sgn(\widehat{\xi}_1[\kappa_{\xi_1}]) \delta_1 \widehat{\zeta}_1 \end{aligned}$$

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

Let us analyze the form of *SVD* representations of the theoretical matrices $\mathbf{\Theta}_{\Gamma c} = \mathbf{\Gamma} \mathbf{c}^T$ and $\mathbf{\Theta}_{\Lambda d} = \mathbf{\Lambda} \mathbf{d}^T$. Each matrix being the product of two vectors has the rank equal 1, and only one singular value is not zero, i.e.,

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

and

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

thus

$$\mathbf{\Theta}_{\Gamma c} = \sigma_1 \mu_1 \nu_1^T, \tag{14}$$

where $\|\mu_1\|_2 = \|\nu_1\|_2 = 1$. Representation of $\mathbf{\Theta}_{\Gamma c}$ given by (14) is obviously unique [10]. To obtain Γ, which fulfills the condition (b) one can take $\Gamma = \mu_1$, or $\Gamma = -\mu_1$. The condition (c) guarantees uniqueness of Γ. The remaining part of decomposition allows for computing $\mathbf{c}$. The vectors $\mathbf{\Lambda}$ and $\mathbf{d}$ can be obtained from $\mathbf{\Theta}_{\Lambda d}$ in a similar way.

The Singular Value Decomposition allows for splitting of aggregated matrices of parameters $\widehat{\mathbf{\Theta}}_{\Gamma c}^{(LS)}$ and $\widehat{\mathbf{\Theta}}_{\Lambda d}^{(LS)}$ into products of two vectors (see (12)) and estimating $\widehat{\mathbf{\Gamma}}_N^{(LS)} \widehat{\mathbf{c}}_N^{(LS)T}$ and $\widehat{\mathbf{\Lambda}}_N^{(LS)} \widehat{\mathbf{d}}_N^{(LS)T}$ according to (13). It was shown in [1] that

$$(\widehat{\mu}_1, \sigma_1 \widehat{\nu}_1) = \arg \min_{c \in R^m, \Gamma \in R^n} \left\| \widehat{\mathbf{\Theta}}_{\Gamma c}^{(LS)} - \mathbf{\Gamma} \mathbf{c}^T \right\|^2. \tag{15}$$

and for the noise-free case ($z_k \equiv 0$) the estimates (13) equal to true system parameters, i.e.,

$$\begin{aligned} \widehat{\mathbf{\Lambda}}_N^{(LS)} &= \mathbf{\Lambda}, \qquad (16) \\ \widehat{\mathbf{\Gamma}}_N^{(LS)} &= \mathbf{\Gamma}, \\ \widehat{\mathbf{c}}_N^{(LS)} &= \mathbf{c}, \\ \widehat{\mathbf{d}}_N^{(LS)} &= \mathbf{d}. \end{aligned}$$

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

$$\begin{aligned} \widehat{\mathbf{\Lambda}}_N^{(LS)} &\to \mathbf{\Lambda}, \qquad (17) \\ \widehat{\mathbf{\Gamma}}_N^{(LS)} &\to \mathbf{\Gamma}, \\ \widehat{\mathbf{c}}_N^{(LS)} &\to \mathbf{c}, \\ \widehat{\mathbf{d}}_N^{(LS)} &\to \mathbf{d}, \end{aligned}$$

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

By taking (10) and (11) into account, the estimation error of the vector θ by the least squares can be expressed as follows

$$\begin{aligned} \Delta_N^{(LS)} &= \widehat{\theta}_N^{(LS)} - \theta = \qquad (18) \\ & \left(\mathbf{\Phi}_N^T \mathbf{\Phi}_N \right)^{-1} \mathbf{\Phi}_N^T \mathbf{Z}_N \\ &= \left(\frac{1}{N} \sum_{k=1}^{N} \phi_k \phi_k^T \right)^{-1} \left(\frac{1}{N} \sum_{k=1}^{N} \phi_k z_k \right). \end{aligned}$$

If $\{z_k\}$ is a zero-mean white noise with finite variance, independent of $\{u_k\}$, then all elements of the vector $\mathbf{Z}_N$ are independent of the elements of the matrix $\mathbf{\Phi}_N$ and from ergodicity of the noise and the process $\{\phi_k\}$ (see Appendix IX-H) it holds that $\Delta_N^{(LS)} \to 0$ with probability 1, as $N \to \infty$. Nevertheless, if $\{z_k\}$ is correlated, i.e., $\mathbf{E} z_k z_{k+i} \neq 0$ for some $i \neq 0$, then the *LS* estimate (11) of θ is not consistent because of the dependence between z_k, a the values $g_l(y_{k-i})$ $(l = 1, ..., q,\ i = 1, ..., p)$ included in ϕ_k. Consequently, the estimates given by (13) are not consistent, too. This conclusion is illustrated in simulation example.

III. Instrumental variables approach

Let us assume that besides $\mathbf{\Phi}_N$ (see (10)) we have given, or we can generate, additional matrix $\mathbf{\Psi}_N$ of instrumental variables, which fulfills (even for correlated z_k) the following conditions:

(C1): $dim\mathbf{\Psi}_N = \dim\mathbf{\Phi}_N$, and the elements of $\mathbf{\Psi}_N = (\psi_1, \psi_2, ..., \psi_N)^T$, where $\psi_k = (\psi_{k,1}, \psi_{k,2}, ..., \psi_{k,m(n+1)+pq})^T$, are commonly bounded, i.e., there exists $0 < \psi_{\max} < \infty$ such that $\left|\psi_{k,j}\right| \leq \psi_{\max}$ $(k = 1...N,\ j = 1...m(n+1) + pq)$ and $\psi_{k,j}$ are ergodic, not necessary zero-mean, processes (see Appendix IX-H)

(C2): there exists $Plim(\frac{1}{N}\mathbf{\Psi}_N^T\mathbf{\Phi}_N) = \mathbf{E}\psi_k\phi_k^T$ and the limit is not singular, i.e., $\det\{\mathbf{E}\psi_k\phi_k^T\} \neq 0$

(C3): $Plim(\frac{1}{N}\mathbf{\Psi}_N^T\mathbf{Z}_N) = \mathbf{E}\psi_k z_k$ and $\mathbf{E}\psi_k z_k = cov(\psi_k, z_k) = 0$ (see Assumption 4).

Lemma 1: The necessary condition for existence of the instrumental variables matrix $\mathbf{\Psi}_N$, which fulfills **(C2)** is asymptotic non-singularity of $\frac{1}{N}\mathbf{\Phi}_N^T\mathbf{\Phi}_N$.

