Discarding Data May Help in System Identification

Pierre Carrette, Georges Bastin, Yves Y. Genin, *Fellow, IEEE*, and Michel Gevers, *Fellow, IEEE*

Abstract— We present results concerning the parameter estimates obtained by prediction error methods in the case of input signals that are insufficiently rich. Such input signals are typical of industrial measurements where occasional stepwise reference changes occur. As is intuitively obvious, the data located around the input signal discontinuities carry most of the useful information. Using singular value decomposition (SVD) techniques, we show that in noise undermodeling situations, the remaining data may introduce large bias on the model parameters with a possible increase of their total mean square error. A data selection criterion is then proposed to discard such poorly informative data to increase the accuracy of the transfer function estimate.

I. Introduction

THE aim of this paper is to analyze in detail the accuracy of the least squares (LS) prediction error method [3] for estimating system model parameters in situations where the system input signals exhibit only a few step discontinuities corresponding to changes in the reference signal (i.e., typical of industrial processes). More precisely, the system under study is assumed to be a single input single output (SISO) ARMAX system, whereas the model structure is chosen to be a SISO ARX model whose input to output dynamics are able to represent that of the true system exactly. To motivate the present study, consider the following ARMAX system:

$$(1 - 0.8z^{-1})y(t) = 0.5z^{-1}u(t) + (1 + 0.8z^{-1} + 0.3z^{-2})e(t) \quad (1)$$

and let us compute the parameter vector $\theta = [\theta_1, \theta_2]^T$ of the following ARX structure:

$$(1 + \theta_1 z^{-1})y(t) = \theta_2 z^{-1}u(t) + \varepsilon(t) \quad (2)$$

on the basis of a finite number N of input-output (I/O) data to obtain the best approximation of the actual system in the LS prediction error sense [3]. In (1) and (2), $u(t)$ and $y(t)$ stand, respectively, for the input and output signals, whereas $e(t)$ denotes a white noise disturbance, and $\varepsilon(t)$ is the modeling error.

Note that the model structure is able to represent the I/O system dynamics exactly but not the noise dynamics. The coefficients of the polynomials acting on $u(t)$ and $y(t)$ in (1) constitute the so-called true parameter vector: $\theta_0 := [-0.8, 0.5]^T$. We shall assume that our objective is to estimate the parameter vector θ_0 as accurately as possible using the model structure (2), i.e., in the presence of unmodeled noise dynamics. It should be stressed that the I/O and noise dynamics have a common denominator both in the true system and in the model structure.

Manuscript received December 28, 1994; revised February 9, 1996. This work was suported by the FRIA. The associate editor coordinating the review of this paper and approving it for publication was Dr. Zhi Ding.

The authors are with CESAME, Université Catholique de Louvain, Bâtiment Euler, B-1348 Louvain-la-Neuve, Belgium.

Publisher Item Identifier S 1053-587X(96)06679-2.

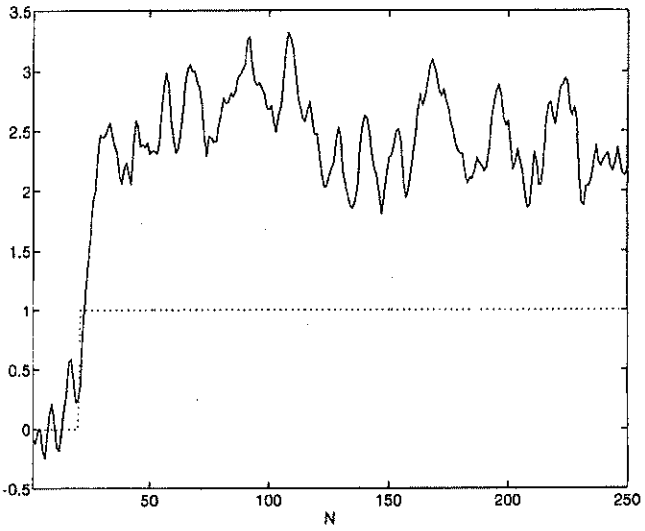


Fig. 1. Input ($\cdots$) and output ($-\!-$) signals of the ARMAX system.

The applied step input and the resulting system output signal are displayed in Fig. 1 in the case of a Gaussian white noise disturbance $e(t)$ with $N(0, 0.01)$ characteristics. The parameters $\hat{\theta}_1$ and $\hat{\theta}_2$ are estimated using a standard LS prediction error criterion with no data filtering and using data sequences of increasing length N. For each N, the bias (with respect to θ_0) and the variance have been estimated using 200 Monte Carlo simulations and are shown in Fig. 2. This figure shows that the variance decreases monotonically with the data length, whereas the bias is seen to be strongly influenced by the input signal: It reaches a minimum just after the step signal instant (which occurs at $N = 20$), and it increases significantly with N from there on. The reason for the bias increase is that in the absence of input excitation, the parameter fit focuses on the modeling of the noise dynamics. Since these cannot be modeled exactly within the given model structure, the parameters (in particular $\hat{\theta}_2$) tend to biased values that attempt to yield the best output predictor within the given model structure.

For each N, the total mean square error (MSE) of the parameters has been computed:

$$\mathrm{MSE}(N) := \sum_{i=1}^{2} [\mathrm{Bias}^2\{\hat{\theta}_i(N)\} + \mathrm{Var}\{\hat{\theta}_i(N)\}]. \quad (3)$$

In Fig. 3, the sum of the squared bias terms ($- \cdot$), the sum of the variance terms ($\cdots$), and the total MSE ($-\!-$) are represented as a function of the data length used in the

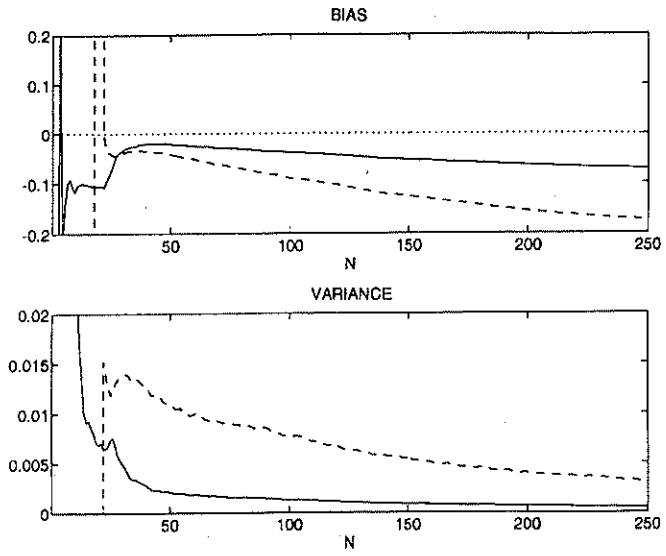


Fig. 2. Bias and variance of the parameters $\hat{\theta}_1$ (——) and $\hat{\theta}_2$ (– –) as a function of N.

