International Journal of Control
Vol. 82, No. 3, March 2009, 440–455

Taylor & Francis
Taylor & Francis Group

Hammerstein system identification by non-parametric instrumental variables

Zygmunt Hasiewicz and Grzegorz Mzyk*

Institute of Computer Engineering, Control and Robotics, Wrocław University of Technology, Wroclaw, Poland

(*Received 8 May 2007; final version received 22 May 2008*)

A mixed, parametric–non-parametric routine for Hammerstein system identification is presented. Parameters of a non-linear characteristic and of ARMA linear dynamical part of Hammerstein system are estimated by least squares and instrumental variables assuming poor *a priori* knowledge about the random input and random noise. Both subsystems are identified separately, thanks to the fact that the unmeasurable interaction inputs and suitable instrumental variables are estimated in a preliminary step by the use of a non-parametric regression function estimation method. A wide class of non-linear characteristics including functions which are not linear in the parameters is admitted. It is shown that the resulting estimates of system parameters are consistent for both white and coloured noise. The problem of generating optimal instruments is discussed and proper non-parametric method of computing the best instrumental variables is proposed. The analytical findings are validated using numerical simulation results.

Keywords: Hammerstein system; parameter estimation; instrumental variables; non-parametric regression; convergence analysis

1. Introduction

Conception of the block-oriented models in system identification (interconnections of static non-linearities and linear dynamics) has been introduced in 1980s by Billings and Fakhouri (1982), as an alternative for Volterra series expansions. It was commonly accepted because of satisfactory approximation capabilities of various real processes and relatively small model complexity. The decentralised approach to the identification of block-oriented complex systems seems to be the most natural and desirable, and as such an approach corresponds directly to the own nature of systems composed of individual elements distinguished in the structure (Narendra and Gallman 1966; Chang and Luus 1971; Bai 2002; Bai and Li 2004) and tries to treat the components 'locally' as independent, autonomous objects. The Hammerstein system, built of a static non-linearity and a linear dynamics connected in a cascade (Figure 1), is the simplest structure in the class and hence for the most part considered in the system identification literature (see e.g. Giannakis and Serpedin (2001) for the bibliography). Unfortunately, the popular parametric methods elaborated for Hammerstein system identification do not allow full decentralisation of the system identification task, i.e. independent identification of a static non-linearity and a linear dynamics – first of all, because of inaccessibility for measurements of the inner interconnection signal. They assume that the description of system components, i.e. of a static non-linearity and a linear dynamics is known up to the parameters (a polynomial model along with a FIR dynamics representation are usually used) and these parameters are 'glued' when using standard input-output data of the overall system for identification purposes (e.g. Billings and Fakhouri 1982; Stoica and Söderström 1982a,b; Lang 1993). On the other hand, in a non-parametric setting (the second class of existing identification methods, see, e.g. Greblicki and Pawlak (1986, 1994); Pawlak and Hasiewicz (1998) no preliminary assumptions concerning the structure of subsystems are used and only the data decide about the obtained characteristics of the system components but then any possible *a priori* knowledge about the true description of subsystems is not exploited, i.e. inevitably lost. We propose a method where the two approaches are combined. Namely, our idea is to join the results obtained for the non-parametric identification of non-linear characteristics in Hammerstein systems (see the articles cited above), in particular by using kernel regression methods (Greblicki and Pawlak 1986, 1994) with parametric knowledge of subsystems and standard results concerning least squares and instrumental variables (see, for instance,

*Corresponding author. Email: grzegorz.mzyk@pwr.wroc.pl

ISSN 0020–7179 print/ISSN 1366–5820 online

DOI: 10.1080/00207170802225930
http://www.informaworld.com

Wong and Polak (1967); Ward (1977) and Söderström and Stoica (1989), taking advantages of both. The article is an extension of Hasiewicz and Mzyk (2004), where the combined parametric–non-parametric algorithm has been proposed for the identification of parameters appearing linearly in the static non-linear element and the estimation of the impulse response of a FIR linear dynamics in Hammerstein system.

The organisation of the article is as follows. In § 2 the identification problem is stated in detail and *a priori* assumptions concerning signals and subsystems of Hammerstein system are formulated. In § 3 the two-stage parametric–non-parametric estimate for a static characteristic is proposed. Stage 1 of the algorithm (non-parametric) consists in non-parametric estimation of interaction inputs $w_k = \mu(u_k)$ (Figure 1) to cope with their inaccessibility for direct measurements. In Stage 2 (parametric), using the obtained estimates $\widehat{w}_k$ of w_k, we identify parameters of non-linearity by minimisation of appropriate empirical loss function. It is shown in particular that the considered loss function is bounded from the above and from the below by two paraboloids. This fact is further exploited for proving consistency and rate of convergence of the parameter estimate of the non-linearity. Section 4 concerns identification of a linear dynamic subsystem. The instrumental variables type estimate is analysed for identification of parameters of ARMA model. Both internal signal and instruments are established by using the non-parametric regression function estimation methodology. The convergence rate is strictly proved and the problem of optimal, in the minimax sense, selection of instrumental variables is solved. Finally, in § 5, the effectiveness of the approach, even under incomplete *a priori* knowledge, is illustrated in simulation examples. For easier readability the proofs of theorems are deferred to the Appendix.

Summing up, the contribution of the article is the following:

(i) We develop a method of decomposing Hammerstein system identification task on fully independent partial identification problems of component subsystems, robust to the lack or inaccuracy of the prior knowledge of complementary subsystem, based on non-parametric estimation of interaction inputs.

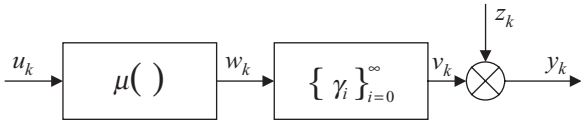

Figure 1. The Hammerstein system – the general form.

(ii) We provide a two-stage identification routine of the static non-linearity with general parametric description not necessarily linear in the parameters exploiting non-parametric estimates of interactions, and discuss parameter identifiability conditions, show consistency of the resulting parameter estimate of the non-linearity and examine the asymptotic rate of convergence.

(iii) We provide an identification algorithm of the linear dynamics of the ARMA-type (i.e. possessing infinite impulse response) with correlated output noise of arbitrary correlation structure, based on the concept of instrumental variables and non-parametric estimation of instruments, and further we give convergence conditions and the asymptotic rate of convergence of adequate system parameter estimates.

(iv) We resolve the problem of the optimum selection of instruments in the minimax sense and propose a suitable routine for estimating optimum instrumental variables.

(v) We verify efficiency of the proposed identification schemes, in particular their robustness to the incomplete or inaccurate knowledge of companion subsystems in the Hammerstein system, by means of computer simulations and show usability of the methods.

As to the items (i)–(iii), they are extensions of the ideas and results presented in our earlier paper (Hasiewicz and Mzyk 2004) towards more general model structures, on the one hand non-linear in the parameters – with respect to the static non-linearity, and with infinite impulse response – and regarding linear dynamics on the other. Particularly, the item (iii) as against the results presented in Hasiewicz and Mzyk (2004), in the context of complex Hammerstein system identification is the same generalisation as instrumental variables are in relation to ordinary least squares in the standard identification tasks of single-element linear dynamic systems with measurable inputs and outputs. In turn, the topic (iv) is specific for instrumental variables approach and the results in (v) illustrate the behaviour of parameter estimates for small and moderate number of measurement data and complete the theoretical asymptotic analysis of the estimates presented in the core of the article where we provide the proofs of consistency of the resulting parameter estimates and discuss the rate of convergence. This is in contrast to most of the literature concerning Hammerstein system identification in the parametric setting, where the proofs of convergence of the presented algorithms are often omitted

(Narendra and Gallman 1966; Chang and Luus 1971; Billings and Fakhouri 1982). We also point out that our proofs of consistency refer to rather complicated nested structures 'non-parametric-in-the-parametric estimate'.

2. Statement of the problem

2.1 *Hammerstein system*

We consider the discrete-time Hammerstein system as in Figure 1, where u_k, y_k and z_k are, respectively, the input, output, and noise at time k, and w_k is the interaction input not available for measurement, which is a typical limitation in the literature (see Billings and Fakhouri (1982) for the discussion).

2.2 *General assumptions*

In this section we assume the following.

Assumption 1: *The input signal* $\{u_k\}$ *is for* $k=\ldots,-1,0,1,\ldots$, *an i.i.d. bounded random process* $|u_k|\leqslant u_{\max}$, *some* $u_{\max}>0$, *and there exists a probability density of* u_k, *say* $\nu(u)$.

Assumption 2: *The non-linear characteristic* $\mu(u)$ *is a Borel measurable bounded function on the interval* $[-u_{\max},u_{\max}]$, *i.e.*

$$\left|\mu(u)\right|\leqslant w_{\max} \tag{1}$$

where $w_{\max}$ *is some positive constant.*

Assumption 3: *The linear dynamics is an asymptotically stable IIR filter*:

$$v_k=\sum_{i=0}^{\infty}\gamma_i w_{k-i} \tag{2}$$

with the unknown impulse response $\{\gamma_i\}_{i=0}^{\infty}$ (*such that* $\sum_{i=0}^{\infty}|\gamma_i|<\infty$).

Assumption 4: *The output noise* $\{z_k\}$ *is a random, arbitrarily correlated process, governed by the general equation*:

$$z_k=\sum_{i=0}^{\infty}\omega_i\varepsilon_{k-i} \tag{3}$$

where $\{\varepsilon_k\}$, $k=\ldots,-1,0,1,\ldots$, *is a bounded stationary zero-mean white noise* ($E\varepsilon_k=0, |\varepsilon_k|\leqslant\varepsilon_{\max}$), *independent of the input signal* $\{u_k\}$, *and* $\{\omega_i\}_{i=0}^{\infty}$ *is unknown*; $\sum_{i=0}^{\infty}|\omega_i|<\infty$. *Hence the noise* $\{z_k\}$ *is a stationary zero-mean and bounded process* $|z_k|\leqslant z_{\max}$, *where* $z_{\max}=\varepsilon_{\max}\sum_{i=0}^{\infty}|\omega_i|$.

Assumption 5: $\mu(u_0)$ *is known at some point* u_0 *and* $\gamma_0=1$.

As it was explained in detail in Hasiewicz and Mzyk (2004), the input-output pair $(u_0,\mu(u_0))$ assumed to be known can refer to arbitrary $u_0\in[-u_{\max},u_{\max}]$, and hence we shall further assume for convenience that $u_0=0$ and $\mu(0)=0$, without loss of generality. The essence of Assumption 5 is explained further in § 3, in the comment on the proposed identification scheme.

2.3 *Parametric prior knowledge of subsystems*

Additional assumptions, specify parametric *a priori* knowledge of $\mu()$ and IIR dynamics which is available in our identification task.

Assumption 6: *As regards the non-linearity we suppose the following*:

A6.1 *The form of a static non-linearity* (1) *is known up to the parameters, i.e. we are given the function* $\mu(u,c)$ *and the set* C *of admissible parameters (it is the pair* $(\mu(u,c),C)$) *such that* $\mu(u,c^*)=\mu(u)$ (Figure 2), *where* $c^*=(c_1^*,c_2^*,\ldots,c_m^*)^T\in C$ *is a vector of the unknown true parameters of the non-linearity.*

A6.2 *The function* $\mu(u,c)$ *is continuous and differentiable in the set* C, *and the gradient* $\nabla_c\mu(u,c)$ *is bounded, i.e.*

$$\left\|\nabla_c\mu(u,c)\right\|\leqslant G_{\max}<\infty\ ,\quad c\in C$$

for each $u\in[-u_{\max},u_{\max}]$.

A6.3 *The admissible set of parameters* C *is bounded and small enough* (*cf Remark* 1 *below*), *and can be considered as a neighbourhood of each admissible parameter vector* $c\in C$, *and in particular, of the true parameter vector* c^*, i.e. $C\equiv\mathcal{O}(c^*)$.