Proof: (for the proof see the Appendix VIII-B). ■

After left hand side multiplying of (10) by $\mathbf{\Psi}_N^T$ we get

$$\mathbf{\Psi}_N^T\mathbf{Y}_N = \mathbf{\Psi}_N^T\mathbf{\Phi}_N\theta + \mathbf{\Psi}_N^T\mathbf{Z}_N.$$

Taking into account conditions **(C1)÷(C3)** we propose to replace the *LS* estimate, given by (11), and computed in stage 1 (see Section II) with the instrumental variables estimate

$$\widehat{\theta}_N^{(IV)} = (\mathbf{\Psi}_N^T\mathbf{\Phi}_N)^{-1}\mathbf{\Psi}_N^T\mathbf{Y}_N. \tag{19}$$

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

IV. Limit properties

For the algorithm (19) the estimation error of aggregated parameter vector θ has the form

$$\begin{aligned} \Delta_N^{(IV)} &= \widehat{\theta}_N^{(IV)} - \theta = \\ &\left(\mathbf{\Psi}_N^T\mathbf{\Phi}_N\right)^{-1}\mathbf{\Psi}_N^T\mathbf{Z}_N \\ &= \left(\frac{1}{N}\sum_{k=1}^N \psi_k\phi_k^T\right)^{-1}\left(\frac{1}{N}\sum_{k=1}^N \psi_k z_k\right). \end{aligned} \tag{20}$$

Theorem 2: Under **(C1)÷(C3)**, the estimate (19) converges in probability to the true parameters of the system, independently of the autocorrelation of the noise, i.e.,

$$P\lim_{N\to\infty}\Delta_N^{(IV)} = 0. \tag{21}$$

Proof: (for the proof see the Appendix VIII-C) ■

Theorem 3: The estimation error $\Delta_N^{(IV)}$ converges to zero with the asymptotic rate $O(\frac{1}{\sqrt{N}})$ in probability (see e.g. Definition 4 in Appendix IX-D), for each strategy of instrumental variables generation, which guarantees fulfilment of **(C1)÷(C3).**

Proof: (for the proof see the Appendix VIII-D) ■

V. Optimal instrumental variables

Theorem 3 gives universal guaranteed asymptotic rate of convergence of the estimate (19). Nevertheless, for moderate number of measurements, the error depends on particular instruments used in application. In this Section, the optimal form of instruments is established for the special case of NARMAX systems, which fulfils the following assumption concerning $\eta()$ and $\{\lambda_j\}_{j=1}^p$.

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

$$\left|\eta(y^{(1)}) - \eta(y^{(2)})\right| \leq r\left|y^{(1)} - y^{(2)}\right|, \tag{22}$$

and

$$\eta(0) = 0. \tag{23}$$

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

$$\alpha = r\sum_{j=1}^p |\lambda_j| < 1. \tag{24}$$

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

$$G_{l,k} \triangleq \mathbf{E}\{g_l(y_k) \mid \{u_i\}_{i=-\infty}^k\} \tag{25}$$

where $l = 1, 2, ...q$ and denote

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

It holds that

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

and the signals

$$\xi_{l,k} = g_l(y_k) - G_{l,k}, \tag{26}$$

for $l = 1, 2, ...q$, and $k = 1, 2, ..., N$, will be interpreted as the "noises".

The equation (1) can be now presented as follows

$$\begin{aligned} y_k &= \sum_{j=1}^p \lambda_j\eta(y_{k-j}) + \sum_{i=0}^n \gamma_i\mu(u_{k-i}) + z_k = \\ &A_k\left(\{y_{k-j}\}_{j=1}^p\right) + B_k\left(\{u_{k-i}\}_{i=1}^n\right) + C_k\left(u_k\right) + z_k, \end{aligned} \tag{27}$$

where $A_k\left(\{y_{k-j}\}_{j=1}^p\right) = \sum_{j=1}^p \lambda_j\eta(y_{k-j})$, and $B_k\left(\{u_{k-i}\}_{i=1}^n\right) = \sum_{i=1}^n \gamma_i\mu(u_{k-i})$, $C_k\left(u_k\right) = \gamma_0\mu(u_k)$. 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 (25) has the following interpretation

$$\begin{aligned} G_{l,k} = E\{&g_l(C_k\left(u_k\right) + A_k\left(\{y_{k-j}\}_{j=1}^p\right) \\ &+B_k\left(\{u_{k-i}\}_{i=1}^n\right) + z_k) \mid \{u_i\}_{i=-\infty}^k\}, \end{aligned} \tag{28}$$

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

Property (P1): The ”disturbances” $\left\{\xi_{l,k}\right\}_{k=1}^N$ given by (26) are independent of the input process $\{u_k\}$ and are all ergodic (see Appendix IX-H).

Mutual independence of $\left\{\xi_{l,k}\right\}_{k=1}^N$ and $\{u_k\}_{k=-\infty}^\infty$ is a direct consequence of definition (25). On the basis of Assumptions 3, 4 and 5 we conclude, that the output $\{y_k\}_{k=1}^N$ of the system is bounded and ergodic. Thanks to Assumption (1), concerning the nonlinear characteristics, the processes $\{g_l\left(y_k\right)\}_{k=1}^N$ and $\{G_{l,k}\}_{k=1}^N$ $(l = 1, 2, ..., q)$ are also bounded and ergodic. Consequently, the "noises" $\left\{\xi_{l,k}\right\}_{k=1}^N$ $(l = 1, 2, ..., q)$, as the sums of ergodic processes, are ergodic too (see (26)).

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

By definition (26) of $\xi_{l,k}$ we simply have

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

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

$$\psi_{k,j} = H_j(\{u_i\}_{i=-\infty}^k), \tag{29}$$

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

Owing to **(P1)** and **(P2)** we obtain

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

Property (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 $\left|\psi_{k,j}\right| = |H_j(u_k)| < \psi_{\max} < \infty$ (see 3), then

$$\frac{1}{N}\sum_{k=1}^{N}\psi_k z_k \to \mathbf{E}\psi_k z_k \tag{30}$$

with probability 1, as $N \to \infty$; (cf. 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 (see assumptions of Theorem 20). To prove (30), making use of Theorem 20 in Appendix IX-H we must show, that $r_{s_{k,j}}(\tau) \to 0$, as $|\tau| \to \infty$. Let us notice that the autocovariance function of z_k $(\mathbf{E}z_k = 0)$

$$r_z(\tau) = \mathbf{E}\left[(z_k - \mathbf{E}z)(z_{k+\tau} - \mathbf{E}z)\right] = \mathbf{E}z_k z_{k+\tau}, \tag{31}$$

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

$$r_z(\tau) \to 0 \tag{32}$$

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

$$\begin{aligned}r_{s_{k,j}}(\tau) &= \mathbf{E}\left[(s_{k,j} - \mathbf{E}s_{k,j})(s_{k+\tau,j} - \mathbf{E}s_{k,j})\right] = \\ &= \mathbf{E}\left[\psi_{k,j}\psi_{k+\tau,j}z_k z_{k+\tau}\right] = cr_z(\tau),\end{aligned} \tag{33}$$

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

$$r_{s_{k,j}}(\tau) \to 0 \tag{34}$$

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

$$\frac{1}{N}\sum_{k=1}^{N}s_{k,j} \to \mathbf{E}s_{k,j} \tag{35}$$

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

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

$$\frac{1}{N}\sum_{k=1}^{N}\psi_k\phi_k^T \to \mathbf{E}\psi_k\phi_k^T, \tag{36}$$

with probability 1 as $N \to \infty$, where ψ_k is given by (29); compare the condition **(C2)**.