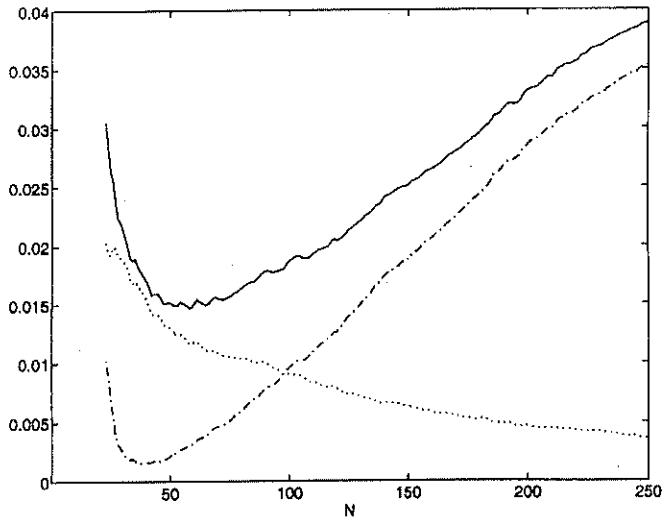


Fig. 3. Squared bias (– ·), variance (···) and mean square error (——) of the estimated parameters.

estimation. The most interesting feature of our simulation is that the total MSE presents a minimum around $N = 50$, i.e., some time after the input step instant. We have also computed, for each N, the frequency integrated total mean square error of the I/O transfer function:

$$\text{MSE}(G, N) := \int E\{|G_0(\omega) - G(\omega, \hat{\theta}(N))|^2\}\, d\omega$$

where $G(\omega, \hat{\theta}) = \hat{\theta}_2 z^{-1}/(1 + \hat{\theta}_1 z^{-1})$ with $z = e^{j\omega}$ and $G_0(\omega) = G(\omega, \theta_0)$ are the model and system I/O transfer functions, respectively. In Fig. 4, the squared bias (– ·) and the variance (···) contributions of the model I/O transfer function error and the total MSE(G, N) (——) are represented as a function of the data length used in the estimation. Obviously, these curves exhibit the same behavior as those in Fig. 3: The MSE of the parameters and of the transfer function estimate reaches a minimum around $N = 50$ and increases thereafter.

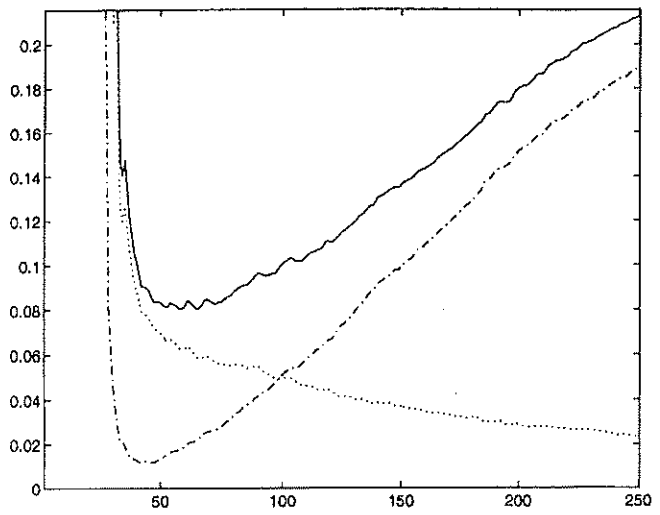


Fig. 4. Squared bias (– ·), variance (···) and mean square error (——) of the estimated I/O transfer function.

Thus, if the objective is the accuracy of the parameters of the I/O model or of its transfer function estimate, these simulations suggest that using more than, say, 50 data deteriorates the performance. Instead, one should stop the parameter estimation relatively early after the input step instant to prevent the increase in parameter bias from exceeding the decrease in their variance.

This example serves as a motivation. More generally, the present paper provides evidence that when the input data record is not very rich, it is better to focus on particular time intervals of data sets to identify the I/O part of unknown systems in situations where there is a common polynomial to the I/O and noise model description and where there is noise under modeling. The analysis performed in this paper is limited to ARX models because they lead to a theoretically tractable estimation problem; besides, they are very commonly used in system identification.

Our theoretical analysis can be summarized as follows. Using the singular value decomposition (SVD) of the model regressor matrix, we introduce the eigenparameters of the model. The estimates of these parameters are almost independent of one another, and they lead to the original model parameters *via* orthogonal transformations. The accuracy of these eigenparameters are expressed using the singular values of the model regressor matrix. Depending on the richness of the input data, these singular values may take widely different values. This implies very different accuracies of the eigenparameters. It turns out that the accuracy with which the actual parameters are estimated is determined by the accuracy of the most poorly estimated eigenparameter. Therefore, the accuracy of the model parameters as a function of the input data is essentially obtained by monitoring the effect of the input data on the bias and variance of the most poorly estimated eigenparameter. Our analysis will show that in noise undermodeling situations, the estimated model parameters of the I/O dynamics may show up large bias with respect to the corresponding system parameters if data located far away from the input step changes are used in the estimation. This is due to the common parameters in the I/O and noise

dynamics of the equation error model structures such as ARX, ARMAX, or ARARX. Given that the use of data located in time intervals where the input is not rich can deteriorate the MSE of the estimated parameters, we propose a data selection criterion that is based on the time evolution of the energy of the data associated with the most poorly estimated model eigenparameter.

The paper is organized as follows. In Section II, we present the system and the model structure as well as the data characteristics considered in the paper. In Section III, we introduce the parameter estimation method based on the minimization of the model prediction errors, and we solve this parameter estimation problem using the SVD of the model regressor matrix. This decomposition actually splits the estimation accuracy into well and poorly estimated eigenparameters. The statistical behavior of these parameters is analyzed in Section IV and linked to that of the original model parameters. In Section V, we describe in detail simulations of the parameter estimation procedure applied to the motivating example presented above. Finally, a data selection criterion based on the data excitation capabilities along the data set length is proposed in Section VI. The efficiency of this selection criterion is illustrated on the same example with input signals exhibiting step-like behavior.

II. System, Model, and Data

In this section, we discuss the structure of the "true system" and of its associated parametric model as well as the characteristics of the data set used to identify this system.

The main feature of our modeling setup is that the I/O system dynamics can be modeled exactly within the given model structure, whereas the system noise dynamics cannot. Futhermore, the input signals are assumed to exhibit only a few step discontinuities.

The true system is a stable SISO ARMAX system written as

$$A_0(z)y(t) = B_0(z)u(t) + C_0(z)e(t) \tag{4}$$

where

- $y(t)$ output signal
- $u(t)$ deterministic input signal
- $e(t)$ Gaussian white noise (i.e., $N(0,\sigma^2)$)

while $(A_0, B_0, C_0)(z)$ are polynomials of order (n_a, n_b, n_c) in the delay operator z^{-1} with the classical normalization $(A_0, B_0, C_0)(\infty) = (1, 0, 1)$. Moreover, the system stability assumption requires $A_0(z)$ to have no roots in z^{-1} inside the unit circle.

We choose to identify this system using an ARX model structure of the form

$$A(z)y(t) = B(z)u(t) + \varepsilon(t) \tag{5}$$

where $(A, B)(z)$ are polynomials of order (n_a, n_b) in z^{-1} with $(A, B)(\infty) = (1, 0)$.