A6.4 *Each parameter* $c\in C$, *and in particular the true parameter vector* $c^*\in C$ *is identifiable in the set* C, *i.e. there exists a sequence of inputs* (*design points*) $\overline{u}_1,\overline{u}_2,\ldots,\overline{u}_{N_0}$, *independent of* c, *such that*

$$\mu(\overline{u}_n,c)=\mu(\overline{u}_n,c^*),\quad n=1,2,\ldots,N_0\Rightarrow c=c^* \tag{4}$$

for each pair $(c,c^*)\in C\times C$.

Remarks below present some comments and motivations concerning Assumptions A6.3 and A6.4.

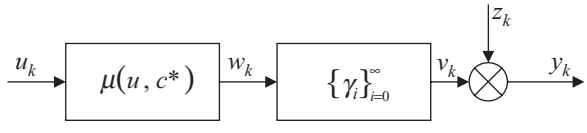

Figure 2. The Hammerstein system with $\mu(u)=\mu(u,c^*)$.

Remark 1: Assumptions A6.1 and A6.2 constitute general (and typical) parametric *a priori* information about the real static non-linearity $\mu(u)$. However many practical problems involve non-linear effects that are prone to a high amount of detailed structure and hence are difficult to describe parametrically by a single model as in Assumption A6.1. In circumstances where single and smooth (Assumption A6.2) model is planned to be used to handle non-linearity, it should be kept in mind that in the real world its use is possible only locally, both with respect to the inputs (Assumption 1) and the parameters (Assumption A6.3), and can require a good deal of expertise. For comprehensive discussion of this and related issues we refer the reader to Ruppert, Wand and Carroll (2003).

Remark 2: As regards Assumption A6.4 for $c \in C$, where Assumptions A6.2 and A6.3 allow a one-term Taylor expansion of the form

$$\mu(\overline{u}_n, c) = \mu(\overline{u}_n, c^*) + \nabla_c^T \mu(\overline{u}_n, c^*)(c - c^*) + o(\|c - c^*\|)$$

and $o(\|c - c^*\|)$ is negligible, we have factually

$$\mu(\overline{u}_n, c) - \mu(\overline{u}_n, c^*) = \nabla_c^T \mu(\overline{u}_n, c^*)(c - c^*), \qquad (5)$$

hence the identifiability condition (4) can be rewritten as

$$\nabla_c^T \mu(\overline{u}_n, c^*)(c - c^*) = 0, \quad n = 1, 2, \ldots, N_0 \Rightarrow c = c^*$$

or that

$$J\big(\overline{u}_1, \overline{u}_2, \ldots, \overline{u}_{N_0}; c^*\big)(c - c^*) = 0$$

implies $c = c^*$, where $J(\overline{u}_1, \overline{u}_2, \ldots, \overline{u}_{N_0}; c^*) = [\nabla_c \mu(\overline{u}_1, c^*), \nabla_c \mu(\overline{u}_2, c^*), \ldots, \nabla_c \mu(\overline{u}_{N_0}, c^*)]^T$ is the Jacobian matrix. This means that – from the viewpoint of the true (but unfortunately unknown) parameter vector c^* – the input sequence $\{\overline{u}_n;\ n = 1, 2, \ldots, N_0\}$ in (4) should be selected such that

$$\operatorname{rank} J\big(\overline{u}_1, \overline{u}_2, \ldots, \overline{u}_{N_0}; c^*\big) = \dim c \qquad (6)$$

i.e. $N_0 \geq \dim c$ (the necessary condition) and

$$\det J^T\big(\overline{u}_1, \overline{u}_2, \ldots, \overline{u}_{N_0}; c^*\big) J\big(\overline{u}_1, \overline{u}_2, \ldots, \overline{u}_{N_0}; c^*\big) > 0$$

To get in turn independent of c^* and of any $c \in C$ (universal) choice of proper $\overline{u}_n$'s one should need that

$$\inf_{c \in C} \big\{\det J^T(\overline{u}_1, \overline{u}_2, \ldots, \overline{u}_{N_0}; c) J(\overline{u}_1, \overline{u}_2, \ldots, \overline{u}_{N_0}; c)\big\} > 0 \qquad (7)$$

In the particular case of linear-in-the-parameters static characteristics $\mu(u, c) = \phi^T(u)c$ discussed in Hasiewicz and Mzyk (2004), where $\phi(u) = (f_1(u), f_2(u), \ldots, f_m(u))^T$ is a vector of linearly independent basis functions, these requirements take the form (compare Remark 2 in Hasiewicz and Mzyk (2004)

$$\operatorname{rank}\big(\phi(\overline{u}_1), \phi(\overline{u}_2), \ldots, \phi(\overline{u}_{N_0})\big)^T = \dim c$$

or

$$\det \Phi_{N_0}^T(\overline{u}_1, \overline{u}_2, \ldots, \overline{u}_{N_0}) \Phi_{N_0}(\overline{u}_1, \overline{u}_2, \ldots, \overline{u}_{N_0}) > 0$$

where $\Phi_{N_0}(\overline{u}_1, \overline{u}_2, \ldots, \overline{u}_{N_0}) = (\phi(\overline{u}_1), \phi(\overline{u}_2), \ldots, \phi(\overline{u}_{N_0}))^T$.

Assumption 7: *About the linear element* (2) *we suppose that it is an* ARMA(s, p) *block, i.e. that it can be described in the parametric form by the following difference equation*:

$$v_k = b_0 w_k + \cdots + b_s w_{k-s} + a_1 v_{k-1} + \cdots + a_p v_{k-p};\ p \geqslant s \qquad (8)$$

with known orders s, p and unknown parameters b_0, $b_1, \ldots, b_s$ and $a_1, a_2, \ldots, a_p$, or equivalently as $A(q^{-1})v_k = B(q^{-1})w_k$, where $B(q^{-1}) = b_0 + b_1 q^{-1} + \cdots + b_s q^{-s}$, $A(q^{-1}) = 1 - a_1 q^{-1} - \cdots - a_p q^{-p}$ and q^{-1} is a backward shift operator (see Figure 3). Because of Assumption 5 it holds that $b_0 = 1$.

As it will be seen, when identifying dynamic subsystem, in the intermediate step Assumption 6 will be weakened to Assumption 2, i.e. the possessed parametric knowledge of the static element will be ignored. Conversely, we shall see that Assumption 7 may be weakened to Assumption 3 during identification process of the static part. The aim is to discover the true parameters of subsystems, respectively $c^* = (c_1^*, c_2^*, \ldots, c_m^*)^T$ and $\theta = (b_0, b_1, \ldots, b_s, a_1, a_2, \ldots, a_p)^T$, using a set of input-output data $\{(u_k, y_k)\}$ collected from the whole system in an identification experiment.

3. Estimation of the non-linearity parameters

Let $\overline{u}_1, \overline{u}_2, \ldots, \overline{u}_{N_0}$ be such that the unknown c^* is identifiable (see Remark 2). Denote

$$W_{N_0} = (w_1, w_2, \ldots, w_{N_0})^T \qquad (9)$$

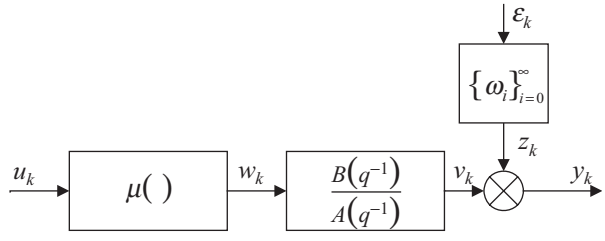

Figure 3. The Hammerstein system with the ARMA linear element.

where $w_n = \mu(\overline{u}_n, c^*)$ for $n = 1, 2, \ldots, N_0$ and introduce the vector

$$\overline{\mu}_{N_0}(c) = \left[\mu(\overline{u}_1, c),\ \mu(\overline{u}_2, c), \ldots, \mu(\overline{u}_{N_0}, c)\right]^T$$

Obviously, the square of the Euclidean norm of the difference between $\overline{\mu}_{N_0}(c)$ and W_{N_0}, i.e. the index

$$Q_{N_0}(c) = \left\|\overline{\mu}_{N_0}(c) - W_{N_0}\right\|^2$$

takes minimum value only for the vector of true parameters c^* of the non-linearity, i.e. the minimiser

$$c^* = \arg\min \left\|\overline{\mu}_{N_0}(c) - W_{N_0}\right\|^2$$

is unique and hence the identification routine may be based on the minimisation of the residual sum of squares

$$Q_{N_0}(c) = \sum_{n=1}^{N_0} [w_n - \mu(\overline{u}_n, c)]^2; \quad c \in C \tag{10}$$

Because of $w_n = \mu(\overline{u}_n, c^*)$ and due to (5) in Remark 2, we have

$$Q_{N_0}(c) = \sum_{n=1}^{N_0} \left|\nabla_c^T \mu(\overline{u}_n, c^*)(c - c^*)\right|^2$$

and by the Schwartz inequality we ascertain that

$$Q_{N_0}(c) \leqslant \left\|J(\overline{u}_1, \overline{u}_2, \ldots, \overline{u}_{N_0}; c^*)\right\|_F^2 \cdot \|c - c^*\|^2$$

where

$$\left\|J(\overline{u}_1, \overline{u}_2, \ldots, \overline{u}_{N_0}; c^*)\right\|_F^2 = \sum_{n=1}^{N_0} \left\|\nabla_c \mu(\overline{u}_n, c^*)\right\|^2$$

is squared Frobenius norm of the Jacobian matrix $J(\overline{u}_1, \overline{u}_2, \ldots, \overline{u}_{N_0}; c^*)$. Hence, by virtue of Assumption A6.2 we get

$$\left\|J(\overline{u}_1, \overline{u}_2, \ldots, \overline{u}_{N_0}; c^*)\right\|_F^2 \leqslant N_0 G_{\max}^2 \triangleq D$$

which yields that

$$Q_{N_0}(c) \leqslant D \cdot \|c - c^*\|^2, \quad \forall c \in C \tag{11}$$

On the other hand, since

$$Q_{N_0}(c) = (c - c^*)^T J^T(\overline{u}_1, \overline{u}_2, \ldots, \overline{u}_{N_0}; c^*) \times J(\overline{u}_1, \overline{u}_2, \ldots, \overline{u}_{N_0}; c^*)(c - c^*)$$

and for the selected $\overline{u}_1, \overline{u}_2, \ldots, \overline{u}_{N_0}$ the matrix $J^T(\overline{u}_1, \overline{u}_2, \ldots, \overline{u}_{N_0}; c^*)J(\overline{u}_1, \overline{u}_2, \ldots, \overline{u}_{N_0}; c^*)$ is symmetric and positive-definite (by fulfillment of the identifiability condition for the selected input series $\overline{u}_1, \overline{u}_2, \ldots, \overline{u}_{N_0}$; see Remark 2), there exists a $\delta > 0$ such that

$$Q_{N_0}(c) \geqslant \delta \cdot \|c - c^*\|^2, \quad \forall c \in C \tag{12}$$

This follows from the observation that

$$\delta \cdot \|c - c^*\|^2 = (c - c^*)^T \delta I (c - c^*)$$

the fact that all eigenvalues of the matrix $J^T(\overline{u}_1, \overline{u}_2, \ldots, \overline{u}_{N_0}; c^*)J(\overline{u}_1, \overline{u}_2, \ldots, \overline{u}_{N_0}; c^*)$, say $\{\lambda_i\}_{i=1}^m$, are positive, $\lambda_i > 0$ for $i = 1, 2, \ldots, m$ (see e.g. the property 6.O, page 335 in Strang (2003), and the fact that eigenvalues of the matrix $J^T(\overline{u}_1, \overline{u}_2, \ldots, \overline{u}_{N_0}; c^*) \times J(\overline{u}_1, \overline{u}_2, \ldots, \overline{u}_{N_0}; c^*) - \delta I$ are $\{\lambda_i - \delta\}_{i=1}^m$ (see Lemma 5 in Appendix), whence it suffices to take some $0 < \delta \leqslant \min\{\lambda_i\}_{i=1}^m$ to assure (12), or, in other words, positive semidefiniteness of the matrix $J^T(\overline{u}_1, \overline{u}_2, \ldots, \overline{u}_{N_0}; c^*) \times J(\overline{u}_1, \overline{u}_2, \ldots, \overline{u}_{N_0}; c^*) - \delta I$.