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

$$y_k = \eta(y_{k-1}) + \sum_{i=0}^{n}\gamma_i\mu(u_{k-i}) + z_k, \tag{37}$$

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

$$|\eta(y)| \leq a|y|, \tag{38}$$

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

$$\delta_k = \sum_{i=0}^{n}\gamma_i\mu(u_{k-i}) + z_k, \tag{39}$$

we get

$$y_k = \eta(y_{k-1}) + \delta_k. \tag{40}$$

Since the input $\{u_k\}$ is the 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 (32)), we conclude that also $r_\delta(\tau) \to 0$, as $|\tau| \to \infty$. The equation (40) can be written in the following form

$$y_k = \delta_k + \eta\left\{\delta_{k-1} + \eta\left[\delta_{k-2} + \eta\left(\delta_{k-3} + ...\right)\right]\right\}. \tag{41}$$

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

$$c_k = \frac{\eta(y_k)}{y_k} \tag{42}$$

with $\frac{0}{0}$ treated as 0. From (38) we have that

$$|c_k| \leq a < 1, \tag{43}$$

and using c_k, the equation (41) can be rewritten as follows

$$y_k = \delta_k + c_{k-1}\left(\delta_{k-1} + c_{k-2}\left(\delta_{k-2} + c_{k-3}\left(\delta_{k-3} + ...\right)\right)\right),$$

i.e.,

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

where $c_{k,0} \triangleq 1$, and $c_{k,i} = c_{k-1}c_{k-2}...c_{k-i}$. From (43) we conclude that

$$|c_{k,i}| < a^i. \tag{44}$$

Since for $0 < a < 1$ the sum $\sum_{i=0}^{\infty}a^i$ is finite, from (44) we get $\sum_{i=0}^{\infty}|c_{k,i}| < \infty$, and from (39) we simply conclude that for $|\tau| \to \infty$ it holds that $r_y(\tau) \to 0$ and $r_{g_l(y_k)}(\tau) \to 0$, where the processes $g_l(y_k)$ $(l = 1, ..., q)$ are elements of the vector ϕ_k. Thus, for the system with the nonlinearity $\eta()$ as in (38) the processes $\{y_k\}$ and $\{g_l(y_k)\}$ $(l = 1, ..., q)$ fulfills assumption of the ergodic law of large numbers (see Theorem 20), and the property (36) holds.

Property (P5b): Under Assumption 7, the convergence (36) takes place also for the system (1) with $p \geq 1$.

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

$$\|\{x_k\}\| = \lim_{K\to\infty}\sup_{k>K}|x_k|. \tag{45}$$

and let us present the equation (1) in the form

$$y_k = \sum_{j=1}^{p}\lambda_j\eta(y_{k-j}) + \delta_k \tag{46}$$

where δ_k is given by (39).

The proof of property **(P5b)** (for $p > 1$) is based of the following theorem (see [12], page. 53).

Theorem 4: Let $\{y_k^{(1)}\}$ and $\{y_k^{(2)}\}$ be two different output sequences of the system (1) (see also (46)), and $\{\delta_k^{(1)}\}$, $\{\delta_k^{(2)}\}$ be respective aggregated inputs (see (39)). If (23), (22) and (24) are fulfilled, then

$$\frac{1}{1+\alpha}\left\|\{\delta_k^{(1)} - \delta_k^{(2)}\}\right\| \leq \left\|\{y_k^{(1)} - y_k^{(2)}\}\right\| \leq \frac{1}{1-\alpha}\left\|\{\delta_k^{(1)} - \delta_k^{(2)}\}\right\|, \tag{47}$$

where the norm $\|\ \|$ is defined in (45).

From (47) and under conditions (23), (22), (24) the steady state of the system (1) depends only on the steady state of the input $\{\delta_k\}$. The special case of (47) is $\delta_k^{(2)} \equiv 0$, in which $\lim_{K\to\infty}\sup_{k>K}\left|y_k^{(2)}\right| = 0$, and

$$\frac{1}{1+\alpha}\left\|\{\delta_k^{(1)}\}\right\| \leq \left\|\{y_k^{(1)}\}\right\| \leq \frac{1}{1-\alpha}\left\|\{\delta_k^{(1)}\}\right\|. \tag{48}$$

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

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

Moreover, on the basis of (1)–(4) since the process $\{y_k\}$ is bounded, it has finite moments of any orders and the ergotic theorems holds (see Lemma 20 and Lemma 21 in Appendix IX-H). In consequence, the convergence (36) holds.

The properties **(P5a)** and **(P5b)** (see (36), (9) and (29)) can be rewritten for particular elements of ψ_k and ϕ_k in the following way

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

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

Under the property that $\mathbf{E}\left[\psi_{k_1,j}\xi_{l,k_2}\right] = 0$ (see **(P3)**) for instrumental variables generated according to (29) we obtain

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

Denoting (cf. (9))

$$\begin{array}{rcl} \mathbf{\Phi}_N^{\#} & = & (\phi_1^{\#}, \phi_2^{\#}, ..., \phi_N^{\#})^T \\ \phi_k^{\#} & \triangleq & (f_1(u_k), ..., f_m(u_k), ..., f_1(u_{k-n}), ..., f_m(u_{k-n}), \\ & & G_{1,k-1}, ..., G_{q,k-1}, ..., G_{1,k-p}, ..., G_{q,k-p})^T \end{array} \tag{49}$$

where $G_{l,k} \triangleq \mathbf{E}\{g_l(y_k) \mid \{u_i\}_{i=-\infty}^k\}$ (see (25)), and making use of ergodicity of the processes $\{\psi_{k,j}\}$ $(j = 1, ..., m(n+1) + pq)$, $\{f_t(u_k)\}$ $(t = 1, ..., m)$ and $\{G_{l,k}\}$ $(l = 1, ..., q)$ (see (29) and Assumption 3) we get

$$\frac{1}{N}\mathbf{\Psi}_N^T\mathbf{\Phi}_N^{\#} = \frac{1}{N}\sum_{k=1}^{N}\psi_k\phi_k^{\#T} \to \mathbf{E}\psi_k\phi_k^{\#T} \text{ with p. 1}$$

and using (36) we get

$$\frac{1}{N}\mathbf{\Psi}_N^T\mathbf{\Phi}_N = \frac{1}{N}\sum_{k=1}^{N}\psi_k\phi_k^{T} \to \mathbf{E}\psi_k\phi_k^{T} \text{ with p. 1}$$

for the instruments as in (29). Directly from definitions (25) and (49) we conclude that $\mathbf{E}\left[\psi_{k_1,j}g_l(y_{k_2})\right] = \mathbf{E}\left[\psi_{k_1,j}G_{l,k_2}\right]$ and

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

Thus, for any choice of instrumental variables matrix $\mathbf{\Psi}_N$, which fulfills the property **(P3)** (see (29)), the following equivalence takes place asymptotically with probability 1, as $N \to \infty$

$$\frac{1}{N}\mathbf{\Psi}_N^T\mathbf{\Phi}_N^{\#} = \frac{1}{N}\mathbf{\Psi}_N^T\mathbf{\Phi}_N. \tag{50}$$

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

$$\Delta_N^{(IV)} = \widehat{\theta}_N^{(IV)} - \theta = \left(\frac{1}{N}\mathbf{\Psi}_N^T\mathbf{\Phi}_N\right)^{-1}\left(\frac{1}{N}\mathbf{\Psi}_N^T\mathbf{Z}_N\right).$$

Introducing

$$\begin{array}{rcl} \mathbf{\Gamma}_N & \triangleq & \left(\frac{1}{N}\mathbf{\Psi}_N^T\mathbf{\Phi}_N\right)^{-1}\frac{1}{\sqrt{N}}\mathbf{\Psi}_N^T \\ \mathbf{Z}_N^* & \triangleq & \frac{\frac{1}{\sqrt{N}}\mathbf{Z}_N}{z_{\max}} \end{array}$$