Note that the degrees of the polynomials constituting the input to output dynamics of the system and of the model are identical (i.e., (n_a, n_b)); thus, the system I/O dynamics can be modeled exactly. The model parameters to estimate are the coefficients of the $A(z)$ and $B(z)$ polynomials. Their total number is equal to $n_p = n_a + n_b$. By contrast, the system noise dynamics does not belong to the model set. Let us then denote by $\mu(t)$ the unmodeled part of the noise, i.e., $\mu(t) := (C_0(z) - 1)e(t)$.

We will assume in this paper that our aim is to identify the I/O transfer function as accurately as possible from open-loop data, despite the fact that the system noise dynamics are undermodeled. Thus, we will want the coefficients of $A(z)$ and $B(z)$ (i.e., the parameters) to converge as close as possible to those of $A_0(z)$ and $B_0(z)$, which will be called the "true parameters."

At any sample time t, an output prediction $\hat{y}(t)$ can be associated with the model equation (5) by the relation

$$\hat{y}(t) = B(z)u(t) - (A(z) - 1)y(t).$$

With the help of the regressor vector $\phi(t) = [-y(t-1), \cdots, -y(t-n_a), u(t-1), \cdots, u(t-n_b)]^T$, we can rewrite the system and the prediction equation in the following way:

$$\begin{aligned} y(t) &= \phi^T(t)\theta_0 + \mu(t) + e(t) \\ \hat{y}(t,\theta) &= \phi^T(t)\theta \end{aligned} \tag{6}$$

where $\theta = [a_1, \cdots, a_{n_a}, b_1, \cdots, b_{n_b}]^T$ is the $(n_p, 1)$ model parameter vector, θ_0 is the corresponding true parameter vector, and $\hat{y}(t,\theta)$ is the predicted output based on any approximation θ of θ_0. The output of both the system and the model is readily seen to depend linearly on the past input and output data constituting the regressor vector $\phi(t)$. This is actually a key property of ARX models. The coefficients of these linear combinations are precisely those of the polynomial pairs $(A_0(z), B_0(z))$ and $(A(z), B(z))$. Finally, all the regressor vectors $\phi(t)$ (with $t = 1 \cdots N$) will be assumed to be known in full in order to ignore the initialization transient phase.

In vector form, (6) can be reformulated as

$$\begin{aligned} y &= \Phi\theta_0 + \mu + e \\ \hat{y}(\theta) &= \Phi\theta \end{aligned} \tag{7}$$

where $y = [y(1), \cdots, y(N)]^T$ is the system output vector, $\hat{y}(\theta)$ is the predicted output vector at θ, and $\Phi = [\phi(1), \cdots, \phi(N)]^T$ is the (N, n_p) regressor matrix, whereas $e = [e(1), \cdots, e(N)]^T$ and $\mu = [\mu(1), \cdots, \mu(N)]^T$ are the white noise and the noise unmodeling vectors, respectively.

We now formulate our assumption concerning the data set at hand. We assume that the input signal record $\{u(t), t = 0 \cdots N-1\}$ is taken as a part of a persistently exciting sequence, which is denoted $s(t)$ of order n_p (i.e., $PE(n_p)$). This means that for all t, there exists m such that (see [5] and [3])

$$\alpha I_{n_p} < \sum_{k=t}^{t+m} \psi(k)\psi(k)^T < \beta I_{n_p} \tag{8}$$

for some positive α, β with $\psi(k) = [s(k-1), \cdots, s(k-n_p)]^T$ and I_{n_p}, which is the identity matrix of order n_p. Furthermore, we consider situations where the value of m required to make the left-hand inequality hold in (8) can be much larger than the time constants of the system and where only a finite data record of length $N > n_p$ is available for the input signal $u(t)$ in such a way that the regressor matrix Φ has full column

rank but is poorly conditioned. For example, the available input data record contains only a few step changes that are separated by long periods where the input is kept constant. This situation is typical of industrial processes for which the only excitations correspond to occasional reference changes. While such signals make the regressor matrix full rank, they typically make it poorly conditioned.

III. Optimal Estimation Vector Solution

The parameter estimation approach used in this paper is the classical LS estimate of the linear model (7): It consists of minimizing the mean square of the model prediction errors over all possible values of the parameter vector θ.

With the prediction errors defined as $\varepsilon(\theta) := y - \hat{y}(\theta)$, the LS cost function takes the form

$$C(\theta, N) = \frac{\|\varepsilon(\theta)\|_2^2}{2N}$$

where $\|\cdot\|_2$ denotes the $\mathcal{L}_2$ norm. The optimal solution vector $\hat{\theta}(N)$, which is unique if the regressor matrix Φ has full rank, results from the following minimization:

$$\hat{\theta}(N) := \arg \min_{\theta \in \mathcal{R}^{n_p}} \{C(\theta, N)\}. \tag{9}$$

In this paper, we are interested in assessing the accuracy of this solution in terms of the statistical behavior of $\hat{\theta}(N)$ with respect to the data set length N.

The solution of this LS problem can be written in terms of the pseudo-inverse Φ^+ (see [6, ch. 3]) of the regressor matrix Φ as $\hat{\theta} = \Phi^+ y$. Using the SVD techniques, we can split the Φ matrix into

$$\underset{(N,n_p)}{\Phi} = \underset{(N,r)}{U} \; \underset{(r,r)}{\Sigma} \; \underset{(r,n_p)}{V^T} \tag{10}$$

where $\Sigma = \text{diag}(\sigma_1, \cdots, \sigma_r)$ is the singular value matrix with $\sigma_i^2 = \lambda_i(\Phi^T\Phi) > 0$ for $i = 1 \cdots r$, and $r = \text{rank}(\Phi)$. V and U are left-orthogonal matrices (i.e., $U^TU = V^TV = I_r$), respectively, called the right and left singular vector matrices of Φ. The pseudoinverse of Φ then takes the form $\Phi^+ = V\Sigma^{-1}U^T$. As the regressor matrix Φ is assumed to have full column rank (see Section II), $r = n_p$ in (10).

The SVD of Φ allows one to put into light the negative consequences resulting from insufficiently informative data sets.

Let us first reformulate the system and the model equations (7) with the help of the right singular vector matrix V of Φ, as follows:

$$\begin{aligned} y &= \Phi_V \theta_{0V} + (\mu + e) \\ \hat{y}(\theta) &= \Phi_V \theta_V \end{aligned} \tag{11}$$

where $\Phi_V := \Phi V = U\Sigma$ is the (N, n_p) eigenregressor matrix, θ_V is the $(n_p, 1)$ eigenparameter vector, and $\theta_{0V} := V^T\theta_0$ is the corresponding true eigenparameter vector. Note that each column of Φ_V is orthogonal to all the others because U and Σ are, respectively, left-orthogonal and diagonal matrices.