The bounds (11) and (12) yield together that in the parameter set C it holds that

$$\delta \cdot \|c - c^*\|^2 \leqslant Q_{N_0}(c) \leqslant D \cdot \|c - c^*\|^2 \tag{13}$$

or equivalently (due to $Q_{N_0}(c^*) = 0$) that

$$\delta \cdot \|c - c^*\|^2 \leqslant Q_{N_0}(c) - Q_{N_0}(c^*) \leqslant D \cdot \|c - c^*\|^2 \tag{14}$$

This shows that in the case under discussion the loss function $Q_{N_0}(c)$ with $c \in C$, computed for the input sequence $\overline{u}_1, \overline{u}_2, \ldots, \overline{u}_{N_0}$, is located in a region bounded by two paraboloids originated at c^*, and certifies identifiability of the parameter vector c^* for the selected input points $\{\overline{u}_n; n = 1, 2, \ldots, N_0\}$ (as $Q_{N_0}(c) = 0$ implies $c = c^*$). We stress that in our consideration $Q_{N_0}(c)$ in itself does not need to be a convex function in the set C.

Because of inaccessibility of interactions $\{w_n\}_{n=1}^{N_0}$ appearing in (10), the direct minimisation of (10) w.r.t. c is not possible. Instead an indirect two-stage procedure can be proposed where, in Stage 1, a non-parametric estimation of w_ns is carried out yielding the estimates $\widehat{w}_n$, and next, in Stage 2, a parametric minimisation of $Q_{N_0}(c)$ is completed employing the obtained $\widehat{w}_n$ instead of w_n. The arising identification scheme is then the following.

Select $\overline{u}_1, \overline{u}_2, \ldots, \overline{u}_{N_0}$ so that the unknown parameter vector c^* be identifiable, i.e. due to the selection rule (7) in Remark 2.

Stage 1 (non-parametric): On the basis of M input-output measurement data $\{(u_k, y_k)\}_{k=1}^M$, for the selected N_0 input points $\{\overline{u}_n;\ n = 1, 2, \ldots, N_0\}$ estimate the corresponding interactions $\{w_n = \mu(\overline{u}_n, c^*);\ n = 1, 2, \ldots, N_0\}$ as

$$\widehat{w}_{n,M} = \widehat{R}_M(\overline{u}_n) - \widehat{R}_M(0), \tag{15}$$

where $\widehat{R}_M(u)$ is a non-parametric estimate of the regression function $R(u)=E[y_k|u_k=u]$.

Stage 2 (parametric): Plug in the estimates $\widehat{w}_{n,M}$ obtained in Stage 1 to the loss function (10) in place of w_n and minimise the quality index:

$$\widehat{Q}_{N_0,M}(c)=\sum_{n=1}^{N_0}\left[\widehat{w}_{n,M}-\mu(\overline{u}_n,c)\right]^2 \tag{16}$$

getting the solution $\hat{c}_{N_0,M}$. Take the computed $\hat{c}_{N_0,M}$ as the estimate of c^*.

The idea behind Stage 1 originates from the fact that under Assumptions 2–4 it holds that (see, e.g. Greblicki and Pawlak (1986))

$$R(u)=\gamma_0\mu(u)+d$$

where $d=E\mu(u_1)\sum_{i=1}^{\infty}\gamma_i$ and $\mu(\,)$ is the Hammerstein system non-linearity, and under Assumptions 5 and 6 along with the fact that by Assumption $\mu(0)=0$ we further get that

$$\mu(u,c^*)=R(u)-R(0)$$

Remark 3: In fact, Assumption 5 and the above routine can be generalised as in general fashion it holds that

$$E\{y_{k+l}|u_k=u\}=\gamma_l\mu(u)+\delta_l$$

where $\delta_l=E\mu(u_1)\cdot\sum_{i\neq l}\gamma_i$, i.e. the lagged regression of the output on the input in Hammerstein system is a scalled and properly shifted version of the non-linear characteristic $\mu()$, i.e. factually we can estimate $\mu(u)$ for each time-lag l between output and input, for which $\gamma_l\neq 0$. Certainly, the best choice of l is for the $|\gamma_l|$ being maximal. For further discussion we refer to Mzyk (2007).

Such a routine is consistent, i.e. provides the estimate $\hat{c}_{N_0,M}$ which converges (as $M\to\infty$) to the true parameter vector c^* provided that $\widehat{R}_M(u)$ is a consistent estimate of the regression function $R(u)$. The following theorem refers to this property.

Theorem 1: *Assume that the computed* $\hat{c}_{N_0,M}$ *is unique and for each* M, $\hat{c}_{N_0,M}\in C$. *If in Stage* 1 (*non-parametric*), *for some* $\tau>0$, *it holds that*

$$\widehat{R}_M(\overline{u}_n)=R(\overline{u}_n)+O(M^{-\tau})\ \textit{in probability} \tag{17}$$

as $M\to\infty$ *for* $n=1,2,\ldots,N_0$ *and for* $\overline{u}_n=0$ *then*

$$\hat{c}_{N_0,M}=c^*+O(M^{-\tau})\ \textit{in probability} \tag{18}$$

as $M\to\infty$.

Proof: Owing to (10) and (16), we see that

$$\widehat{Q}_{N_0,M}(c)=Q_{N_0}(c)+2\sum_{n=1}^{N_0}(w_n-\mu(\overline{u}_n,c))\left(\widehat{w}_{n,M}-w_n\right)+\sum_{n=1}^{N_0}\left(\widehat{w}_{n,M}-w_n\right)^2$$

and hence (by Cauchy inequality $(x+y)^2\leqslant 2x^2+2y^2$) we obtain

$$\begin{aligned}\left|\widehat{Q}_{N_0,M}(c)-Q_{N_0}(c)\right| &\leqslant\left(2\sum_{n=1}^{N_0}\left|\mu(\overline{u}_n,c^*)-\mu(\overline{u}_n,c)\right|\right)\cdot\Delta_M(0)\\ &\quad+2\sum_{n=1}^{N_0}\left|\mu(\overline{u}_n,c^*)-\mu(\overline{u}_n,c)\right|\cdot\Delta_M(\overline{u}_n)\\ &\quad+2N_0\Delta_M^2(0)+2\sum_{n=1}^{N_0}\Delta_M^2(\overline{u}_n)\end{aligned}$$

where $\Delta_M(\overline{u}_n)=|\widehat{R}_M(\overline{u}_n)-R(\overline{u}_n)|$. In particular, for each $c\in\overline{C}$ (a closure of C) by continuity of $\mu(u,c)$ (Assumption A6.2) we have $|\mu(\overline{u}_n,c)-\mu(\overline{u}_n,c^*)|\leqslant A_n$, where A_n is a constant (possibly depending on the input point $\overline{u}_n$), and eventually we get

$$\begin{aligned}\sup_{c\in\overline{C}}\left|\widehat{Q}_{N_0,M}(c)-Q_{N_0}(c)\right| &\leqslant 2A\left[N_0\Delta_M(0)+\sum_{n=1}^{N_0}\Delta_M(\overline{u}_n)\right]\\ &\quad+2\left[N_0\Delta_M^2(0)+\sum_{n=1}^{N_0}\Delta_M^2(\overline{u}_n)\right]\end{aligned} \tag{19}$$

where $A=\max\{A_1,A_2,\ldots,A_{N_0}\}$. Under assumptions of theorem

$$\sup_{c\in\overline{C}}\left|\widehat{Q}_{N_0,M}(c)-Q_{N_0}(c)\right|\to 0\ \text{in probability}$$

as $M\to\infty$, and thus $\hat{c}_{N_0,M}\to c^*$ in probability as M grows (cf Vapnik 1982; van der Vaart 1988). Asymptotically (i.e. for M large) we have (see (12))

$$Q_{N_0}(\hat{c}_{N_0,M})\geqslant\delta\left|\hat{c}_{N_0,M}-c^*\right|^2$$

and hence, for each $\varepsilon>0$, it holds that

$$P\left\{\left\|\hat{c}_{N_0,M}-c^*\right\|>(\varepsilon/\delta)^{1/2}\right\}\leqslant P\left\{Q_{N_0}(\hat{c}_{N_0,M})>\varepsilon\right\}$$

Since further

$$\begin{aligned}Q_{N_0}(\hat{c}_{N_0,M}) &=Q_{N_0}(\hat{c}_{N_0,M})-Q_{N_0}(c^*)\\ &\leqslant 2\sup_{c\in C}\left|\widehat{Q}_{N_0,M}(c)-Q_{N_0}(c)\right|\\ &\leqslant 2\sup_{c\in\overline{C}}\left|\widehat{Q}_{N_0,M}(c)-Q_{N_0}(c)\right|\end{aligned}$$

we conclude (18) including (19) and (17). □

The theorem says that under our assumptions both convergence and the guaranteed rate of convergence of the non-parametric estimate $\widehat{R}_M(u)$ are conveyed on the estimate $\hat{c}_{N_0,M}$ of the parameter vector c^*, i.e. the estimate $\hat{c}_{N_0,M}$ converges to c^* in the same sense and with the same guaranteed speed as $\widehat{R}_M(u)$ to $R(u)$. The proof shows that each kind of probabilistic convergence of $\hat{c}_{N_0,M}$ could potentially be considered in Theorem 1. We confined ourselves to the convergence in probability as such particular type of convergence has been widely examined in the literature concerning non-parametric estimation of non-linearities (regression functions) for Hammerstein systems (Greblicki and Pawlak 1986, 1994; Greblicki 1989). As is well known, the non-parametric rate of convergence $O(M^{-\tau})$ appearing in (17) is usually of slower order than typical parametric rate of convergence $O(M^{-1/2})$ in probability, i.e. as a rule $0 < \tau < 1/2$ (cf e.g. Härdle 1990). Thus in our method, we can lose a little (for $\tau \approx 1/2$) the convergence speed of the estimate of c^*, but instead the successful recovering of c^* can be performed without prior knowledge (and hence even under false knowledge) of a companion dynamic part of the system. Moreover, for smooth non-linearities $\mu(u) = \mu(u, c^*)$ the convergence rate can be made arbitrarily close to $O(M^{-1/2})$ by applying proper non-parametric regression function estimates $\widehat{R}_M(u)$ (cf e.g. Greblicki 1989; Greblicki and Pawlak 1994; Hasiewicz, Pawlak and Śliwiński 2005).

The class of consistent non-parametric regression function estimates $\widehat{R}_M(u)$ which can be employed in Stage 1, and were elaborated up to now in the system identification literature for Hammerstein systems, encompasses kernel and orthogonal series estimates, including wavelet estimates possessing excellent adaptation properties and being able to attain the best possible non-parametric rate of convergence in the sense of Stone (cf Stone 1980; Hasiewicz et al. 2005). An excellent introduction to the broad field of non-parametric regression is provided by Takezawa (2006). For a general treatment of nonparametric regression function estimation methods, we refer the reader to e.g. Härdle (1990); Wand and Jones (1995); Efromovich (1999).

As a simple example from a broad variety of accessible non-parametric regression function estimates we present below the kernel estimate, very easy for computation.