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

$$\Delta_N^{(IV)} = z_{\max}\mathbf{\Gamma}_N\mathbf{Z}_N^*. \tag{51}$$

with the Euclidean norm of $\mathbf{Z}_N^*$

$$\|\mathbf{Z}_N^*\| = \sqrt{\sum_{k=1}^{N}\left(\frac{\frac{1}{\sqrt{N}}z_k}{z_{\max}}\right)^2} = \sqrt{\frac{1}{N}\sum_{k=1}^{N}\left(\frac{z_k}{z_{\max}}\right)^2} \le 1.$$

Let the quality of the instrumental variables be evaluated on the basis of the following criterion (see e.g. [21])

$$Q\left(\mathbf{\Psi}_N\right) = \max_{\|\mathbf{Z}_N^*\| \le 1}\left\|\Delta_N^{(IV)}\left(\mathbf{\Psi}_N\right)\right\|^2 \tag{52}$$

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

Theorem 5: If the Assumptions 1–6, 7 and the condition (29) hold, then the criterion $Q\left(\mathbf{\Psi}_N\right)$ given by (52) attains minimum for the choice

$$\mathbf{\Psi}_N^{\#} = \mathbf{\Phi}_N^{\#} \tag{53}$$

i.e., for each $\mathbf{\Psi}_N$ it holds that

$$\lim_{N\to\infty} Q(\mathbf{\Psi}_N^{\#}) \leqslant \lim_{N\to\infty} Q(\mathbf{\Psi}_N) \text{ with p. 1.}$$

(for the proof see Appendix VIII-E)

Obviously,instrumental variables given by (53) fulfills postulates **(C1)÷(C3)**.

VI. Nonparametric generation of instrumental variables

The optimal matrix of instruments $\mathbf{\Psi}_N^{\#}$ 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 $\mathbf{\Psi}_N^{\#}$ is also difficult, because the elements $G_{l,k}$ depends on infinite number of measurements of the input process. Therefore, we propose the following heuristic method

$$\begin{array}{rcl} \mathbf{\Psi}_N^{(r)\#} & = & (\psi_1^{(r)\#}, \psi_2^{(r)\#}, ..., \psi_N^{(r)\#})^T, \\ \psi_k^{(r)\#} & \triangleq & (f_1(u_k), ..., f_m(u_k), ..., f_1(u_{k-n}), ..., f_m(u_{k-n}), \\ & & G_{1,k-1}^{(r)}, ..., G_{q,k-1}^{(r)}, ..., G_{1,k-p}^{(r)}, ..., G_{q,k-p}^{(r)})^T \end{array}$$

where

$$\begin{array}{rcl} G_l^{(r)} & = & G_l^{(r)}(u^{(0)}, ..., u^{(r)}) \triangleq \mathbf{E}\{g_l(y_j) \mid u_j = u^{(0)}, ..., u_{j-r} = u^{(r)} \\ G_{l,k}^{(r)} & = & G_l^{(r)}(u_k, ..., u_{k-r}). \end{array} \tag{54}$$

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

$$\mathbf{\Psi}_N^{(r)\#} \cong \mathbf{\Psi}_N^{\#}$$

when r grows (this question is treated as open). For $r = 0$ we have

$$\begin{array}{rcl} \mathbf{\Psi}_N & = & \mathbf{\Psi}_N^{(0)\#}, \\ \psi_k^{(0)\#} & \triangleq & (f_1(u_k), .., f_m(u_k), .., f_1(u_{k-n}), .., f_m(u_{k-n}), \\ & & R_1(u_{k-1}), .., R_q(u_{k-1}), .., R_1(u_{k-p}), .., R_q(u_{k-p}))^T \end{array}$$

where

$$R_l(u) = G_l^{(0)}(u) = \mathbf{E}\{g_l(y_k)\} \mid u_k = u\}. \tag{55}$$

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

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

the regression functions in (55) can be written as

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

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. The most natural method for generation of $\mathbf{\Psi}_N^{(r)\#}$ is thus the kernel method.

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. [4])

$$\widehat{R}_{l,M}(u) = \frac{\frac{1}{M}\sum_{i=1}^{M}\left(x_{l,i}K\left(\frac{u-u_i}{h(M)}\right)\right)}{\frac{1}{M}\sum_{i=1}^{M}K\left(\frac{u-u_i}{h(M)}\right)}, \tag{56}$$

where $K()$ is a kernel function, and $h()$ – the bandwidth parameter.

Further considerations will be based on the following two theorems, proved in [4] and [6].

Theorem 6: If $h(M) \to 0$ and $Mh(M) \to \infty$ for $M \to \infty$, and $K(v)$ is one of $\exp(-|v|)$, $\exp(-v^2)$, or $\frac{1}{1+|v|^{1+\delta}}$, then

$$\frac{\frac{1}{M}\sum_{i=1}^{M}\left(y_iK\left(\frac{u-u_i}{h(M)}\right)\right)}{\frac{1}{M}\sum_{i=1}^{M}K\left(\frac{u-u_i}{h(M)}\right)} \to \mathbf{E}\{y_i \mid u_i = u\} \tag{57}$$

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

Theorem 7: If both the regression $\mathbf{E}\{y_i \mid u_i = u\}$, 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 above theorems, let us additionally take the following assumption.

Assumption 8: The functions $g_1(y)$,...,$g_q(y)$, $f_1(u)$,...,$f_m(u)$ 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 (56) is correlated. Let us decompose the sums in numerator and denominator in (56) 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.

$$\begin{aligned} L(\{(u_i, x_{l,i})\}_{i=1}^{M}) &\triangleq \frac{1}{M}\sum_{i=1}^{M} x_{l,i}K\left(\frac{u-u_i}{h(M)}\right) = \frac{1}{r}\sum_{t=1}^{r} s_t \\ W(\{u_i\}_{i=1}^{M}) &\triangleq \frac{1}{M}\sum_{i=1}^{M} K\left(\frac{u-u_i}{h(M)}\right) = \frac{1}{r}\sum_{t=1}^{r} w_t \end{aligned} \tag{58}$$

with

$$\begin{aligned} s_t &= \frac{1}{\frac{M}{r}} \sum_{\{i:0<ir+t\leq M\}} x_{l,ir+t}K\left(\frac{u-u_{ir+t}}{h(M)}\right), \\ w_t &= \frac{1}{\frac{M}{r}} \sum_{\{i:0<ir+t\leq M\}} K\left(\frac{u-u_{ir+t}}{h(M)}\right). \end{aligned} \tag{59}$$

The components of the sum (58) 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 (59) are i.i.d. Each of the sub-sums $\{s_t\}$ has the same probability density, but uses different subset of measurements. All of them includes $\overline{M} = \frac{M}{r}$ data. For simplicity let us write

$$\begin{aligned} s_t &= \frac{1}{\overline{M}} \sum_{\{i:0<ir+t\leq M\}} x_{l,ir+t}K\left(\frac{u-u_{ir+t}}{H(\overline{M})}\right) \\ w_t &= \frac{1}{\overline{M}} \sum_{\{i:0<ir+t\leq M\}} K\left(\frac{u-u_{ir+t}}{H(\overline{M})}\right) \end{aligned} \tag{60}$$

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

$$H(\overline{M}) = cM^{\alpha} = c\left(M^{\frac{1-\frac{1}{\chi(M)}}{1-\frac{1}{\chi(M)}}}\right)^{\alpha} = c\left(\overline{M}^{\alpha}\right)^{\frac{1}{1-\frac{1}{\chi(M)}}} = O(\overline{M}^{\alpha}) \tag{61}$$

and for $\overline{M} \to \infty$, it holds that

$$H(\overline{M}) \to 0 \text{ and } \overline{M}H(\overline{M}) \to \infty. \tag{62}$$

From (60), (61), (62) and Theorem 6, for $r \to \infty$ we get

$$P\lim_{\overline{M}\to\infty}\left(\frac{s_t}{w_t}\right) = \frac{P\lim_{\overline{M}\to\infty}(s_t)}{P\lim_{\overline{M}\to\infty}(w_t)} = \frac{a(u)}{b(u)} = R_l(u)$$

for each $t = 1, 2, ..., r$, 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 that

$$P\lim_{M\to\infty}\left(\widehat{R}_{l,M}(u)\right) = R_l(u). \tag{63}$$

Under Assumption 8, from the property (61) and Theorem 7 we conclude that for $h(M) = cM^{-\frac{1}{5}}$ the rate of convergence of (56) is $O(M^{-\frac{2}{5}})$ in probability.