The optimal estimate can then be expressed either in terms of the eigenparameter vector $\hat{\theta}_V$ (i.e., $= \Phi_V^+ y$)

$$\hat{\theta}_V = \theta_{0V} + \Sigma^{-1}U^T(\mu + e) \tag{12}$$

or in terms of the original parameter vector $\hat{\theta}$ (i.e., $= V\hat{\theta}_V$): $\hat{\theta} = \theta_0 + V\Sigma^{-1}U^T(\mu + e)$, which is seen to consist of n_p independent linear combinations of the optimal eigenparameter vector $\hat{\theta}_V$. In the following sections, the elements of the optimal eigenparameter vector $\hat{\theta}_V$ will be shown to be almost independent of each other.

IV. Statistical Analysis of the Parameters

In this section, we derive asymptotic expressions for the first two probability moments of the eigenparameter and of the parameter vectors, respectively. To begin with, let us introduce the excitation assumption.

A. Excitation Assumption

Recall from (10) that for a fixed value of N, the eigenvalues of the matrix $\Phi^T\Phi$, i.e., the square of the singular values of Φ, are denoted

$$\sigma_i^2 = \lambda_i(\Phi^T\Phi) \qquad i = 1 \cdots n_p \tag{13}$$

such that $\text{diag}(\sigma_i^2) = V^T(\Phi^T\Phi)V$, where V is the corresponding eigenvector matrix, which is also called the right singular vector matrix of Φ. Note that each σ_i^2 is monotonically nondecreasing with N in view of a known property of the sum of positive semi-definite matrices (see [6, Corollary 4.9]) due to Weyl. Each σ_i^2 is a stochastic quantity in view of the noise contributions to the regressor matrix Φ; the input signal $u(t)$ is considered to be deterministic.

Similarly, the eigenvalues of the matrix $E\{\Phi^T\Phi\}$, where $E\{\cdot\}$ is the expectation operator over the noise characteristics, are denoted

$$s_i^2 := \lambda_i(E\{\Phi^T\Phi\}) \qquad i = 1 \cdots n_p \tag{14}$$

such that $\text{diag}(s_i^2) = \mathcal{V}^T(E\{\Phi^T\Phi\})\mathcal{V}$, where $\mathcal{V}$ is the corresponding eigenvector matrix.

Our excitation assumption on the input signal stipulates that

$$\text{the } \Phi \text{ matrix is such that } \sigma^2 \ll \sigma_i^2 \text{ for } i = 1 \cdots n_p \tag{15}$$

where σ^2 is the variance of the white noise in (4). This excitation assumption imposes that each eigensubspace energy σ_i^2 is much larger than the system noise power σ^2; in other words, for the chosen data record of length N, the input-induced energy dominates the noise power in each eigensubspace. Under this excitation assumption, it can be shown that

$$\mathcal{V}^{(i)T}\mathcal{V}^{(i)} \approx_p 1 \quad \text{and} \quad \frac{\sigma_i^2}{s_i^2} \approx_p 1 \qquad i = 1 \cdots n_p \tag{16}$$

where $\mathcal{V}^{(i)}$ is the ith column of $\mathcal{V}$, and $\approx_p$ denotes approximation in a wide probability sense.[1] This means that each $\mathcal{V}^{(i)}$ can be considered to be an eigenvector of $\Phi^T\Phi$ with s_i^2 as eigenvalue. Note also that in view of the approximation in a

[1] A random variable $x \in \mathcal{R}$ is said to approach the real constant $x_0 \neq 0$ in a wide probability sense (i.e., $x \approx_p x_0$ or $x_0 \approx_p x$) if and only if $E\{(x/x_0 - 1)^2\} \ll 1$.

wide probability sense, the left part of (16) and the respective orthonormality of the $\mathcal{V}^{(i)}$ and $V^{(i)}$ vectors implies that

$$E\{(V^{(i)T}\mathcal{V}^{(j)})^2\} \ll 1 \qquad i,j = 1\cdots n_p \quad \text{with} \quad i \neq j. \tag{17}$$

The expressions (16) are of special interest in the present context and will be used in the sequel of this paper in the form of the following transitivity relations with $i = 1\cdots n_p$:

$$E\{V^{(i)}[\cdot]\} \approx_p V^{(i)}E\{[\cdot]\} \quad \text{and} \quad E\left\{\frac{[\cdot]}{\sigma_i^2}\right\} \approx_p \frac{E\{[\cdot]\}}{\sigma_i^2} \tag{18}$$

In other words, it turns out that under this excitation assumption, $V^{(i)}$ and σ_i^2 are approximately deterministic.

Now, instead of showing that (15) implies (16), we demonstrate that (15) implies the following equivalent expression:

$$E\left\{\left(\frac{\mathcal{V}^{(i)T}(\Phi^T\Phi)\mathcal{V}^{(j)}}{s_i s_j} - \delta_{ij}\right)^2\right\} \ll 1 \qquad i,j = 1\cdots n_p \tag{19}$$

with δ_{ij} as the Kronecker symbol. The equivalence between (19) and (16) originates from the fact that using both parts of (16) and (17) in the left-hand side of (19) straightforwardly leads to such probabilistic expression (see the definition of the wide probability sense approximation). Note also that the expression (19) is identical to

$$\mathrm{Var}\left\{\frac{\mathcal{V}^{(i)T}(\Phi^T\Phi)\mathcal{V}^{(j)}}{s_i s_j}\right\} \ll 1 \qquad i,j = 1\cdots n_p. \tag{20}$$

The reason for this is that by use of (14) together with the fact that only $\Phi^T\Phi$ is random, we have $E\{\mathcal{V}^{(i)T}(\Phi^T\Phi)\mathcal{V}^{(j)}\} = \delta_{ij}s_i^2$. Thus, the left-hand side of (19) is the expectation of the square of a zero mean random variable, i.e., its variance. Now, in order to demonstrate (19), let us evaluate this variance. To begin with, we split $\Phi_{\mathcal{V}}^{(i)} = \Phi\mathcal{V}^{(i)}$ into two parts. Considering the column $\Phi_{\mathcal{V}}^{(i)}$ as a signal over time (i.e., $\Phi_{\mathcal{V}}^{(i)}(t)$), we have the following relations:

$$\begin{aligned}\Phi_{\mathcal{V}}^{(i)}(t) &= \mathcal{V}_u^{(i)}(z)u(t) - \mathcal{V}_y^{(i)}(z)y(t)\\ &= \left[\mathcal{V}_u^{(i)}(z) - \mathcal{V}_y^{(i)}(z)\frac{B_0(z)}{A_0(z)}\right]u(t)\\ &\quad + \left[-\mathcal{V}_y^{(i)}(z)\frac{C_0(z)}{A_0(z)}\right]e(t)\\ &= G_i(z)u(t) + H_i(z)e(t)\end{aligned}$$

with

$$\begin{aligned}\mathcal{V}_u^{(i)}(z) &= \mathcal{V}_{(n_a+1)i}z^{-1} + \cdots + \mathcal{V}_{(n_a+n_b)i}z^{-n_b} \quad \text{and}\\ \mathcal{V}_y^{(i)}(z) &= \mathcal{V}_{1i}z^{-1} + \cdots + \mathcal{V}_{n_a i}z^{-n_a}.\end{aligned}$$