Example: Kernel regression estimate (studied in Greblicki and Pawlak (1986), (1994) has the form

$$\widehat{R}_M(u) = \frac{\sum_{k=1}^{M} y_k K[(u-u_k)/h(M)]}{\sum_{k=1}^{M} K[(u-u_k)/h(M)]} = \sum_{k=1}^{M}\left[\frac{K[(u-u_k)/h(M)]}{\sum_{k=1}^{M} K[(u-u_k)/h(M)]}\right]\cdot y_k \qquad (20)$$

where $K(u)$ is a kernel (weighting) function and $h(M)$ is a bandwidth parameter controlling the range of data used for estimating a regression function $R(u)$ at a given point u. Standard examples are $K(u) = I_{[-0.5,0.5]}(u)$, $(1-|u|)I_{[-1,1]}(u)$ or $(1/\sqrt{2\pi})e^{-u^2/2}$ and $h(M) = \text{const} \cdot M^{-\alpha}$ with $0 < \alpha < 1$. Owing to the convergence results provided in Wand and Jones (1995), we find out that for each of the above kernels $K(u)$ it holds that $\widehat{R}_M(u) \to R(u)$ in probability as $M \to \infty$, and that the convergence takes place at every point $u \in \text{Cont}(\mu, \nu)$, the set of continuity points of $\mu(u)$ and $\nu(u)$, at which $\nu(u) > 0$ where $\nu(u)$ is a probability density of the system input (assumed to exist, cf Assumption 1). Applying in particular the Gaussian kernel $K(u) = (1/\sqrt{2\pi})e^{-u^2/2}$ and taking, according to the recommendation in Greblicki and Pawlak (1994), $h(M) \sim M^{-1/5}$ we attain in Stage 1 the convergence rate $|\widehat{R}_M(u) - R(u)| = O(M^{-2/5})$ in probability very close to $O(M^{-1/2})$, provided that $\mu(u)$ and $\nu(u)$ are at the selected estimation points $u \in \{0, \overline{u}_n;\ n = 1, 2, \ldots, N_0\}$ (see (15) and (17)) two times or more continuously differentiable functions and $\nu(u) > 0$ there (cf Greblicki and Pawlak 1986, 1994).

As regards the parametric optimisation task (unconstrained non-linear least squares) to be solved in Stage 2 we can use to this end for instance the standard Levenberg-Marquardt method. Minimisation of $\widehat{Q}_{N_0,M}(c)$ in Stage 2 of the identification procedure is then performed with the use of the following iterative routine:

$$\hat{c}^{(i+1)}_{N_0,M} = \hat{c}^{(i)}_{N_0,M} - \left[J^T\left(\hat{c}^{(i)}_{N_0,M}\right)J\left(\hat{c}^{(i)}_{N_0,M}\right) + \lambda_i I\right]^{-1} \times J^T\left(\hat{c}^{(i)}_{N_0,M}\right)r\left(\hat{c}^{(i)}_{N_0,M}\right)$$

where $J(c) = J(\overline{u}_1, \overline{u}_2, \ldots, \overline{u}_{N_0}; c^*)$ is the Jacobian matrix of the form $J(c)[n,j] = (\partial r_n(c)/\partial c_j)$ $(n = 1, 2, \ldots, N_0;\ j = 1, 2, \ldots, m)$, where $r_n(c) = \mu(\overline{u}_n, c) - \widehat{w}_{n,M}$ and $r(c) = (r_1(c), r_2(c), \ldots, r_{N_0}(c))^T$. The λ_is are weighting coefficients (continuation parameters) modified according to the rule:

$$\lambda_{i+1} = \begin{cases} \lambda_i \cdot \nu, & \text{if } \widehat{Q}_{N_0,M}\left(\hat{c}^{(i+1)}_{N_0,M}\right) \geqslant \widehat{Q}_{N_0,M}\left(\hat{c}^{(i)}_{N_0,M}\right) \\ \lambda_i/\nu, & \text{otherwise} \end{cases}$$

where $\nu > 1$ (see Moré (1977); Seber and Wild (1989); Press, Teukolsky, Vetterling and Flannery (1992) for various implementations and discussion of Levenberg-Marquardt method).

4. Identification of ARMA linear dynamics by non-parametric instrumental variables

4.1 *Least squares* versus *instrumental variables*

Since $v_k = y_k - z_k$ (Figure 3) thus the noisy dynamics output y_k can be expressed as (cf (8))

$$y_k = \vartheta_k^T \theta + \bar{z}_k \tag{21}$$

where $\theta = (b_0, b_1, \ldots, b_s, a_1, a_2, \ldots, a_p)^T$ is a vector of unknown true parameters of the linear dynamics ($p \geqslant s$), $\vartheta_k = (w_k, w_{k-1}, \ldots, w_{k-s}, y_{k-1}, y_{k-2}, \ldots, y_{k-p})^T$ is a generalised input vector and

$$\bar{z}_k = z_k - a_1 z_{k-1} - \cdots - a_p z_{k-p}$$

is a proper, zero-mean and stationary resultant disturbance. For a set of N generalised input-output data $\{(\vartheta_k, y_k)\}$, we can write concisely

$$Y_N = \Theta_N \theta + Z_N$$

where $Y_N = (y_1, y_2, \ldots, y_N)^T$, $\Theta_N = (\vartheta_1, \vartheta_2, \ldots, \vartheta_N)^T$ and $Z_N = (\bar{z}_1, \bar{z}_2, \ldots, \bar{z}_N)^T$. To this end we need in fact to have at out disposal the input-output data $\{(u_k, y_k)\}$ for $1 - p \leqslant k \leqslant N$, where p is the dynamics order.

Since the matrix Θ_N contains among others the regressors $y_{k-1}, y_{k-2}, \ldots, y_{k-p}$ and the noise $\{\bar{z}_k\}$ is not white, the popular least squares estimate of θ, of the form

$$\widehat{\theta}_N^{(LS)} = \left(\Theta_N^T \Theta_N\right)^{-1} \Theta_N^T Y_N \tag{22}$$

is obviously biased (Hannan and Deistler 1998; Stoica and Söderström 2002). As is well known, we can overcome this weakness by using instrumental variables approach, yielding the estimate (Söderström and Stoica 1989)

$$\widehat{\theta}_N^{(IV)} = \left(\Psi_N^T \Theta_N\right)^{-1} \Psi_N^T Y_N \tag{23}$$

where Ψ_N is a matrix of properly selected instruments

$$\begin{aligned} \Psi_N &= (\psi_1, \psi_2, \ldots, \psi_N)^T \\ \psi_k &= (\psi_{k,1}, \psi_{k,2}, \ldots, \psi_{k,s+p+1})^T \end{aligned} \tag{24}$$

such that the following two properties hold

(a) $P\lim_{N\to\infty}((1/N)\Psi_N^T \Theta_N)$ exists and is not singular
(b) $P\lim_{N\to\infty}((1/N)\Psi_N^T Z_N) = 0$

for each $a_1, a_2, \ldots, a_p$, with $Z_N = (\bar{z}_1, \bar{z}_2, \ldots, \bar{z}_N)^T$ and $\bar{z}_k = z_k - a_1 z_{k-1} - \cdots - a_p z_{k-p}$ as given above.

Remark 4: Observe that

$$Z_N = \left[Z_N^{(0)}, Z_N^{(1)}, \ldots, Z_N^{(p)}\right] \cdot \begin{bmatrix} 1 \\ -a_1 \\ \vdots \\ -a_p \end{bmatrix}$$

where

$$Z_N^{(r)} = (z_{1-r}, z_{2-r}, \ldots, z_{N-r})^T, \quad r = 0, 1, \ldots, p \tag{25}$$

and by virtue of (b) it holds that

$$\left[\frac{1}{N}\Psi_N^T Z_N^{(0)}, \frac{1}{N}\Psi_N^T Z_N^{(1)}, \ldots, \frac{1}{N}\Psi_N^T Z_N^{(p)}\right] \cdot \begin{bmatrix} 1 \\ -a_1 \\ \vdots \\ -a_p \end{bmatrix} \to 0$$

in probability as $N \to \infty$, for each $a_1, a_2, \ldots, a_p$. Hence, by lying $a_1 = a_2 = \cdots = a_p = 0$, we obtain in particular that

$$\frac{1}{N}\Psi_N^T Z_N^{(0)} \to 0, \quad \text{in probability as } N \to \infty$$

which implies in consequence that it holds

$$\begin{aligned} &(-a_1)\frac{1}{N}\Psi_N^T Z_N^{(1)} + \cdots + \\ &(-a_p)\frac{1}{N}\Psi_N^T Z_N^{(p)} \to 0 \quad \text{in probability as } N \to \infty \end{aligned}$$

for arbitrary $a_1, a_2, \ldots, a_p$. Consequently, by zeroing $a_j \neq a_r$ and putting $a_r = -1$, we ascertain that

$$\frac{1}{N}\Psi_N^T Z_N^{(r)} \to 0, \quad \text{in probability as } N \to \infty \tag{26}$$

for each $r = 1, 2, \ldots, p$.

Under such conditions the estimation error

$$\Delta_N^{(IV)} = \widehat{\theta}_N^{(IV)} - \theta = \left(\frac{1}{N}\Psi_N^T \Theta_N\right)^{-1} \left(\frac{1}{N}\Psi_N^T Z_N\right) \tag{27}$$

tends to zero (in probability) as $N \to \infty$, i.e. $\widehat{\theta}_N^{(IV)} \to \theta$ in probability as N grows large.

The conditions (a) and (b) require in fact that the elements of Ψ_N be correlated with inputs and simultaneously not correlated with the noise $\{\bar{z}_k\}$. The simplest Ψ_N-generation techniques exploit directly former inputs of linear dynamics (see e.g. Stoica and Söderström 1982a), i.e. we take

$$\psi_{k,i} = w_{k-i+1}$$

which yields

$$\psi_k = (w_k, \ldots, w_{k-s}, w_{k-s-1}, \ldots, w_{k-s-p})^T \tag{28}$$

Other, more sophisticated Ψ_N-generation methods are described in e.g. Söderström and Stoica (1983, 1989); Stoica and Söderström (2002). For clarity of exposition, we shall further confine to the estimate (23) with the instruments (28).

4.2 *Non-parametric instrumental variables*

As it was already pointed out, in the Hammerstein system the needed inputs $w_k, \ldots, w_{k-s-p}$ of the linear dynamics, $w_k = \mu(u_k)$, are not accessible for measurements thus precluding the direct use of the instrumental variables estimate (23), (28). According to our idea, to overcome this drawback we propose to compute appropriate estimates $\widehat{w}_{k,M}$ by a non-parametric technique (in Stage 1; see (15) and example in § 3), similarly as for a static element, and use them instead of unknown w_ks, here for $1-(s+p) \le k \le N$, obtaining the following plug in estimate:

$$\widehat{\theta}_{N,M}^{(IV)} = \left(\widehat{\Psi}_{N,M}^T \widehat{\Theta}_{N,M}\right)^{-1} \widehat{\Psi}_{N,M}^T Y_N \tag{29}$$

where

$$\begin{aligned} \widehat{\Theta}_{N,M} &= (\widehat{\vartheta}_{1,M}, \ldots, \widehat{\vartheta}_{N,M})^T \\ \widehat{\vartheta}_{k,M} &= (\widehat{w}_{k,M}, \ldots, \widehat{w}_{k-s,M}, y_{k-1}, \ldots, y_{k-p})^T \end{aligned} \tag{30}$$

$$\begin{aligned} \widehat{\Psi}_{N,M} &= (\widehat{\psi}_{1,M}, \ldots, \widehat{\psi}_{N,M})^T \\ \widehat{\psi}_{k,M} &= (\widehat{w}_{k,M}, \ldots, \widehat{w}_{k-s,M}, \widehat{w}_{k-s-1,M}, \ldots, \widehat{w}_{k-s-p,M})^T \end{aligned} \tag{31}$$

Remark 5: The estimation points (treated as fixed points in Stage 1)

$$\begin{aligned} &u_1, \ldots, u_{1-s}, u_{1-s-1}, \ldots, u_{1-s-p} \\ &u_2, \ldots, u_{2-s}, u_{2-s-1}, \ldots, u_{2-s-p} \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ &u_N, \ldots, u_{N-s}, u_{N-s-1}, \ldots, u_{N-s-p} \end{aligned}$$

i.e.

$$u_{1-(s+p)}, u_{2-(s+p)}, \ldots, u_0; u_1, u_2, \ldots, u_N$$

for which $w_{k-r} = \mu(u_{k-r}) = \mu(u_{k-r}, c^*)$ have to be estimated are generated at random, by the input law, and $\{(u_{k-r}, y_{k-r});\ k = 1, 2, \ldots, N$ and $r = 0, 1, \ldots, s+p\}$ induce the 'data' $\{(w_{k-r}, y_{k-r});\ k = 1, 2, \ldots, N$ and $r = 0, 1, \ldots, s+p\}$ required for the identification of linear dynamics. Further, $\{(u_k, y_k)\}_{k=1}^N$ (corresponding to $r = 0$) constitute first part of the data collection $\{(u_k, y_k)\}_{k=1}^M$, $M \gg N$, used for non-parametric estimation of w_{k-r}s in Stage 1. The time instant index $k = 1$ is, by assumption, attributed to the beginning of recording the input-output data employed to this end.