VII. The 3-stage identification

Taking into account the conclusions from Section VI, in particular the form of optimal instruments $\mathbf{\Psi}_N^*$, we propose the following identification procedure.

Stage 1 (nonparametric): Using $M + \max(n,p)$ measurements $\{(u_i, y_i)\}_{i=1-\max(n,p)}^{M}$ generate empirical matrix of instruments $\widehat{\Psi}_{N,M}^* = (\widehat{\psi}_{1,M}^*, \widehat{\psi}_{2,M}^*, ...\widehat{\psi}_{N,M}^*)^T$, where

$$\begin{aligned} \widehat{\psi}_{k,M}^* = & \ (f_1(u_k), ..., f_m(u_k), ..., f_1(u_{k-n}), ... \\ & ..., f_m(u_{k-n}), \widehat{R}_{1,M}(u_{k-1}), ..., \\ & \widehat{R}_{q,M}(u_{k-1}), ..., \widehat{R}_{1,M}(u_{k-p}), ..., \widehat{R}_{q,M}(u_{k-p}))^T \end{aligned} \tag{64}$$

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

Stage 2 (parametric): Estimate the aggregated parameter vector (8)

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

by the instrumental variables method

$$\widehat{\theta}_{N,M}^{*(IV)} = \left(\widehat{\Psi}_{N,M}^{*T}\Phi_N\right)^{-1}\widehat{\Psi}_{N,M}^{*T}\mathbf{Y}_N \tag{65}$$

where $\mathbf{Y}_N = (y_1, y_2, ..., y_N)^T$, $\Phi_N = (\phi_1, \phi_2, ..., \phi_N)^T$, $\phi_k = (f_1(u_k), ..., f_m(u_k), ..., f_1(u_{k-n}), ..., f_m(u_{k-n}), g_1(y_{k-1}), ..., g_q(y_{k-1}), ..., g_1(y_{k-p}), ..., g_q(y_{k-p}))^T$, (see (9)), 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 $\Theta_{\lambda d} = \Lambda\mathbf{d}^T$ and $\Theta_{\gamma c} = \mathbf{\Gamma}\mathbf{c}^T$.

Stage 3 (decomposition): Compute the SVD (singular value decomposition) 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)

$$\begin{aligned} \widehat{\mathbf{\Lambda}}_N &= sgn(\widehat{\xi}_1[\kappa_{\xi_1}])\widehat{\xi}_1 \qquad \widehat{\mathbf{\Gamma}}_N = sgn(\widehat{\mu}_1[\kappa_{\mu_1}])\widehat{\mu}_1 \\ \widehat{\mathbf{c}}_N &= sgn(\widehat{\mu}_1[\kappa_{\mu_1}])\sigma_1\widehat{\nu}_1 \qquad \widehat{\mathbf{d}}_N = sgn(\widehat{\xi}_1[\kappa_{\xi_1}])\delta_1\widehat{\zeta}_1 \end{aligned} \tag{66}$$

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

Under condition (63) the following theorem holds.

Theorem 8: For the NARMAX system with the characteristic $\eta(y)$ as in Assumption 7 it holds that

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

as $M \to \infty$ and $N \to \infty$, provided that $h(M)$ fulfills assumptions of Theorem 6.

(for the proof – see the Appendix VIII-F)

VIII. Proofs of theorems and lemmas

A. Hammerstein system as a special case of NARMAX system

Lemma 9: 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'

$$x_k \triangleq \sum_{i=0}^{n} b_i \mu(u_{k-i}) + v_k, \tag{67}$$

resembles the difference equation of AR linear model

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

which can be presented equivalently as ([8])

$$y_k = \sum_{l=0}^{\infty} r_l x_{k-l}. \tag{68}$$

Inserting (67) to (68) we obtain that

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

and further

$$y_k = \sum_{q=0}^{\infty} \gamma_q \mu(u_{k-q}) + z_k, \tag{69}$$

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 (69) represents Hammerstein system with infinite impulse response. ■

B. The necessary condition for the 3-Stage algorithm

Lemma 10: If $\det(\mathbf{B}^T \mathbf{A}) \neq 0$, for a given matrices $\mathbf{A}, \mathbf{B} \in \mathbf{R}^{\alpha \times \beta}$ with finite elements, then $\det(\mathbf{A}^T \mathbf{A}) \neq 0$.

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

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

we conclude that one can find the non-zero vector $\boldsymbol{\xi} \in \mathbf{R}^{\beta}$, such that $\mathbf{A}\boldsymbol{\xi} = 0$. Multiplying this equation by $\mathbf{B}^T$ we get $\mathbf{B}^T \mathbf{A} \boldsymbol{\xi} = 0$, and hence $\det(\mathbf{B}^T \mathbf{A}) = 0$. ■

For $\mathbf{A} = \frac{1}{\sqrt{N}} \mathbf{\Phi}_N$ and $\mathbf{B} = \frac{1}{\sqrt{N}} \mathbf{\Psi}_N$ we conclude that the necessary condition for $\frac{1}{N} \mathbf{\Psi}_N^T \mathbf{\Phi}_N$ to be of full rank is $\det(\frac{1}{N} \mathbf{\Phi}_N^T \mathbf{\Phi}_N) \neq 0$, i.e., persistent excitation of $\{\phi_k\}$.

C. Proof of Theorem 2

Proof: From the Slutzky theorem (cf. [3] and Appendix IX-E) we have

$$P \lim_{N \to \infty} (\Delta_N^{(IV)}) = \left(P \lim_{N \to \infty} \left(\frac{1}{N} \mathbf{\Psi}_N^T \mathbf{\Phi}_N \right) \right)^{-1} P \lim_{N \to \infty} \left(\frac{1}{N} \mathbf{\Psi}_N^T \mathbf{Z}_N \right),$$

and directly from the conditions **(C2)** and **(C3)** it holds that

$$P \lim_{N \to \infty} \left(\Delta_N^{(IV)} \right) = 0. \tag{70}$$

■

D. Proof of theorem 3

Proof: Let us define the scalar random variable

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

where $\| \ \|$ denotes any vector norm. We must show that

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

for each $\varepsilon > 0$, each $r_N \to 0$ and $a_N = \frac{1}{\sqrt{N}}$ (see Definition 4). Using Lemma 15, to prove that $\xi_N = O(\frac{1}{\sqrt{N}})$ in probability, we show that $\xi_N = O(\frac{1}{N})$ in the mean square sense. Introducing

$$\begin{aligned} \mathbf{A}_N &= \frac{1}{N} \mathbf{\Psi}_N^T \mathbf{\Phi}_N = \frac{1}{N} \sum_{k=1}^{N} \psi_k \phi_k^T, \\ \mathbf{B}_N &= \frac{1}{N} \mathbf{\Psi}_N^T \mathbf{Z}_N = \frac{1}{N} \sum_{k=1}^{N} \psi_k z_k, \end{aligned}$$

we obtain that

$$\Delta_N^{(IV)} = \mathbf{A}_N^{-1} \mathbf{B}_N. \tag{71}$$

Under Assumptions 1–6 we conclude that the system output y_k is bounded, i.e., $|y_k| < y_{\max} < \infty$. Moreover, under condition **(C1)**, it holds that