Note that the filters $G_i(z)$ and $H_i(z)$ are causal and stable infinite impulse response (IIR) filters with effective length[2] less than N_0, say. In vector form, we may write

$$\Phi_{\mathcal{V}}^{(i)} = G_i u + H_i e \tag{21}$$

[2]The effective length of a stable filter $F(z) = f_0 + f_1 z^{-1} + \cdots$ is defined as the smallest integer N_f such that $f_k \approx 0$ for $k \geq N_f$.

where $G_i, H_i \in \mathcal{R}^{N\times N}$ denote the Toeplitz matrix representation of the filters $G_i(z)$ and $H_i(z)$, respectively, i.e.,

$$G_i \equiv \begin{pmatrix} 0 & 0 & \cdots & 0 & 0\\ g_{i1} & 0 & \cdots & 0 & 0\\ \vdots & \vdots & \ddots & \vdots & \vdots\\ g_{i(N-2)} & g_{i(N-3)} & \cdots & 0 & 0\\ g_{i(N-1)} & g_{i(N-2)} & \cdots & g_{i1} & 0\end{pmatrix}$$

and similarly for H_i. Therefore

$$\begin{aligned}\mathcal{V}^{(i)T}(\Phi^T\Phi)\mathcal{V}^{(j)} &= \Phi_{\mathcal{V}}^{(i)T}\Phi_{\mathcal{V}}^{(j)}\\ &= u^T G_i^T G_j u + e^T H_i^T H_j e\\ &\quad + (u^T G_i^T H_j e + e^T H_i^T G_j u)\end{aligned}$$

and

$$\begin{aligned}E\{\Phi_{\mathcal{V}}^{(i)T}\Phi_{\mathcal{V}}^{(j)}\} &= u^T G_i^T G_j u + E\{e^T H_i^T H_j e\}\\ &= u^T G_i^T G_j u + \eta_{ij}N\sigma^2\end{aligned} \tag{22}$$

with $\eta_{ij} := \mathrm{tr}(H_i^T H_j)/N$, where $\mathrm{tr}(M) := \Sigma_i M_{ii}$ denotes the trace of any square matrix M. Note that (22) equals s_i^2 for $i = j$, i.e.,

$$s_i^2 = \|G_i u\|_2^2 + \eta_{ii}N\sigma^2 \tag{23}$$

while it is zero for $i \neq j$ by (14). We may then write

$$\begin{aligned}\Phi_{\mathcal{V}}^{(i)T}\Phi_{\mathcal{V}}^{(j)} &= E\{\Phi_{\mathcal{V}}^{(i)T}\Phi_{\mathcal{V}}^{(j)}\} + (e^T H_i^T H_j e - \eta_{ij}N\sigma^2)\\ &\quad + (u^T G_i^T H_j e + e^T H_i^T G_j u).\end{aligned} \tag{24}$$

Now, the variance of (24) is

$$\begin{aligned}\mathrm{Var}\{\Phi_{\mathcal{V}}^{(i)T}\Phi_{\mathcal{V}}^{(j)}\} &= E\{[(e^T H_i^T H_j e - \eta_{ij}N\sigma^2)\\ &\quad + (u^T G_i^T H_j e + e^T H_i^T G_j u)]^2\}\\ &= [\overline{\eta}_{ij}N\sigma^2 + \|(H_j^T G_i + H_i^T G_j)u\|_2^2]\sigma^2\end{aligned}$$

where

$$\overline{\eta}_{ij} := \mathrm{tr}(H_i^T H_j H_j^T H_i + H_i^T H_j H_i^T H_j)/N.$$

Finally, we may overbound the left-hand side of (19) as follows:

$$\begin{aligned}&\mathrm{Var}\left\{\frac{\Phi_{\mathcal{V}}^{(i)T}\Phi_{\mathcal{V}}^{(j)}}{s_i s_j}\right\}\\ &= \frac{\sigma^2}{s_i^2 s_j^2}[\overline{\eta}_{ij}N\sigma^2 + \|(H_j^T G_i + H_i^T G_j)u\|_2^2] \qquad (25)\\ &\leq \frac{\sigma^2}{s_i^2 s_j^2}[\overline{\eta}_{ij}N\sigma^2 + 2(\|H_j\|_2^2\|G_i u\|_2^2 + \|H_i\|_2^2\|G_j u\|_2^2)] \qquad (26)\\ &\leq \frac{2N_0\sigma^2}{s_i^2 s_j^2}[N_0\eta_{ii}\eta_{jj}N\sigma^2 + \eta_{jj}\|G_i u\|_2^2 + \eta_{ii}\|G_j u\|_2^2] \qquad (27)\\ &\leq 2N_0\left(N_0\sqrt{\eta_{ii}\eta_{jj}}\frac{\sigma^2}{s_i s_j} + \eta_{jj}\frac{\sigma^2}{s_j^2} + \eta_{ii}\frac{\sigma^2}{s_i^2}\right) \qquad (28)\\ &\ll 1 \qquad (29)\end{aligned}$$

in view of (15) for $N > N_0$. Indeed, (25) originates from the following 2-norm properties: $\|a+b\|_2^2 \le 2(\|a\|_2^2 + \|b\|_2^2)$ and $\|Aa\|_2 \le \|A\|_2\|a\|_2$ for any vectors a, b and matrix A (see [6, ch. 2] for more details). On the other hand, (26) and (27) come from matrix norm properties. First, for any matrix $F \in \mathcal{R}^{N\times N}$, we have

$$\begin{aligned}\mathrm{tr}\{FF\}/N &\le \mathrm{tr}\{F^TF\}/N \quad (F \text{ square})\\ &\le \|F\|_2^2 \quad (\text{matrix norm inequality})\end{aligned}$$

Second, for any Toeplitz matrix $F \in \mathcal{R}^{N\times N}$ representing a stable filter $F(z) = f_0 + f_1 z^{-1} + \cdots$ of effective length N_f with $N_f \ll N$ (i.e., $(F)_{ij} := f_{i-j}$ for $i, j = 1 \cdots N$ with $f_k = 0$ for $k < 0$ and $f_k \approx 0$ for $k \ge N_f$, see above), we can write

$$\begin{aligned}\|F\|_2^2 &= \max_{\|x\|_2=1} \|Fx\|_2^2\\ &\le (|f_0| + \cdots + |f_{N_f-1}|)^2 \quad (F \text{ Toeplitz})\\ &\le N_f(f_0^2 + \cdots + f_{N_f-1}^2) \quad (\text{Cauchy-Schwartz})\\ &\approx N_f \mathrm{tr}\{F^TF\}/N \quad (N_f \ll N).\end{aligned}$$

To sum up, we have the following inequalities:

$$\frac{\mathrm{tr}\{FF\}}{N} \le \frac{\mathrm{tr}\{F^TF\}}{N} \le \|F\|_2^2 \le N_f \frac{\mathrm{tr}\{F^TF\}}{N}. \tag{30}$$

Hence, $\|H_i\|_2^2 \le N_0\eta_{ii}$, whereas $\overline{\eta}_{ij} \le 2N_0^2\eta_{ii}\eta_{jj}$ (i.e., F identical to $H_i^TH_j$ in (30) and $\|H_i^TH_j\|_2 \le \|H_i\|_2\|H_j\|_2$ from a matrix 2-norm property). In (28), we used the definition of s_i^2 and s_j^2 (see (23)) and the following inequality: $a^2/(a^2+b^2) \le 1$ for any real a, b. Finally, (29) is due to (15).