Such a strategy yields a consistent estimate $\widehat{\theta}_{N,M}^{(IV)}$ of θ (see Theorem 2 and 3 and proof in Appendix 1). It is a well-known fact in the linear system identification theory that instrumental variables (28) guarantee fulfillment of (b) and of the non-singularity condition (a) under whiteness of the signal $\{w_k\}$. In this context along with Assumption 1, the non-parametric estimation of interactions $w_k (= \mu(u_k))$ and next application of the obtained estimates in (29) seems to be justified and rather natural. In contrast to the method presented in Stoica and Söderström (1982a) where the instruments were constructed by the specific linear filtering of the overall system input process $\{u_k\}$, our transformation of the inputs $\{u_k\}$ is more direct though non-linear (see e.g. (20)). Referring to (27) and the conditions (a) and (b) generally imposed on the instruments Ψ_N, for $\widehat{\Theta}_{N,M}$ and $\widehat{\Psi}_{N,M}$ given in (30), (31) and appearing in (29) we can show that the following theorems hold.

Theorem 2: *If the non-parametric estimate* $\widehat{R}_M(u)$ *is bounded, converges pointwise to the regression function* $R(u)$ *and at the estimation points* $u \in \{0, u_{k-r};$ *for* $k = 1, 2, \ldots, N$ *and* $r = 0, 1, \ldots, s+p\}$ (*see* (15), (30), (31)) *the error behaves like*

$$\left|\widehat{R}_M(u) - R(u)\right| = O(M^{-\tau}) \text{ in probability} \tag{32}$$

then

(a′) $P\lim_{M,N\to\infty}((1/N)\widehat{\Psi}_{N,M}^T \widehat{\Theta}_{N,M})$ *does exist and is not singular*

(b′) $P\lim_{M,N\to\infty}((1/N)\widehat{\Psi}_{N,M}^T Z_N) = 0$

provided that $NM^{-\tau} \to 0$.

Proof: See Appendix 1.

Theorem 3: *Under assumptions of Theorem* 2, *for the estimate* (29) *with* $\widehat{\Theta}_{N,M}$ *and the instruments* $\widehat{\Psi}_{N,M}$ *as in* (30) *and* (31) *it holds that*

$$\widehat{\theta}_{N,M}^{(IV)} \to \theta \text{ in probability} \tag{33}$$

as $N, M \to \infty$, *provided that* $NM^{-\tau} \to 0$. *Particularly, for* $M \sim N^{(1+\alpha)/\tau}$, $\alpha > 0$, *the asymptotic rate of convergence is*

$$\left\|\widehat{\theta}_{N,M}^{(IV)} - \theta\right\| = O\left(N^{-\min\left(\frac{1}{2}, \alpha\right)}\right) \text{ in probability} \tag{34}$$

Proof: Theorem may be proved in a similar manner as Theorems 3 and 4 in Hasiewicz and Mzyk (2004), with obvious substitutions, and hence the proof is here omitted.

We observe that since in (32) in general we have $0 < \tau < 1/2$, hence to fulfill the conditions (a′) and (b′) far more input-output data, M, must be used in Stage 1 for non-parametric estimation of interactions $\{w_k\}$ than the inputs and instruments, N, for computation of the estimate $\widehat{\theta}_{N,M}^{(IV)}$ in Stage 2. This can be explained by the necessity of proper reduction of estimation errors of w_k and slower convergence of non-parametric estimates. For $M \sim N^{(1+\alpha)/\tau}$, any $\alpha > 0$, we get consistent estimate $\widehat{\theta}_{N,M}^{(IV)}$ (see (33)) and for $\alpha \geqslant 1/2$ the estimate attains (w.r.t. N) the best possible parametric rate of convergence $O(N^{-1/2})$ in probability (see (34)).

4.3 *Selection of optimal instruments*

Since the accuracy (27) of the IV method obviously depends on the employed instruments, the natural question is how to choose the instruments in an optimal manner. We shall shortly discuss this issue below and provide a suboptimal 'practical' rule for selection of the best (in some sense) instrumental variables for the Hammerstein system. The rule is based on non-parametric estimates of the interactions $\{w_k\}$ and of the noise-free outputs $\{v_k\}$ (see Figure 3). Denote

$$\Gamma_N \triangleq \left(\frac{1}{N}\Psi_N^T\Theta_N\right)^{-1}\frac{1}{\sqrt{N}}\Psi_N^T$$
$$Z_N^* \triangleq \frac{\frac{1}{\sqrt{N}}Z_N}{\bar{z}_{\max}}$$

where $\bar{z}_{\max}$ is an upper bound on the absolute value of the noise $\{\bar{z}_k\}$ (see (21) and Assumption 4). The estimation error (27) can be clearly rewritten as

$$\Delta_N^{(IV)} = \bar{z}_{\max}\Gamma_N Z_N^* \tag{35}$$

Note that the[1] Euclidean norm $\|\cdot\|_2$ of the normalised noise vector Z_N^* is now ≤ 1:

$$\left\|Z_N^*\right\|_2 = \sqrt{\sum_{k=1}^{N}\left(\frac{\frac{1}{\sqrt{N}}\bar{z}_k}{\bar{z}_{\max}}\right)^2} = \sqrt{\frac{1}{N}\sum_{k=1}^{N}\left(\frac{\bar{z}_k}{\bar{z}_{\max}}\right)^2} \leq 1$$

Let us apply the following instruments' quality index

$$Q(\Psi_N) = \max_{|Z_N^*|_2\leq 1}\left|\Delta_N^{(IV)}(\Psi_N)\right|_2^2 \tag{36}$$

We write here $\Delta_N^{(IV)}(\Psi_N)$ to emphasise the dependence of the error $\Delta_N^{(IV)}$ on the choice of Ψ_N, i.e. the instruments ψ_k. The following theorem can be proved.

Theorem 4: *For Hammerstein systems, the index* $Q(\Psi_N)$ *is asymptotically optimal for the instrumental matrix*

$$\Psi_N^* = \left(\psi_1^*, \psi_2^*, \ldots, \psi_N^*\right)^T \text{ with}$$
$$\psi_k^* = (w_k, w_{k-1}, \ldots, w_{k-s}, v_{k-1}, v_{k-2}, \ldots, v_{k-p})^T \tag{37}$$

where w_k, $w_{k-1}, \ldots, w_{k-s}$ *are interactions and* v_{k-1}, $v_{k-2}, \ldots, v_{k-p}$ *are noise-free outputs of the system (Figure 3), i.e. for* Ψ_N^* *as in* (37) *and all other admissible choices of* Ψ_N *as in* (24) *it holds that*

$$\lim_{N\to\infty} Q\left(\Psi_N^*\right) \leqslant \lim_{N\to\infty} Q(\Psi_N) \quad \text{with probability 1} \tag{38}$$

Proof: See Appendix 2.

Theorem 4 is only of theoretical value because of inaccessibility in the system of $\{w_k\}$ and $\{v_k\}$. However it provides a guideline concerning the best choice of instruments, which can be used as a starting point for setting up a 'practical' routine for synthesis of the instruments. Namely, using a non-parametric technique (particularly, non-parametric estimates $\widehat{w}_k$ of w_k) and taking account of the form of the optimal instruments (37), we can propose a natural scheme where instrumental variable vectors are computed as

$$\widehat{\psi}_{k,M}^* = (\widehat{w}_{k,M}, \widehat{w}_{k-1,M}, \ldots, \widehat{w}_{k-s,M}, \hat{v}_{k-1,M}, \hat{v}_{k-2,M}, \ldots, \hat{v}_{k-p,M})^T \tag{39}$$

where $\widehat{w}_{k,M}$s are non-parametric estimates (15) of w_ks and $\hat{v}_{k,M}$s are non-parametric estimates of v_ks calculated as

$$\hat{v}_{k,M} = \sum_{i=0}^{F}\widehat{\gamma}_{i,M}\widehat{w}_{k-i,M}$$

with (Greblicki and Pawlak 1986)

$$\widehat{\gamma}_{i,M} = \widehat{\varkappa}_{i,M}/\widehat{\varkappa}_{0,M},$$
$$\widehat{\varkappa}_{i,M} = \frac{1}{M}\sum_{k=1}^{M-i}(y_{k+i} - \bar{y})(u_k - \overline{u}),$$
$$\bar{y} = \frac{1}{M}\sum_{k=1}^{M} y_k, \quad \overline{u} = \frac{1}{M}\sum_{k=1}^{M} u_k$$

and F being a chosen 'cut-off level' of the infinite length impulse response $\{\gamma_i\}$ of the linear dynamics in the Hammerstein system (cf (2)). The simulations presented in § 5 confirm efficiency of such a scheme; however comprehensive theoretical analysis, in particular examination of the influence of the selection of the F value on the asymptotic behaviour of the estimate $\widehat{\theta}_{N,M}^{(IV)}$ with $\widehat{\Psi}_{N,M} = \widehat{\Psi}_{N,M}^*$ $(= (\widehat{\psi}_{1,M}^*, \widehat{\psi}_{2,M}^*, \ldots, \widehat{\psi}_{N,M}^*)^T)$ is an open question and is left for future research.

5. Simulation examples

5.1 *Identification of the Hammerstein system by the use of the Levenberg-Marquardt method and the kernel non-parametric instrumental variables*

The Hammerstein system composed of the following blocks was examined:

(a) static non-linearity (in a parametric form)

$$\mu(u,c) = c_1u + c_2u^2 + c_3u^3 + e^{c_4u} \overset{c=c^*}{=} u + u^2 - u^3 + e^u$$

with the unknown true parameter vector $c = c^* = (1, 1, -1, 1)^T$, and

(b) linear dynamics (in a parametric form)

$$v_k = b_0 w_k + b_1 w_{k-1} + a_1 v_{k-1} + a_2 v_{k-2}$$

governed by the equation

$$v_k = w_k + w_{k-1} + 0.5 v_{k-1} + 0.25 v_{k-2}$$

with the unknown vector of true parameters $\theta = (1, 1, 0.5, 0.25)^T$.

The system was excited by an i.i.d. uniformly distributed input signal $u_k \sim U(-3, 3)$, and disturbed by the correlated ARMA noise $z_k = 0.7 z_{k-1} + \varepsilon_k + \varepsilon_{k-1}$, where $\varepsilon_k \sim U(-\delta, \delta)$ and magnitude δ was changed and set as $\delta = 0.1$, 0.2 and 0.3.

In the identification step of the non-linear part, it was also assumed we know in advance that $\mu(0) = 1$, i.e. that the value of static characteristic $\mu(u, c^*) = \mu(u)$ is known at the point $u_0 = 0$ (see Assumption 5 in § 2). Owing to § 3, particularly to the relation $R(u) = \mu(u) + d$ (for $\gamma_0 = 1$), this requires a slight modification of the w_ns estimation routine (15), namely its revision to the form

$$\widehat{w}_{n,M} = \widehat{R}_M(\overline{u}_n) - \widehat{R}_M(0) + \mu(0)$$

i.e. to estimate the w_ns we actually need to use in Stage 1 the rule $\widehat{w}_{n,M} = \widehat{R}_M(\overline{u}_n) - \widehat{R}_M(0) + 1$ instead of (15).