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

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

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

for $j = m(n+1)+1, ..., m(n+1)+pq$, so each element of $\mathbf{A}_N$ is bounded. Similarly, one can show boundedness of elements of the vector $\mathbf{B}_N$. The norm of the error error $\Delta_N^{(IV)}$ given by (71) can be evaluated as follows

$$\begin{aligned} \xi_N &= \left\| \Delta_N^{(IV)} \right\| = \left\| \left(\frac{1}{N} \mathbf{\Psi}_N^T \mathbf{\Phi}_N \right)^{-1} \left(\frac{1}{N} \mathbf{\Psi}_N^T \mathbf{Z}_N \right) \right\| \\ &\leq \left\| \left(\frac{1}{N} \mathbf{\Psi}_N^T \mathbf{\Phi}_N \right)^{-1} \right\| \left\| \frac{1}{N} \mathbf{\Psi}_N^T \mathbf{Z}_N \right\| \leq \\ &\leq c \left\| \frac{1}{N} \mathbf{\Psi}_N^T \mathbf{Z}_N \right\| = c \left\| \frac{1}{N} \sum_{k=1}^{N} \psi_k z_k \right\| \end{aligned}$$

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

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

and hence

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

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

$$\mathbf{E}\xi_N^2 \le \alpha^2 c^2 \dim\psi_k \sum_{i=1}^{\dim\psi_k} \frac{1}{N^2} E\left(\sum_{k=1}^{N} \psi_{k,i} z_k\right)^2 =$$

$$= \alpha^2 c^2 \dim\psi_k \sum_{i=1}^{\dim\psi_k} \frac{1}{N^2}\mathbf{E}\left[\sum_{k_1=1}^{N}\sum_{k_2=1}^{N} \psi_{k_1,i}\psi_{k_2,i} z_{k_1} z_{k_2}\right] \le$$

$$\le \alpha^2 c^2 \dim\psi_k \sum_{i=1}^{\dim\psi_k} \frac{1}{N^2}\sum_{k_1=1}^{N}\sum_{k_2=1}^{N} \left|\mathbf{E}\left[\psi_{k_1,i}\psi_{k_2,i}\right]\right| \left|\mathbf{E}\left[z_{k_1} z_{k_2}\right]\right| \le$$

$$\le \alpha^2 c^2 \left(\dim\psi_k\right)^2 \frac{\psi_{\max}^2}{N}\left[\left|r_z(0)\right| + 2\sum_{\tau=1}^{N}\left(1-\frac{\tau}{N}\right)\left|r_z(\tau)\right|\right]$$

$$\le \frac{C}{N}\sum_{\tau=0}^{\infty}\left|r_z(\tau)\right|,$$

where

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

Since

$$\left|var\varepsilon\sum_{\tau=0}^{\infty}\sum_{i=0}^{\infty}\omega_i\omega_{i+\tau}\right| \le var\varepsilon\sum_{\tau=0}^{\infty}\sum_{i=0}^{\infty}\left|\omega_i\right|\left|\omega_{i+\tau}\right|$$

$$\le var\varepsilon\sum_{i=0}^{\infty}\left|\omega_i\right|\sum_{i=0}^{\infty}\left|\omega_{i+\tau}\right| < \infty,$$

we conclude that

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

where $D = Cvar\varepsilon\left|\sum_{\tau=0}^{\infty}\sum_{i=0}^{\infty}\omega_i\omega_{i+\tau}\right|$. ■

E. Proof of theorem 5

Proof: To simplify presentation let $z_{\max} = 1$. From (51) we get

$$\left\|\Delta_N^{(IV)}(\mathbf{\Psi}_N)\right\|^2 = \Delta_N^{(IV)T}(\mathbf{\Psi}_N)\Delta_N^{(IV)}(\mathbf{\Psi}_N) = \mathbf{Z}_N^{*T}\mathbf{\Gamma}_N^T\mathbf{\Gamma}_N\mathbf{Z}_N^*,$$

and the maximum value of cumulated error is

$$Q(\mathbf{\Psi}_N) = \max_{\left\|\mathbf{Z}_N^*\right\|\le 1}\left(\Delta_N^{(IV)T}(\mathbf{\Psi}_N)\Delta_N^{(IV)}(\mathbf{\Psi}_N)\right)$$

$$= \max_{\left\|\mathbf{Z}_N^*\right\|\le 1}\left\langle \mathbf{Z}_N^*, \mathbf{\Gamma}_N^T\mathbf{\Gamma}_N\mathbf{Z}_N^*\right\rangle =$$

$$= \left\|\mathbf{\Gamma}_N\right\|^2 = \lambda_{\max}\left(\mathbf{\Gamma}_N^T\mathbf{\Gamma}_N\right),$$

where $\|\ \|$ is the spectral matrix norm induced by the Euclidean vector norm, and $\lambda_{\max}()$ denotes the biggest eigenvalue of the matrix. Since [17],[21]

$$\lambda_{\max}\left(\mathbf{\Gamma}_N^T\mathbf{\Gamma}_N\right) = \lambda_{\max}\left(\mathbf{\Gamma}_N\mathbf{\Gamma}_N^T\right),$$

from definition of $\mathbf{\Gamma}_N$ we conclude that

$$\max_{\left\|\mathbf{Z}_N^*\right\|\le 1}\left(\Delta_N^{(IV)T}(\mathbf{\Psi}_N)\Delta_N^{(IV)}(\mathbf{\Psi}_N)\right)$$

$$= \max_{\|\zeta\|\le 1}\left\langle \zeta, \mathbf{\Gamma}_N\mathbf{\Gamma}_N^T\zeta\right\rangle$$

$$= \max_{\|\zeta\|\le 1}\left\langle \zeta, \left(\frac{1}{N}\mathbf{\Psi}_N^T\mathbf{\Phi}_N\right)^{-1}\left(\frac{1}{N}\mathbf{\Psi}_N^T\mathbf{\Psi}_N\right)\left(\frac{1}{N}\mathbf{\Phi}_N^T\mathbf{\Psi}_N\right)^{-1}\zeta\right\rangle.$$