Hence, under the excitation assumption (15), the expression (19) is satisfied. It then follows that (18) is valid. This allows us to indistinctly use, in the sequel of the paper, σ_i^2 for s_i^2 and $V^{(i)}$ for $\mathcal{V}^{(i)}$ in a second-order probability sense. Finally, note the respective contributions of the input and noise signals to the eigenvalues $s_i^2 = E\{\|\Phi_V^{(i)}\|_2^2\}$ (i.e., (23)) of the matrix $E\{\Phi^T\Phi\}$: s_i^2 increases with N due to the stationary noise power (i.e., $\sim \sigma^2$) but also with $\|G_iu\|_2^2$ (which is input signal dependent). Actually, this latter term brings high contributions at the time instants for which the input signal excites the ith eigensubspace determined *via* the corresponding eigenvector $\mathcal{V}^{(i)}$. Moreover, even in estimation problems with insufficient excitation, these nonpersistent contributions can easily dominate the noise energy as illustrated in Section V. This is in contrast with classical excitations (i.e., pseudo-random input signals [1], [2]) for which s_i^2 can be expressed as Ns_{0i}^2 with s_{0i}^2 constant because the $\Phi_V^{(i)}$'s are stationary over the data length N.

B. Eigenparameters: $\hat{\theta}_V(N)$

Assuming that $e(t)$ is a Gaussian white noise (i.e., $\sim N(0, \sigma^2)$) and that the excitation assumption (15) holds, we can compute the first two probability moments of the eigenparameter vector distribution (see (12)):

$$\begin{aligned}\hat{\theta}_{Vi} &= \theta_{0Vi} + \sigma_i^{-1}U^{(i)T}(\mu + e)\\ &= \theta_{0Vi} + \frac{\Phi_V^{(i)T}(\mu+e)}{\sigma_i^2}.\end{aligned}$$

Using the same decomposition of $\Phi_V^{(i)}$ as in (21) and the transitivity relations (18), we successively obtain

$$\begin{aligned}E\{\hat{\theta}_{Vi}(N)\} &\approx_p \theta_{0Vi} + \frac{E\{(u^TG_i^T + e^TH_i^T)(\mu+e)\}}{\sigma_i^2(N)}\\ &\approx_p \theta_{0Vi} + \alpha_i N\frac{\sigma^2}{\sigma_i^2(N)}.\end{aligned} \tag{31}$$

The last expression uses the independence between u and $\mu+e$ as well as between H_ie and e. Recall that $\mu = (C_0 - I)e$ (i.e., $\mu(t) = (c_{01}z^{-1} + c_{02}z^{-2} + \cdots)e(t)$) with $C_0 \in \mathcal{R}^{N\times N}$, which is the Toeplitz matrix representation of $C_0(z)$, and denote $\alpha_i := \mathrm{tr}\{H_i^T(C_0 - I)\}/N$.

Similarly

$$\begin{aligned}&\mathrm{Cov}\{\hat{\theta}_V(N)\}_{ij}\\ &\quad\approx_p \frac{E\{(\Phi_V^{(i)T}(\mu+e))((\mu+e)^T\Phi_V^{(j)})\} - \alpha_i\alpha_jN^2\sigma^4}{\sigma_i^2(N)\sigma_j^2(N)}\\ &\quad\approx_p [E\{\Phi_V^{(i)T}E\{(\mu+e)(\mu+e)^T\}\Phi_V^{(j)}\}\\ &\qquad + E\{\Phi_V^{(i)T}E\{\mu\Phi_V^{(j)T}\}\mu\}]\frac{1}{\sigma_i^2(N)\sigma_j^2(N)}\\ &\quad\approx_p [u^TG_i^TC_0C_0^TG_ju + (\kappa_{ij} + \overline{\kappa}_{ij})N\sigma^2]\frac{\sigma^2}{\sigma_i^2(N)\sigma_j^2(N)}\end{aligned} \tag{32}$$

where

$$\begin{aligned}\kappa_{ij} &:= \mathrm{tr}\{H_i^TC_0C_0^TH_j\}/N \quad \text{and}\\ \overline{\kappa}_{ij} &:= \mathrm{tr}\{H_i^T(C_0 - I)H_j^T(C_0 - I)\}/N\end{aligned}$$

are appropriate constants.

Let us point out the following:

- The mean of the eigen-parameter $\hat{\theta}_{Vi}$ is independent of the other $\hat{\theta}_{Vj}$ (for $j \ne i$). However, the noise undermodeling term introduces some correlation between the $\hat{\theta}_{Vi}$'s. Actually, if the true system is ARX, the eigenparameter $\hat{\theta}_{Vi}$ is unbiased and uncorrelated with the others: $E\{\hat{\theta}_{Vi}(N)\} \approx 0$ and $\mathrm{Cov}\{\hat{\theta}_V(N)\}_{ij} \approx_p \delta_{ij}\sigma^2/\sigma_i^2(N)$. The reason for this is that $\alpha_i = \overline{\kappa}_{ij} = 0$ and $\kappa_{ij} = \eta_{ij}$ when $\mu = 0$ (i.e., $C_0 = I$).
- The bias of $\hat{\theta}_{Vi}(N)$ (i.e., $\hat{\theta}_{Vi} - \theta_{0Vi}$) may become large if the associated singular value $\sigma_i^2(N)$ behaves like $\eta_{ii}N\sigma^2$ (see (23)); this means that there is no significant input energy in the ith right singular subspace associated to $V^{(i)}$. This bias then tends to α_i/η_{ii}.
- The variance of $\hat{\theta}_{Vi}(N)$ decreases with N because of the monotonic nondecreasing property of the corresponding singular value $\sigma_i(N)$ of Φ. However, this decrease appears to be insignificant in case of almost steady $\sigma_i(N)$ as will be seen in Section V.
- Classical pseudo-random input signals lead to estimated eigenparameters that have constant bias (i.e., a bias $\alpha_i\sigma^2/\sigma_{0i}^2$ that is significantly smaller than α_i/η_{ii} for high signal-to-noise ratio) and variances that decrease like N^{-1} for $\sigma_i^2(N) \approx N\sigma_{0i}^2$.
- In the presence of a singular value $\sigma_{i_{\min}}^2$ significantly smaller than the others (i.e., $\sigma_{i_{\min}} \ll \sigma_i$ with $i \ne i_{\min}$), the corresponding eigenparameter, which is denoted by

$\hat{\theta}_{V i_{\min}}$, is the *most poorly* estimated one because it has the largest bias and variance.

Let us then consider the parameter vector $\hat{\theta}$, which contains the actual model coefficients of real interest.