Since dim $c = 4$, in Stage 1 of our procedure, according to the discussion in § 3 (see Remark 2), we choose for simplicity $N_0 = 4$. The kernel estimate described in Example in § 3, with the compactly supported window kernel:

$$K(x) = \begin{cases} \dfrac{1}{2}, & |x| < 1 \\ 0, & |x| \geqslant 1 \end{cases},$$

and $h(M) = 0.44 M^{-1/5}$ computed according to the rule recommended in Greblicki and Pawlak 1994 (see § 8, p. 145). In turn, in Stage 2, Levenberg-Marquardt methodology (sketched out in § 3) has been applied for minimisation of the appropriate loss function of the form (cf (16))

$$\widehat{Q}_{N_0,M}(c) = \sum_{n=1}^{4} \left[c_1 \overline{u}_n + c_2 \overline{u}_n^2 + c_3 \overline{u}_n^3 + e^{c_4 \overline{u}_n} - \widehat{w}_{n,M} \right]^2 = \sum_{n=1}^{4} r_n^2(c)$$

being the empirical version of the index (10), with the Jacobian matrix $J(\overline{u}_1, \overline{u}_2, \ldots, \overline{u}_{N_0}; c)$ (see Remark 2), denoted for shortness as $J(c)$, of the form

$$J(c) = \begin{bmatrix} \dfrac{\partial \mu(\overline{u}_1, c)}{\partial c_1} & \dfrac{\partial \mu(\overline{u}_1, c)}{\partial c_2} & \dfrac{\partial \mu(\overline{u}_1, c)}{\partial c_3} & \dfrac{\partial \mu(\overline{u}_1, c)}{\partial c_4} \\ \dfrac{\partial \mu(\overline{u}_2, c)}{\partial c_1} & \dfrac{\partial \mu(\overline{u}_2, c)}{\partial c_2} & \dfrac{\partial \mu(\overline{u}_2, c)}{\partial c_3} & \dfrac{\partial \mu(\overline{u}_2, c)}{\partial c_4} \\ \dfrac{\partial \mu(\overline{u}_3, c)}{\partial c_1} & \dfrac{\partial \mu(\overline{u}_3, c)}{\partial c_2} & \dfrac{\partial \mu(\overline{u}_3, c)}{\partial c_3} & \dfrac{\partial \mu(\overline{u}_3, c)}{\partial c_4} \\ \dfrac{\partial \mu(\overline{u}_4, c)}{\partial c_1} & \dfrac{\partial \mu(\overline{u}_4, c)}{\partial c_2} & \dfrac{\partial \mu(\overline{u}_4, c)}{\partial c_3} & \dfrac{\partial \mu(\overline{u}_4, c)}{\partial c_4} \end{bmatrix}$$

$$= \begin{bmatrix} \dfrac{\partial r_1}{\partial c_1} & \dfrac{\partial r_1}{\partial c_2} & \dfrac{\partial r_1}{\partial c_3} & \dfrac{\partial r_1}{\partial c_4} \\ \dfrac{\partial r_2}{\partial c_1} & \dfrac{\partial r_2}{\partial c_2} & \dfrac{\partial r_2}{\partial c_3} & \dfrac{\partial r_2}{\partial c_4} \\ \dfrac{\partial r_3}{\partial c_1} & \dfrac{\partial r_3}{\partial c_2} & \dfrac{\partial r_3}{\partial c_3} & \dfrac{\partial r_3}{\partial c_4} \\ \dfrac{\partial r_4}{\partial c_1} & \dfrac{\partial r_4}{\partial c_2} & \dfrac{\partial r_4}{\partial c_3} & \dfrac{\partial r_4}{\partial c_4} \end{bmatrix}$$

$$= \begin{bmatrix} \overline{u}_1 & \overline{u}_1^2 & \overline{u}_1^3 & \overline{u}_1 e^{\overline{u}_1 c_4} \\ \overline{u}_2 & \overline{u}_2^2 & \overline{u}_2^3 & \overline{u}_2 e^{\overline{u}_2 c_4} \\ \overline{u}_3 & \overline{u}_3^2 & \overline{u}_3^3 & \overline{u}_3 e^{\overline{u}_3 c_4} \\ \overline{u}_4 & \overline{u}_4^2 & \overline{u}_4^3 & \overline{u}_4 e^{\overline{u}_4 c_4} \end{bmatrix}$$

which is of full rank, i.e. $\det(J^T(c) J(c)) > 0$, for each $c_4 \neq 0$, and as far as the input points $\overline{u}_1, \ldots, \overline{u}_4$ (freely selected by experimenter) are non-zero and different. In the experiment we set $\overline{u}_1 = -2$, $\overline{u}_2 = -1$, $\overline{u}_3 = 1$ and $\overline{u}_4 = 2$ for computing the non-parametric pointwise estimates $\widehat{w}_{1,M}, \ldots, \widehat{w}_{4,M}$ of the unknown $w_1 = \mu(\overline{u}_1, c^*), \ldots, w_4 = \mu(\overline{u}_4, c^*)$. For $\lambda_0 = 1/1024$ and $\nu = 8$, after 100 iterations, for each M ranging over $100, \ldots, 1000$ and $\delta = 0.1$, 0.2, 0.3, we obtained the estimation errors $\Delta_c(M) = (|\hat{c}_{N_0,M} - c^*|_2 / \|c^*\|_2) \cdot 100\%$ shown in Figure 4. The plots illustrate rather small dependence of the estimation accuracy on the number of data M used for non-parametric estimation of interaction inputs w_n in Stage 1 for each examined intensity of noise δ when $M > 800$, and show in particular that $\Delta_c(M) < 5\%$ is guaranteed for each δ if $M > 400$ observations.

The error of the estimate (23) of the linear dynamics parameters θ was computed for various strategies of the choice of instrumental variables and $R = 10$ independent trials for each strategy. For each instrumental variables choice, we assumed that $M = bN^{2.75}$ with the scale factor $b = 1.54 \cdot 10^{-4}$ (see Theorem 3) using the rule proposed in Hasiewicz

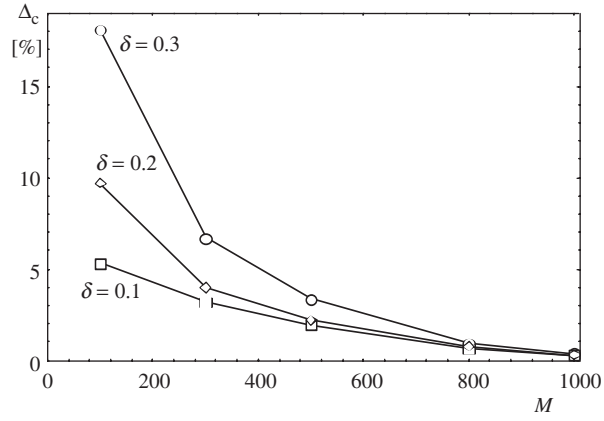

Figure 4. Estimation error for Levenberg-Marquardt method.

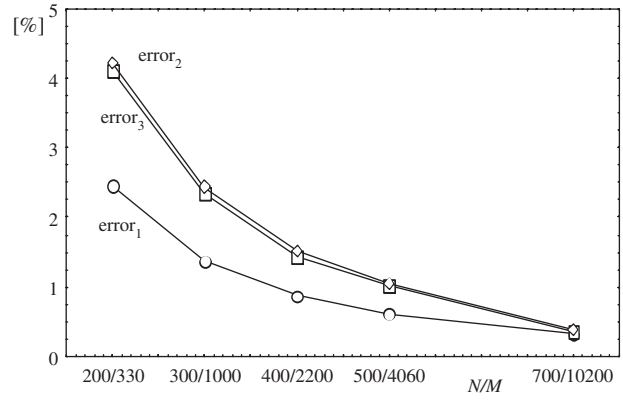

Figure 5. Linear dynamics estimation error *versus* number of measurements.

and Mzyk 2004 (see §6, p. 1373) and changed the data length M over $330, \ldots, 10200$. This yielded the number of instruments N (see (24)) ranging, respectively, from 200 to 700, and the ratio N/M as shown in Figure 5. First, the perfect optimal instruments Ψ_N^* (37) were used under working hypothesis of accessibility of signals $\{w_k\}$ and $\{v_k\}$ in the system. The computed $\text{error}_1(N) = \max_{r=1,\ldots,R}(\|\widehat{\theta}_N^{(r)}(\Psi_N^{*(r)}) - \theta\|_2/\|\theta\|_2) \cdot 100\%$ may be clearly treated as an empirical lower bound of the estimation errors for other choices of instrumental variables (see (38)). Next, the strategy of non-parametric generation of instruments (31) has been applied using kernel estimates of w_ks (of the same form as for the static element), and the $\text{error}_2(N) = \max_{r=1,\ldots,R}(\|\widehat{\theta}_{N,M}^{(r)}(\widehat{\Psi}_{N,M}^{(r)}) - \theta\|_2/\|\theta\|_2) \cdot 100\%$ was computed. Finally, the sub-optimal instruments given by (39) have been applied with similar estimates of w_ks, and the impulse response cut-off level $F = 4$, yielding the $\text{error}_3(N) = \max_{r=1,\ldots,R}(\|\widehat{\theta}_{N,M}^{(r)}(\widehat{\Psi}_{N,M}^{*(r)}) - \theta\|_2/\|\theta\|_2) \cdot 100\%$. The above three errors are shown in Figure 5. According to our expectation, the best – though not realisable – is the choice of $\Psi_N = \Psi_N^*$. The 'empirical' strategies $\widehat{\Psi}_{N,M}$ and $\widehat{\Psi}_{N,M}^*$ are worse, and in the sense of the relative error used in the experiment rather indistinguishable, but the same level of estimation accuracy as for the static system, i.e. not $<5\%$, is achieved already for $M = 330$ input-output data and $N = 200$ instruments. Moreover, the difference between the relative estimation errors for the proposed 'empirical' and optimal instruments is in general not $>1.5\%$. Hence, in our example, all strategies of instruments selection are in fact comparable, which in particular justifies the proposed strategies based on non-parametric estimation of instruments.