On the basis of (50), it holds that

$$\max_{\left\|\mathbf{Z}_N^*\right\|\le 1}\left(\Delta_N^{(IV)T}(\mathbf{\Psi}_N)\Delta_N^{(IV)}(\mathbf{\Psi}_N)\right)$$

$$= \max_{\|\zeta\|\le 1}\left\langle \zeta, \left(\frac{1}{N}\mathbf{\Psi}_N^T\mathbf{\Phi}_N^{\#}\right)^{-1}\left(\frac{1}{N}\mathbf{\Psi}_N^T\mathbf{\Psi}_N\right)\left(\frac{1}{N}\mathbf{\Phi}_N^{\#T}\mathbf{\Psi}_N\right)^{-1}\zeta\right\rangle,$$

with probability 1, as $N \to \infty$, where $\mathbf{\Phi}_N$ and $\mathbf{\Phi}_N^{\#}$ are given by (9) and (49), respectively. Using Lemma 13 for $\mathbf{M}_1 = \frac{1}{\sqrt{N}}\mathbf{\Phi}_N^{\#}$ and $\mathbf{M}_2 = \frac{1}{\sqrt{N}}\mathbf{\Psi}_N$ we get

$$\zeta^T\mathbf{\Gamma}_N\mathbf{\Gamma}_N^T\zeta \ge \zeta^T\left(\frac{1}{N}\mathbf{\Phi}_N^{\#T}\mathbf{\Phi}_N^{\#}\right)^{-1}\zeta$$

for each vector ζ, and consequently

$$Q\left(\mathbf{\Psi}_N\right) = \max_{\|\zeta\|\le 1}\left(\zeta^T\mathbf{\Gamma}_N\mathbf{\Gamma}_N^T\zeta\right) \ge \max_{\|\zeta\|\le 1}\left(\zeta^T\left(\frac{1}{N}\mathbf{\Phi}_N^{\#T}\mathbf{\Phi}_N^{\#}\right)^{-1}\zeta\right).$$

For $\mathbf{\Psi}_N = \mathbf{\Phi}_N^{\#}$, it holds that

$$\max_{\|\zeta\|\le 1}\left(\zeta^T\mathbf{\Gamma}_N\mathbf{\Gamma}_N^T\zeta\right) = \max_{\|\zeta\|\le 1}\left(\zeta^T\left(\frac{1}{N}\mathbf{\Phi}_N^{\#T}\mathbf{\Phi}_N^{\#}\right)^{-1}\zeta\right)$$

and the criterion $Q\left(\mathbf{\Psi}_N\right)$ attains minimum. The choice $\mathbf{\Psi}_N = \mathbf{\Phi}_N^{\#}$ is thus asymptotically optimal. ■

F. Proof of theorem 8

Proof: The estimation error (65) 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 $\widehat{\theta}_N^{*(IV)} = \left(\Psi_N^{*T}\Phi_N\right)^{-1}\Psi_N^{*T}\mathbf{Y}_N$, and Ψ_N^* is defined by (53) and (49). From the triangle inequality, for each norm $\|\ \|$ it holds that

$$\left\|\Delta_{N,M}^{(IV)}\right\| \le \left\|\widehat{\theta}_{N,M}^{*(IV)} - \widehat{\theta}_N^{*(IV)}\right\| + \left\|\widehat{\theta}_N^{*(IV)} - \theta\right\|. \qquad (72)$$

On the basis of Theorem 2

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

as $N \to \infty$. To prove 8, let us analyze the component $\left\|\widehat{\theta}_{N,M}^{*(IV)} - \widehat{\theta}_N^{*(IV)}\right\|$ in (72) to show that, for fixed N, it tends to zero in probability, as $M \to \infty$. let us denote (cf. [9], page. 116-117)

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

From (63) we have that

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

as $M \to \infty$, and particularly

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

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 Banach Theorem (see [12], Theorem 5.8., page. 106) we get

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

Since $r_M \to 0$ in probability as $M \to \infty$, we finally conclude that

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

as $M \to \infty$, for each N. ∎

IX. Algebra toolbox

A. SVD decomposition

Theorem 11: [10] For each $\mathbf{A} \in \mathbf{R}^{m,n}$ it exists the unitary matrices $\mathbf{U} \in \mathbf{R}^{m,m}$, and $\mathbf{V} \in R^{n,n}$, such that

$$\mathbf{U}^T \mathbf{A} \mathbf{V} = \mathbf{\Sigma} = diag(\sigma_1, ..., \sigma_l), \tag{73}$$

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

$$\begin{aligned} \sigma_1 &\geq \sigma_2 \geq ... \geq \sigma_r > 0 \\ \sigma_{r+1} &= ... = \sigma_l = 0 \end{aligned}$$

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

The numbers $\sigma_1, ..., \sigma_l$ are called the singular values of the matrix $\mathbf{A}$. Solving (73) with respect to $\mathbf{A}$ we obtain

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T = \sum_{i=1}^{r} \mathbf{u}_i \sigma_i \mathbf{v}_i^T = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T, \tag{74}$$

where $\mathbf{u}_i$ and $\mathbf{v}_i$ denote i-th columns of $\mathbf{U}$ and $\mathbf{V}$, respectively [10].

B. Factorization theorem

Theorem 12: [17] Each positive definite matrix $\mathbf{M}$ can be shown in the form

$$\mathbf{M} = \mathbf{P}\mathbf{P}^T$$

where $\mathbf{P}$ (root of $\mathbf{M}$) is nonsingular.

C. Technical lemma

Lemma 13: (for the proof see [21]) let $\mathbf{M}_1$ and $\mathbf{M}_2$ be two matrices with the same dimensions. If $\left(\mathbf{M}_1^T \mathbf{M}_1\right)^{-1}$, $\left(\mathbf{M}_1^T \mathbf{M}_2\right)^{-1}$ and $\left(\mathbf{M}_2^T \mathbf{M}_1\right)^{-1}$ exist, then

$$\mathbf{D}_N = \left(\mathbf{M}_2^T \mathbf{M}_1\right)^{-1} \mathbf{M}_2^T \mathbf{M}_2 \left(\mathbf{M}_1^T \mathbf{M}_2\right)^{-1} - \left(\mathbf{M}_1^T \mathbf{M}_1\right)^{-1}$$

jest is nonnegative definite, i.e., for each ζ it holds that

$$\zeta^T \mathbf{D}_N \zeta \geq 0.$$

D. Types of convergence of random sequences

Definition 1: [3]The sequence of random variables $\{\varkappa_k\}$ converges, for $k \to \infty$, with probability 1 (strongly) to $\varkappa^*$, if

$$P(\lim_{k\to\infty} \varkappa_k = \varkappa^*) = 1.$$

Definition 2: [3]The sequence of random variables $\{\varkappa_k\}$ converges, for $k \to \infty$, in probability (weakly) to $\varkappa^{\#}$, if

$$\lim_{k\to\infty} P(\left|\varkappa_k - \varkappa^{\#}\right| > \varepsilon) = 0,$$

for each $\varepsilon > 0$. The $\varkappa^{\#}$ is denoted as a probabilistic limit of $\varkappa_k$

$$\text{Plim}_{k\to\infty} \varkappa_k = \varkappa^{\#}. \tag{75}$$

Lemma 14: If $\varkappa_k \to \varkappa$ with probability 1, as $k \to \infty$, then $\varkappa_k \to \varkappa$ in probability, as $k \to \infty$.

Definition 3: [3] The sequence of random variables $\{\varkappa_k\}$ converges, for $k \to \infty$, in the mean square sense to $\varkappa^*$, if

$$\lim_{k\to\infty} \mathbf{E}(\varkappa_k - \varkappa^*)^2 = 0.$$

Definition 4: [5] The sequence of random variables $\{\varkappa_k\}$ has the rate of convergence $O(e_k)$ in probability as $k \to \infty$, where $\{e_k\}$ is deterministic number sequence which tends to zero, i.e.,

$$\varkappa_k = O(e_k) \text{ in probability,}$$

if $\left\{\frac{\varkappa_k}{e_k}\chi_k\right\} \to 0$ in probability for each number sequence $\{\chi_k\}$, such that $\lim_{k\to\infty} \chi_k = 0$.