C. Parameters: $\hat{\theta}(N)$

Using the excitation assumption (leading to (18)) and the first two probability moments of the eigenparameters computed above, we can evaluate the first two probability moments of the original parameters $\hat{\theta} = V\hat{\theta}_V$. This leads to

$$E\{\hat{\theta}(N)\} \approx_p V E\{\hat{\theta}_V(N)\}$$
$$\text{Cov}\{\hat{\theta}(N)\} \approx_p V \text{ Cov}\{\hat{\theta}_V(N)\} V^T$$

As each $\hat{\theta}_k$ depends on the most poorly estimated eigenparameter $\hat{\theta}_{V i_{\min}}$, we deduce that the lowest singular value $\sigma_{i_{\min}}$ is the dominant factor that determines the accuracy of every parameter $\hat{\theta}_k$. Furthermore, in case of $\sigma_{i_{\min}} \ll \sigma_i$ with $i \neq i_{\min}$, we can write

$$E\{\hat{\theta}_k - \theta_{0k}\} \approx_p (V^{(i_{\min})})_k \frac{\alpha_{i_{\min}} N\sigma^2}{\sigma^2_{i_{\min}}(N)} \qquad k = 1 \cdots n_p$$
$$\text{Var}\{\hat{\theta}_k\} \approx_p (V^{(i_{\min})})_k^2 \text{ Var}\{\hat{\theta}_{V i_{\min}}\}$$

where $(V^{(i_{\min})})_k$ denotes the kth component of the $i_{\min}$th right singular vector associated with $\sigma_{i_{\min}}$, and $\text{Var}\{\hat{\theta}_{V i_{\min}}\}$ is obtained from (32) with $i = j = i_{\min}$.

Actually, the probability distributions of both parameter vectors is asymptotically (with $\sigma_i^2(N)$ instead of the classical N) Gaussian with the computed first moments as characteristics. However, let us stress once more that these results are valid only if the excitation assumption (15) is satisfied. Depending on the eigenregressor matrix behavior, this could need a large time interval ($N \gg 1$) to be true.

V. Simulation

In this section, we give more details concerning the simulation presented in the introduction, building on the concepts and results worked out in the preceding sections.

The input and output signals are shown in Fig. 1. The columns of the associated regressor matrix Φ are made up of the vectors $-y$ and u, respectively. The SVD of the regressor matrix Φ is performed for each value of the data set length N. The columns $\Phi_V^{(i)}$ of the eigenregressor matrix are made up of two signals represented in the upper part of Fig. 5 for $N = 250$; the associated squared singular values $\sigma_i^2(N)$ are displayed in the lower part of the same figure. Actually, these squared singular values are normalized with respect to their final values shown on the same figure: These are $\sigma_1^2(250) = 1629$ and $\sigma_2^2(250) = 5.228$.

It is seen that the SVD has split the regressor matrix Φ into two columns $\Phi_V^{(i)}$ behaving very differently with N: The first one (——), which is denoted $\Phi_V^{(1)}$, is similar to the step input signal; the other one (– –), $\Phi_V^{(2)}$ is significantly nonzero only at the jumping part of the data set. From an energy viewpoint, we remark that the $\sigma_i^2(N)$'s have specific behaviors: $\sigma_1^2(N)$, which is associated with $\Phi_V^{(1)}$, increases

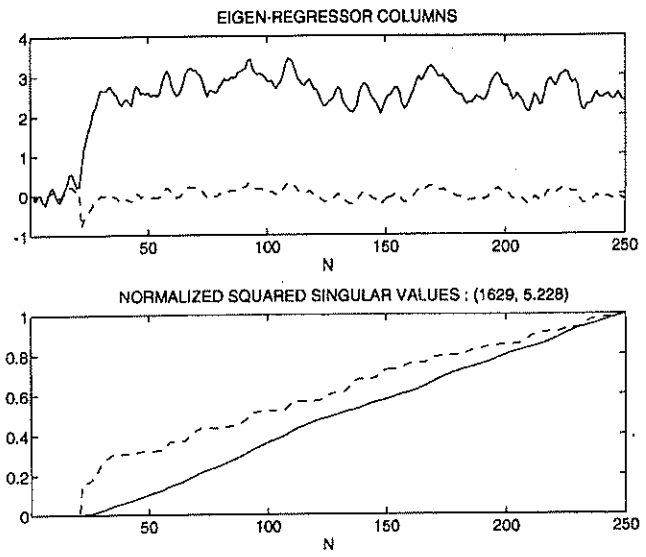


Fig. 5. Eigenregressors $\Phi_V^{(i)}$ and associated normalized energies $\sigma_i^2(N)$.

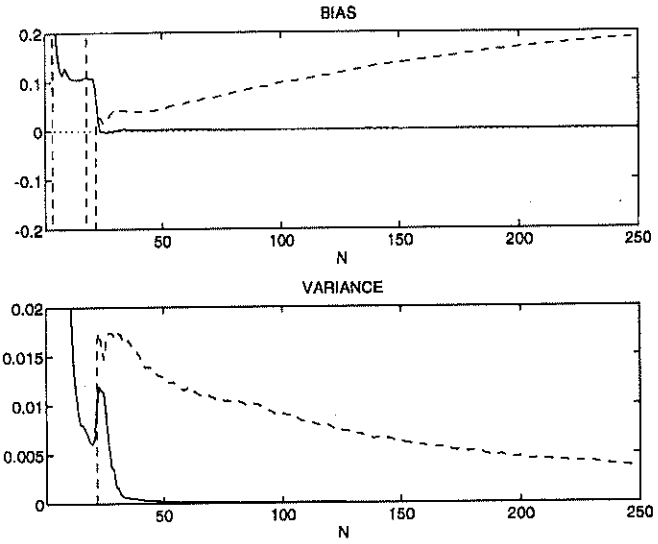


Fig. 6. Bias and variance of the $\hat{\theta}_{V1}$ (——) and $\hat{\theta}_{V2}$ (– –) eigenparameters with N.

linearly with N, whereas $\sigma_2^2(N)$ jumps to 30% of its final value just after the jump in the step input instant, and then ($N > 50$) increases very slowly and linearly with N due to the noise (see the $\eta_{ii} N\sigma^2$ term in (23)).

Finally, note the difference in the final energy of the eigenregressors, i.e., $\sigma_1^2 \sim 300\sigma_2^2$ with only 250 data samples, together with the inequality $\sigma_2^2(50) \gg \sigma^2$ supporting our excitation assumption (see (15)) even for a small number of data.

Since the system is not in the model set, we have seen in Section IV that the parameter vector $\hat{\theta}(N)$ must be biased with respect to the true parameter vector θ_0. To investigate this question, let us consider the statistical behavior of the estimated parameters. Therefore, we make Monte-Carlo simulations over 200 experiments to estimate the means and the variances of the eigenparameters and parameters, $\hat{\theta}_{Vi}$ and $\hat{\theta}_i$, respectively. Thus, we replace $E\{\cdot\}$, which is defined in

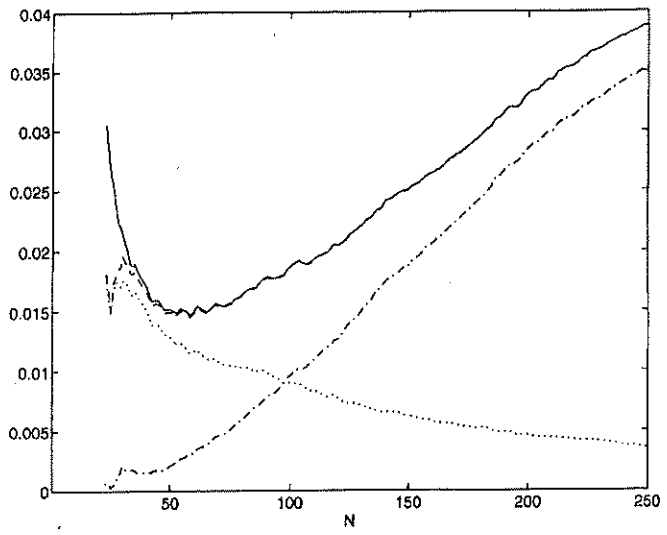


Fig. 7. Mean square error (——) of the parameters of the ARX model: See text for details.