5.2 *Non-linearity recovering of the Hammerstein system under incorrect* **a priori** *knowledge of the linear dynamics*

In the second experiment, we compared efficiency of our two-stage routine with the instrumental variables approach proposed by Stoica and Söderström (1982a) (called further SS algorithm) in the case when *a priori* knowledge about the structure of the linear dynamics is inaccurate. In the Hammerstein system we assumed the linear-in-the-parameters static non-linearity (as needed for comparison purposes)

$$\mu(u, c) = c_1 u + c_2 u^2 + c_3 u^3 \stackrel{c \equiv c^*}{=} u + u^2 - u^3$$

with the unknown parameter vector $c = c^* = (1, 1, -1)^T$ and the dynamic part of the form

$$v_k = 0.5 v_{k-1} + 0.25 v_{k-1} + w_k + w_{k-1}$$

but took for the identification purposes the incorrect parametric model of the linear dynamics in the form

$$v_k = a_1 v_{k-1} + b_0 w_k$$

The i.i.d. system input and output noise were generated as $u_k \sim U(-10, 10)$ and $\varepsilon_k \sim U(-1, 1)$, and the aim was only to estimate true parameters c^* of the non-linearity $\mu(u, c)$. To compare efficiency of our two stage identification algorithm with the SS instrumental variables given in Stoica and Söderström (1982a) the overall input-output parametric model of Hammerstein system needed in Stoica and Söderström (1982a) has been computed

$$v_k^{(m)} = A_1 v_{k-1} + B_{01} u_k + B_{02} u_k^2 + B_{03} u_k^3 \triangleq \phi_k^{(ss)T} p$$

where

$$\phi_k^{(ss)} = \left(v_{k-1}, u_k, u_k^2, u_k^3\right)^T \quad \text{and} \quad p = (p_1, p_2, p_3, p_4)^T$$

with

$$p_1 = A_1 = a_1 \quad p_2 = B_{01} = b_0 c_1$$
$$p_3 = B_{02} = b_0 c_2 \quad p_4 = B_{03} = b_0 c_3$$

and implemented in the IV estimate described in Stoica and Söderström (1982a):

$$\hat{p}_N^{(ss)} = \left(\Psi_N^{(ss)T}\Theta_N\right)^{-1}\Psi_N^{(ss)T}Y_N \tag{40}$$

where $\Psi_N^{(ss)} = (\psi_1^{(ss)}, \psi_2^{(ss)}, \ldots, \psi_N^{(ss)})^T$ and the instruments are $\psi_k^{(ss)} = (u_{k-1}, u_k, u_k^2, u_k^3)^T$. In turn, owing to the particular linear-in-the-parameters form of the non-linearity, according to the discussion in Hasiewicz and Mzyk (2004) (see Remark 2 there), for performing the non-parametric stage in our two-stage procedure (Stage 1) we set $N_0 = 10 \geqslant \dim c = 3$, and arbitrarily take the input points $\overline{u}_i = -9 + 2(i-1)$ forming an equidistant grid on the input domain $(-10, 10)$, and moreover guaranteeing fulfillment of the condition $\operatorname{rank}\Phi_{N_0} = \dim c$ where $\Phi_{N_0} = (\phi_1, \phi_2, \ldots, \phi_{N_0})^T$, $\phi_k = (\overline{u}_k, \overline{u}_k^2, \overline{u}_k^3)^T$. This is because for the static non-linearity as in the example in the parametric stage (Stage 2) the linear least squares can be used getting the estimate (Hasiewicz and Mzyk 2004)

$$\hat{c}_{N_0,N} = \left(\Phi_{N_0}^T\Phi_{N_0}\right)^{-1}\Phi_{N_0}^T\widehat{W}_{N_0,N}$$

where $\widehat{W}_{N_0,N} = (\widehat{w}_{1,N}, \widehat{w}_{2,N}, \ldots, \widehat{w}_{N_0,N})^T$ with the non-parametric kernel estimates $\widehat{w}_{k,N}$. For comparison purposes, in the kernel estimates of w_ns (see (15), (20)) we assumed $M = N$ (with N the same as for (40)) taking due to (Greblicki and Pawlak 1994) $h_{opt}(N) = 4.1N^{-0.2}$. Next we compared appropriate relative estimation errors of the non-linearity parameters produced by both methods:

$$\Delta_c = \frac{\left\|\hat{c}_{N_0,N} - c^*\right\|_2}{\|c^*\|_2}\cdot 100\% \quad \text{and}$$
$$\Delta_c^{(ss)} = \frac{\left\|\widehat{p'}_N^{(ss)} - c^*\right\|_2}{\|c^*\|_2}\cdot 100\%$$

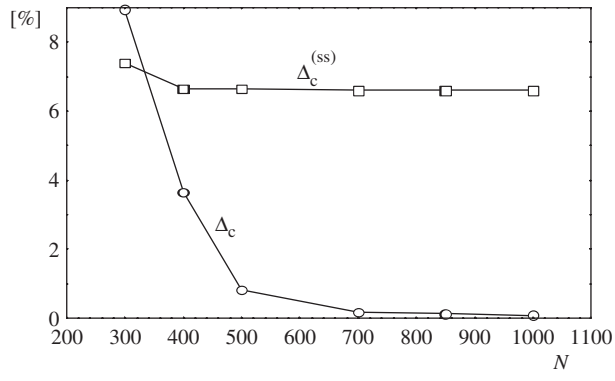

Figure 6. Comparison of the two-stage parametric–non-parametric and SS algorithm.

where $\widehat{p'}_N^{(ss)} = ((\hat{p}_{2,N}^{(ss)}/b_0), (\hat{p}_{3,N}^{(ss)}/b_0), (\hat{p}_{4,N}^{(ss)}/b_0))^T$ for growing data length N. The results are presented in Figure 6. We see that our combined parametric-non-parametric approach is here truly robust to the incorrect knowledge of the parametric description of the linear subsystem, in contrast to the SS algorithm in Stoica and Söderström (1982a) where the essential systematic estimation error (bias) appears (see Figure 6).

6. Conclusions

The two-stage parametric-non-parametric approach for the identification of Hammerstein systems by instrumental variables method proposed in the article reveals the following noteworthy properties: (1) small *a priori* knowledge about the random signals needed for the identification scheme to work and generality of the identification framework, as any particular model of the input and noise is not assumed; (2) broad applicability, as vast class of non-linear characteristics (not necessarily linear-in-the-parameters) and ARMA-type linear dynamics (with infinite impulse response) is admitted; (3) full decomposition of the Hammerstein system identification task, as identification of subsystems is performed in fact in a completely decentralised (local) manner; (4) robustness to the partial inaccuracy of *a priori* knowledge, as successful identification of one system component can be performed in spite of imprecise parametric information of the other; (5) computational simplicity, as parametric identification of subsystems can be performed independently (locally) by using standard identification routines with well elaborated software, and non-parametric stage needs only elementary computations. The disadvantage is that some deterioration of convergence speed of the parameter estimates, in comparison with the best possible parametric rate of convergence, can occur in the method. This is caused by slower convergence of non-parametric techniques employed in the first (preliminary) stage in the approach.

In the article, general conditions are provided for obtaining successful symbiosis of parametric and non-parametric identification methods, i.e. of getting hybrid parametric-non-parametric identification algorithms which guarantee achievement of consistent and effective estimates (cf Theorems 1–3) in a convenient way and at no great expense.

Note

1. From now on the Euclidean norm will be denoted by $\|\cdot\|_2$ to avoid ambiguity.

References

Bai, E.W. (2002), 'A Blind Approach to the Hammerstein-Wiener Model Identification', *Automatica*, 38, 967–979.

Bai, E.W., and Li, D. (2004), 'Convergence of the Iterative Hammerstein System Identification Algorithm', *IEEE Transactions on Automatic Control*, 49, 1929–1940.

Billings, S.A., and Fakhouri, S.Y. (1982), 'Identification of Systems Containing Linear Dynamic and Static Nonlinear Elements', *Automatica*, 18, 15–26.

Chang, F.H.I., and Luus, R. (1971), 'A Non-iterative Method for Identification Using Hammerstein Model', *IEEE Transactions on Automatic Control*, AC-16, 464–468.

Efromovich, S. (1999), *Nonparametric Curve Estimation*, New York: Springer.

Giannakis, G.B., and Serpedin, E. (2001), 'A Bibliography on Nonlinear System Identification', *Signal Processing*, 81, 533–580.

Greblicki, W. (1989), 'Nonparametric Orthogonal Series Identification of Hammerstein Systems', *International Journal of Systems Science*, 20, 2355–2367.

Greblicki, W., and Pawlak, M. (1986), 'Identification of Discrete Hammerstein Systems Using Kernel Regression Estimates', *IEEE Transactions on Automatic Control*, 31, 74–77.

——. (1994), 'Cascade Non-Linear System Identification by a Non-Parametric Method', *International Journal of System Science*, 25, 129–153.

Hannan, E.J., and Deistler, M. (1998), *The Statistical Theory of Linear Systems*, New York: J. Wiley & Sons.

Härdle, W. (1990), *Applied Nonparametric Regression*, Cambridge, UK: Cambridge University Press.

Hasiewicz, Z., and Mzyk, G. (2004), 'Combined Parametric-Nonparametric Identification of Hammerstein Systems', *IEEE Transactions on Automatic Control*, 49, 1370–1375.

Hasiewicz, Z., Pawlak, M., and Śliwiński, P. (2005), 'Nonparametric Identification of Nonlinearities in Block-Oriented Systems by Orthogonal Wavelets with Compact Support', *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, 52, 427–442.

Lang, Z.Q. (1993), 'Controller Design Oriented Model Identification Method for Hammerstein System', *Automatica*, 29, 767–771.

Moré, J.J. (1977), *The Levenberg-Marquardt Algorithm: Implementation and Theory Series of Lecture Notes in Mathematics*, New York: Springer-Verlag, pp. 105–116.

Mzyk, G. (2007), 'Generalised Kernel Regression Estimate for the Identification of Hammerstein Systems', *International Journal of Applied Mathematics and Computer Science*, 17, 189–197.

Narendra, K.S., and Gallman, P.G. (1966), 'An Iterative Method for the Identification of Nonlinear Systems Using the Hammerstein Model', *IEEE Transactions on Automatic Control*, AC-11, 546–550.

Pawlak, M., and Hasiewicz, Z. (1998), 'Nonlinear System Identification by the Haar Multiresolution Analysis', *IEEE Transactions on Circuits and Systems*, 45, 945–961.

Press, W.H., Teukolsky, S.A., Vetterling, W., and Flannery, B. (1992), *Numerical Recipes in C* (2nd ed.), Cambridge, UK : Cambridge University Press.

Rao, C.R. (1973), *Linear Statistical Inference and its Applications* (2nd ed.), New York: John Wiley & Sons.

Ruppert, D., Wand, M.P., and Carroll, R.J. (2003), *Semiparametric Regression*, New York: Cambridge University Press.

Seber, G.A.F., and Wild, C.J. (1989), *Nonlinear Regression Series of Probability and Mathematical Statistics*, New York: Wiley.

Söderström, T., and Stoica, P. (1983), *Instrumental Variable Methods for System Identification*, New York: Springer Verlag.

——. (1989), *System Identification*, Englewood Cliffs, NJ: Prentice Hall.

Stoica, P., and Söderström, T. (1982a), 'Instrumental-Variable Methods for Identification of Hammerstein Systems', *International Journal of Control*, 35, 459–476.

——. (1982b), 'Comments on the Wong and Polak Minimax Approach to Accuracy Optimisation of Instrumental Variable Methods', *IEEE Transactions on Automatic Control*, 27, 1138–1139.

——. (2002), 'Instrumental Variable Methods for System Identification', *Circuits Systems Signal Processing*, 21, 1–9.

Stone, C.J. (1980), 'Optimal Rates of Convergence for Nonparametric Regression', *Annals of Statistics*, 8, 1348–1360.

Strang, G. (2003), *Introduction to Linear Algebra*, Wellesley: Wellesley-Cambridge Press.

Takezawa, K. (2006), *Introduction to Nonparametric Regression*, Hoboken: Wiley.

van der Vaart, A.W. (1988), *Asymptotic Statistics*, Cambridge: Cambridge University Press.

Vapnik, V. (1982), *Estimation of Dependences Based on Empirical Data*, Berlin: Springer.

Wand, M.P., and Jones, H.C. (1995), *Kernel Smoothing*, London: Chapman and Hall.

Ward, R. (1977), 'Notes on the Instrumental Variable Method', *IEEE Transactions on Automatic Control*, 22, 482–484.

Wong, K., and Polak, E. (1967), 'Identification of Linear Discrete Time Systems Using the Instrumental Variable Method', *IEEE Transactions on Automatic Control*, AC-12, 707–718.