Definition 5: [5] The sequence of random variables $\{\varkappa_k\}$ has the rate of convergence $O(e_k)$ in the mean square sense, as $k \to \infty$, if it exists the constant $0 \leq c < \infty$, such that

$$\mathbf{E}\varkappa_k^2 \leq ce_k.$$

Lemma 15: [5] If $\varkappa_k = O(e_k)$ in the mean square sense, then $\varkappa_k = O(\sqrt{e_k})$ in probability.

E. Slutzky theorem

Theorem 16: ([17]) If $\text{Plim}_{k\to\infty} \varkappa_k = \varkappa^{\#}$ and the function $g()$ is continuous, then

$$P \lim_{k\to\infty} g(\varkappa_k) = g(\varkappa^{\#}).$$

F. Chebychev's inequality

Lemma 17: ([3], page 106) For each constant c, each random variable X and each $\varepsilon > 0$ it holds that

$$P\left\{|X - c| > \varepsilon\right\} \leq \frac{1}{\varepsilon^2} \mathbf{E}\left(X - c\right)^2.$$

In particular, for $c = \mathbf{E}X$

$$P\left\{|X - \mathbf{E}X| > \varepsilon\right\} \leq \frac{1}{\varepsilon^2} varX.$$

G. Persistent excitation

Definition 6: ([20]) The stationary random process $\{\alpha_k\}$ is strongly persistently exciting of orders $n \times m$, (denote $SPE(n, m)$) if the matrix

$$R_{\varkappa}(n, m) = \mathbf{E} \begin{bmatrix} \varkappa_k \\ : \\ \varkappa_{k-n+1} \end{bmatrix} \begin{bmatrix} \varkappa_k \\ : \\ \varkappa_{k-n+1} \end{bmatrix}^T$$

where $\varkappa_k = \begin{bmatrix} \alpha_k & \alpha_k^2 & .. & \alpha_k^m \end{bmatrix}^T$, is of full rank.

Lemma 18: ([20]) The i.i.d. process $\{\alpha_k\}$ is $SPE(n, m)$ for each n and m.

Lemma 19: ([20]) Let $x_k = H(q^{-1})u_k$, $H(q^{-1})$ be 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.

H. Ergodic processes

Definition 7: ([19]) The stationary stochastic process $\{\varkappa_k\}$ is ergodic with respect to the first and the second order moments if

$$\begin{aligned} \frac{1}{N}\sum_{k=1}^{N} \varkappa_k &\to \mathbf{E}\varkappa_k \\ \frac{1}{N}\sum_{k=1}^{N} \varkappa_k \varkappa_{k+\tau} &\to \mathbf{E}\varkappa_k \varkappa_{k+\tau} \end{aligned}$$

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

Theorem 20: (see [16], or [19]) Let us assume that $\{\varkappa_k\}$ is a discrete-time random process with finite variance. If the autocorrelation of $\{\varkappa_k\}$ is such that $r_{\varkappa}(\tau) \to 0$ for $|\tau| \to \infty$, then

$$\frac{1}{N}\sum_{k=1}^{N} \varkappa_k \to \mathbf{E}\varkappa \tag{76}$$

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

Theorem 21: (cf. [16], or [19]) If the two random processes $\{\varkappa_{1,k}\}$ and $\{\varkappa_{2,k}\}$ have finite fourth order moments and $r_{\varkappa_1}(\tau) \to 0$, $r_{\varkappa_2}(\tau) \to 0$ as $|\tau| \to \infty$, then

$$\frac{1}{N}\sum_{k=1}^{N} \varkappa_{1,k}\varkappa_{2,k} \to \mathbf{E}\varkappa_{1,k}\varkappa_{2,k}$$

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

I. Modified triangle inequality

Lemma 22: [3] If X and Y are k-dimensional random vectors, then $P\left[\|X+Y\| \geqslant \varepsilon\right] \leqslant P\left[\|X\| \geqslant \frac{\varepsilon}{2}\right] + P\left[\|Y\| \geqslant \frac{\varepsilon}{2}\right]$ for each vector norm $\|\bullet\|$ and each $\varepsilon > 0$.

Proof: Let us define the following random events: A: $\|X+Y\| \geqslant \varepsilon$, B: $\|X\| + \|Y\| \geqslant \varepsilon$, C: $\|X\| \geqslant \frac{\varepsilon}{2}$, D: $\|Y\| \geqslant \frac{\varepsilon}{2}$. Obviously $A \Longrightarrow B$ and $B \Longrightarrow (C \smile D)$. Thus $A \subset B \subset (C \smile D)$ and $P\left(A\right) \leqslant P\left(B\right) \leqslant P(C \smile D) \leqslant P(C) + P(D)$. ■

References

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

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

[3] Y.S. Chow and H. Teicher. *Probability theory: independence, interchangeability, martingales*. Springer Verlag, 2003.

[4] W. Greblicki, A. Krzyzak, and M. Pawlak. Distribution-free pointwise consistency of kernel regression estimate. *The Annals of Statistics*, 12(4):1570–1575, 1984.

[5] W. Greblicki and M. Pawlak. Fourier and Hermite series estimates of regression functions. *Annals of the Institute of Statistical Mathematics*, 37(1):443–454, 1985.

[6] W. Greblicki and M. Pawlak. Identification of discrete Hammerstein systems using kernel regression estimates. *IEEE Transactions on Automatic Control*, 31:74–77, 1986.

[7] R. Haber. Structural identification of quadratic block-oriented models based on estimated Volterra kernels. *International journal of systems science*, 20(8):1355–1380, 1989.

[8] E.J. Hannan and M. Deistler. The statistical theory of linear systems. *New York*, pages 161–222, 1988.

[9] Z. Hasiewicz. *Identyfikacja sterowanych systemów o złożonej strukturze, (in Polish)*. Wydawn. Politechniki Wrocławskiej, 1993.

[10] A. Kiełbasiński and H. Schwetlick. *Numeryczna algebra liniowa: wprowadzenie do obliczeń zautomatyzowanych, (in Polish)*. Wydawnictwa Naukowo-Techniczne, 1992.

[11] D.R. Kincaid and E.W. Cheney. *Numerical analysis: mathematics of scientific computing*, volume 2. Amer Mathematical Society, 2002.

[12] J. Kudrewicz. *Analiza funkcjonalna dla automatyków i elektroników, (in Polish)*. Państwowe Wydawn. Naukowe, 1976.

[13] G. Mzyk. Application of instrumental variable method to the identification of Hammerstein-Wiener systems. In *Proceedings of the 6th International Conference MMAR*, pages 951–956, 2000.

[14] G. Mzyk. Kernel-based instrumental variables for NARMAX system identification. *Proceedings of the ICSES*, pages 469–475, 2001.

[15] G. Mzyk. Zastosowanie metody zmiennych instrumentalnych do identyfikacji systemów Hammersteina-Wienera. *Pomiary Automatyka Kontrola*, (7/8):35–40, 2001.

[16] B. Ninness. Strong laws of large numbers under weak assumptions with application. *Automatic Control, IEEE Transactions on*, 45(11):2117–2122, 2000.

[17] C.R. Rao. 1973. linear statistical inference and its applications.

[18] T. Söderström and P. Stoica. *Instrumental variable methods for system identification*, volume 161. Springer-Verlag Berlin, 1983.

[19] T. Söderström and P. Stoica. *System Identification*. NJ: Prentice Hall, Englewood Cliffs, 1989.

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

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

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