Appendix 1. Proof of Theorem 2

(a′) Denoting $A_{N,M} = \widehat{\Psi}_{N,M} - \Psi_N$ and $B_{N,M} = \widehat{\Theta}_{N,M} - \Theta_N$ we have

$$\begin{aligned}\frac{1}{N}\widehat{\Psi}_{N,M}^T\widehat{\Theta}_{N,M} &= \frac{1}{N}(\Psi_N + A_{N,M})^T(\Theta_N + B_{N,M}) \\ &= \frac{1}{N}\Psi_N^T\Theta_N + \frac{1}{N}A_{N,M}^T\Theta_N + \frac{1}{N}\Psi_N^T B_{N,M} \\ &\quad + \frac{1}{N}A_{N,M}^T B_{N,M} \end{aligned} \tag{41}$$

Since in the case under consideration $(1/N)\Psi_N^T\Theta_N$ fulfils condition (a), to prove (a′) it suffices to show that, for example, the $\|\cdot\|_{1,1}$ norms of the remaining three components in (41), i.e.

$$\xi_{N,M}=\left\|\frac{1}{N}A_{N,M}^T\Theta_N\right\|_{1,1}\quad \chi_{N,M}=\left\|\frac{1}{N}\Psi_N^TB_{N,M}\right\|_{1,1}$$
$$\varkappa_{N,M}=\left\|\frac{1}{N}A_{N,M}^TB_{N,M}\right\|_{1,1}$$

tend to zero in probability as N, $M\to\infty$ and $NM^{-\tau}\to 0$ where $\|\cdot\|_{1,1}$ is the matrix norm induced by the 1-vector norm $\|x\|_1=\sum_{i=1}^{\dim x}|x[i]|$. This norm can be used in the proof because in the finite dimensional real space all vector norms are equivalent. To show such a property for $\xi_{N,M}$ let us notice that under assumptions as in § 2 all elements

$$\vartheta_{kl}=\begin{cases} w_{k-l+1}, & \text{for } l\leqslant s+1\\ y_{k-l+s+1}, & \text{for } l>s+1\end{cases}$$

of the matrix Θ_N are bounded in the absolute value, i.e. there exists $0<b<\infty$ such that $|\vartheta_{kl}|<b$ for each $k=1,2,\ldots,N$ and $l=1,2,\ldots,s+p+1$. Since further

$$\frac{1}{N}A_{N,M}^T\Theta_N=\frac{1}{N}\sum_{k=1}^{N}\begin{bmatrix}\widehat{w}_{k,M}-w_k\\ \widehat{w}_{k-1,M}-w_{k-1}\\ \ldots\\ \widehat{w}_{k-s-p,M}-w_{k-s-p}\end{bmatrix}[\vartheta_{k1},\vartheta_{k2},\ldots,\vartheta_{k,s+p+1}]$$

we obtain

$$\xi_{N,M}\leqslant b(s+p+1)\max_k|\widehat{w}_{k,M}-w_k|,\quad 1-(s+p)\leqslant k\leqslant N$$

and hence

$$P(\xi_{N,M}>\varepsilon)\leqslant\sum_{k=1-(s+p)}^{N}P\left(|\widehat{w}_{k,M}-w_k|>\frac{\varepsilon}{b(s+p+1)}\right)$$

Since by assumption $\widehat{w}_{k,M}$s are bounded and the rate of convergence of non-parametric estimate is given by (32), it holds that $E|\widehat{w}_{k,M}-w_k|=O(M^{-\tau})$, $k=1-(s+p),\ldots,N$. Hence, using Markov's inequality we have

$$P(\xi_{N,M}>\varepsilon)\leqslant\frac{C(s+p+1)}{\varepsilon}[NM^{-\tau}+(s+p)M^{-\tau}]\qquad(42)$$

where C is a constant. The right-hand side in (42) tends to zero as N, $M\to\infty$ provided that $NM^{-\tau}\to 0$. The convergence of $\chi_{N,M}\to 0$ and $\varkappa_{N,M}\to 0$ in probability can be proved in the same manner.

(b′) Including that in the case considered $(1/N)\Psi_N^TZ_N$ fulfils condition (b) and that

$$\frac{1}{N}\widehat{\Psi}_{N,M}^TZ_N=\frac{1}{N}\Psi_N^TZ_N+\frac{1}{N}A_{N,M}^TZ_N$$

with $A_{N,M}$ as in part (a′) above, the proof of (b′) can be obtained by exploiting boundedness of $\bar{z}_k$ (i.e. that $|\bar{z}_k|\leqslant(1+|a_1|+|a_2|+\cdots+|a_p|)z_{\max}$; cf Assumption 4) and following the steps of the proof in part (a′) with respect to $\varrho_{N,M}=\|(1/N)A_{N,M}^TZ_N\|_1$.

Appendix 2. Proof of Theorem 4

Let us put for convenience $\bar{z}_{\max}=1$. Taking account of (35) one can write

$$\left\|\Delta_N^{(IV)}(\Psi_N)\right\|_2^2=\Delta_N^{(IV)T}(\Psi_N)\Delta_N^{(IV)}(\Psi_N)=Z_N^{*T}\Gamma_N^T\Gamma_NZ_N^*$$

and hence the quality index (36) for N large enough (see § 4) is

$$Q(\Psi_N)=\max_{\|Z_N^*\|_2\leq 1}\left\langle Z_N^*,\Gamma_N^T\Gamma_NZ_N^*\right\rangle=\|\Gamma_N\|^2=\lambda_{\max}\left(\Gamma_N^T\Gamma_N\right)$$

where $\|\cdot\|$ is a spectral matrix-norm induced by the Euclidean vector-norm, and $\lambda_{\max}(A)$ is the greatest absolute eigenvalue of A. Since (Rao 1973; Chapter 1)

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

we conclude that

$$Q(\Psi_N)=\max_{\|\zeta\|_2\leq 1}\left\langle\zeta,\Gamma_N\Gamma_N^T\zeta\right\rangle$$
$$=\max_{\|\zeta\|_2\leq 1}\left\langle\zeta,\left(\frac{1}{N}\Psi_N^T\Theta_N\right)^{-1}\left(\frac{1}{N}\Psi_N^T\Psi_N\right)\left(\frac{1}{N}\Theta_N^T\Psi_N\right)^{-1}\zeta\right\rangle$$

where Ψ_N is an arbitrary instrumental matrix

$$\Psi_N=(\psi_1,\psi_2,\ldots,\psi_N)^T$$

of the form as in (24) fulfilling the conditions (a) and (b).

Since

$$\Theta_N=(\vartheta_1,\vartheta_2,\ldots,\vartheta_N)^T$$

where

$$\vartheta_k=(w_k,w_{k-1},\ldots,w_{k-s},y_{k-1},y_{k-2},\ldots,y_{k-p})^T$$

for $k=1,2,\ldots,N$, and

$$\Psi_N^*=\left(\psi_1^*,\psi_2^*,\ldots,\psi_N^*\right)^T$$

where

$$\psi_k^*=(w_k,w_{k-1},\ldots,w_{k-s},v_{k-1},v_{k-2},\ldots,v_{k-p})^T$$

for $k=1,2,\ldots,N$, and moreover (see § 4)

$$y_{k-r}=v_{k-r}+z_{k-r}\ ,\quad r=1,2,\ldots,p$$

one can write in fact

$$\vartheta_k=\psi_k^*+\tilde{z}_k$$

where $\tilde{z}_k$ is the following output noise vector

$$\tilde{z}_k=(0,0,\ldots,0,\ z_{k-1},z_{k-2},\ldots,z_{k-p})^T$$

for $k=1,2,\ldots,N$. This yields eventually

$$\Theta_N=\Psi_N^*+\widetilde{Z}_N$$

where

$$\widetilde{Z}_N=(\tilde{z}_1,\tilde{z}_2,\ldots,\tilde{z}_N)^T$$

and leads further to the relation

$$\frac{1}{N}\Psi_N^T\Theta_N=\frac{1}{N}\Psi_N^T\left(\Psi_N^*+\widetilde{Z}_N\right)=\frac{1}{N}\Psi_N^T\Psi_N^*+\frac{1}{N}\Psi_N^T\widetilde{Z}_N$$

i.e.

$$\frac{1}{N}\Psi_N^T\Theta_N - \frac{1}{N}\Psi_N^T\Psi_N^* = \frac{1}{N}\sum_{k=1}^{N}\Psi_N^T\widetilde{Z}_N$$

Owing to denotation (25) in Remark 4 we see that

$$\widetilde{Z}_N = \left[0, 0, \ldots, 0; \widetilde{Z}_N^{(1)}, \widetilde{Z}_N^{(2)}, \ldots, \widetilde{Z}_N^{(p)}\right]_{N\times[(s+1)+p]}$$

i.e.

$$\frac{1}{N}\Psi_N^T\widetilde{Z}_N = \left[0, 0, \ldots, 0; \frac{1}{N}\Psi_N^T\widetilde{Z}_N^{(1)}, \frac{1}{N}\Psi_N^T\widetilde{Z}_N^{(2)}, \ldots, \frac{1}{N}\Psi_N^T\widetilde{Z}_N^{(p)}\right]$$

and due to (26) in Remark 4 it holds that

$$\frac{1}{N}\Psi_N^T\widetilde{Z}_N \to 0 \text{ in probability as } N\to\infty$$

Hence

$$\frac{1}{N}\Psi_N^T\Theta_N - \frac{1}{N}\Psi_N^T\Psi_N^* \to 0 \text{ in probability as } N\to\infty$$

which means that for each, arbitrarily small, $\varepsilon > 0$ we have

$$P\left\{\left\|\frac{1}{N}\Psi_N^T\Theta_N - \frac{1}{N}\Psi_N^T\Psi_N^*\right\| \geqslant \varepsilon\right\} \to 0 \quad \text{as } N\to\infty$$

or that asymptotically, as $N\to\infty$, it holds that

$$P\left\{\left\|\frac{1}{N}\Psi_N^T\Theta_N - \frac{1}{N}\Psi_N^T\Psi_N^*\right\| \neq 0\right\} \simeq 0$$

yielding subsequently

$$\frac{1}{N}\Psi_N^T\Theta_N \simeq \frac{1}{N}\Psi_N^T\Psi_N^* \tag{43}$$

with probability 1.

Since for N large, with probability 1, $(1/N)\Psi_N^T\Psi_N^*$ can be applied instead of $(1/N)\Psi_N^T\Theta_N$, we obtain

$$Q(\Psi_N) = \max_{\|\zeta\|_2\leq 1}\left\langle \zeta, \left(\frac{1}{N}\Psi_N^T\Psi_N^*\right)^{-1}\left(\frac{1}{N}\Psi_N^T\Psi_N\right)\left(\frac{1}{N}\Psi_N^{*T}\Psi_N\right)^{-1}\zeta\right\rangle$$

with probability 1. Making use now of Lemma 6 in Appendix 3 for $M_1 = (1/\sqrt{N})\Psi_N^*$ and $M_2 = (1/\sqrt{N})\Psi_N$, and including (43) we get asymptotically

$$\zeta^T\Gamma_N\Gamma_N^T\zeta \geq \zeta^T\left(\frac{1}{N}\Psi_N^{*T}\Psi_N^*\right)^{-1}\zeta$$

with probability 1 for each vector ζ. Therefore

$$Q(\Psi_N) = \max_{\|\zeta\|_2\leq 1}\left(\zeta^T\Gamma_N\Gamma_N^T\zeta\right) \geq \max_{\|\zeta\|_2\leq 1}\left(\zeta^T\left(\frac{1}{N}\Psi_N^{*T}\Psi_N^*\right)^{-1}\zeta\right) \tag{44}$$

Since for $\Psi_N = \Psi_N^*$ both sides in (44) are equal, $Q(\Psi_N)$ attains minimum as $\Psi_N = \Psi_N^*$ □

Appendix 3. Technical lemmas

Lemma 5: *If λ_i, $i = 1, 2, \ldots, m$, are eigenvalues of the matrix A then the matrix $A - \delta I$, has the eigenvalues $\lambda_i - \delta$, $i = 1, 2, \ldots, m$.*

Proof: The conclusion follows immediately from the relation

$$\det([A - \delta I] - \lambda I) = \det(A - (\lambda + \delta)I)$$

□

Lemma 6: (Wong and Polak 1967) *Let M_1 and M_2 be matrices with the same dimensions. If there exist $(M_1^TM_1)^{-1}$, $(M_1^TM_2)^{-1}$ and $(M_2^TM_1)^{-1}$, then*

$$D_N = \left(M_2^TM_1\right)^{-1}M_2^TM_2\left(M_1^TM_2\right)^{-1} - \left(M_1^TM_1\right)^{-1}$$

is positive semidefinite, i.e. for each vector ζ it holds that

$$\zeta^TD_N\zeta \geq 0$$