Section IV, by the experimental average $(200)^{-1}\Sigma_{k=1}^{200}(\cdot)_k$.

The bias and the variance of the estimated parameters computed by such Monte-Carlo simulations are shown in Fig. 6 for the eigenparameters and in Fig. 2 for the actual parameters. We can see the following:

- The statistical behavior of the eigenparameters are quite different: $\hat{\theta}_{V1}(N)$, which is associated with the highest singular value $\sigma_1(N)$, has insignificant bias and very low variance, whereas $\hat{\theta}_{V2}(N)$ (which is associated with $\sigma_2(N)$) becomes highly biased with a variance that slowly decreases with N after the input jump ($N > 50$).

  Actually, $\hat{\theta}_{V2}$ is the most poorly estimated eigenparameter (i.e., $i_{\min} = 2$) because it is associated with the smallest singular value $\sigma_2(N)$.
- The actual parameters $\hat{\theta}_i(N)$ become strongly biased with slowly decreasing variances for $N > 50$ because they both depend on the most poorly estimated eigenparameter $\hat{\theta}_{Vi_{\min}} := \hat{\theta}_{V2}$.

These observations are fully consistent with our remarks of Sections IV-B and C.

For what concerns the total MSE (3), it can be written with the help of the estimated eigenparameters $\hat{\theta}_{Vi}(N)$ as

$$\text{MSE}(N) = \sum_{i=1}^{2}[(E\{\hat{\theta}_{Vi}(N) - \theta_{0Vi}\})^2 + \text{Var}\{\hat{\theta}_{Vi}(N)\}]. \tag{33}$$

With the bias and the variances computed by the Monte Carlo simulations, we see in Fig. 7 that a minimum of MSE(N) (——) is reached just after the step instant. Moreover, $\text{MSE}_{V2}(N)$ (− −), that is, the mean square error computed for the most poorly estimated eigenparameter $\hat{\theta}_{V2}(N)$ (i.e., $i = 2$ in (33) for $i_{\min} = 2$) almost exactly matches MSE(N). The squared bias (− ·) and the variance (· · ·) of $\hat{\theta}_{V2}(N)$ are also represented in the figure in order to emphasize the respective contributions of the bias and the variance errors to $\text{MSE}_{V2}(N)$.

Finally, let us remark that a stochastic input signal with a signal-to-noise ratio of about 3.5 dB (i.e., $u(t) \sim N(0, \sigma_u^2)$ with $\sigma_u^2 \approx 2.2\sigma^2$) leads to the same minimum of the MSE as the one achieved in Fig. 7. This means that our step signal, considered during the whole time interval, exhibits poor excitation capabilities because it contains only a few interesting high signal-to-noise ratio samples all located in time around the step instant.

As mentioned in Section I, this simulation provides evidence that it may be sensible to discard parts of the identification data set while estimating the model parameters. These data do not bring enough information about the input to output system dynamics to significantly decrease the variance of the estimated parameters, but, worse, they seriously increase their bias.

VI. Removal of Data

This last remark suggests the idea of selecting appropriate data subsets of the data set that lead to monotonically increasing model parameter accuracy with N, i.e., decreasing MSE(N).

On the basis of the statistical analysis performed in the preceding sections, the interesting time intervals are those for which the information carried by the input data is much larger than that coming from the noise. Indeed, these data subsets contribute to decrease the bias of the parameters because they mainly reflect the I/O system dynamics that belong to the model set. In contrast, the remaining data subsets reflect the noise system dynamics and consequently tend to increase the parameter bias in the case of noise undermodeling and common parameters between the I/O and noise models.

The excitation information exhibited by the data set as a function of time can be read in the singular values of the model regressor matrix Φ (i.e., the energy of its associated eigenregressor columns). From (23), the input and noise signals are seen to contribute to their values in a very different manner (linear with N for the noise). Therefore, from the time variations of these singular values, we can determine the time intervals for which these variations are significantly larger than the noise contribution alone. Moreover, as the accuracy of the model parameters $\hat{\theta}$ depends on the most poorly estimated eigenparameter $\hat{\theta}_{Vi_{\min}}$, it is enough to consider the time variations of its associated singular value $\sigma_{i_{\min}}(N)$. This is why we propose to use a data removal criterion of the form

$$\Delta\sigma_{i_{\min}}^2(N) := \sigma_{i_{\min}}^2(N) - \sigma_{i_{\min}}^2(N-1) < \eta_c \tag{34}$$

for some appropriate threshold value η_c. This selection criterion means that we discard the regressor vector $\phi(N)$ for which the inequality is satisfied. Note that the threshold η_c depends on the power (or variance) of the noise acting on the output: referring to (23), η_c should be chosen at least of order $\eta_{i_{\min}}\sigma^2$ in order to discard the regressor rows whose $i_{\min}$th element is only noise dependent. Actually, this lower bound on η_c can be estimated through the lowest value of the slope of the graph of $\sigma_{i_{\min}}^2(N)$ as a function of the data number N; see, e.g., Fig. 5.

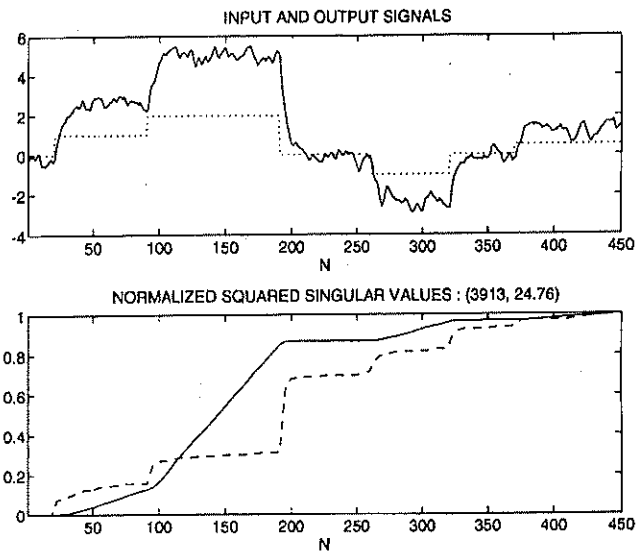


Fig. 8. Input $(\cdots)$ and output (——) signals and associated normalized eigen-regressor energies $\sigma_i^2(N)$.

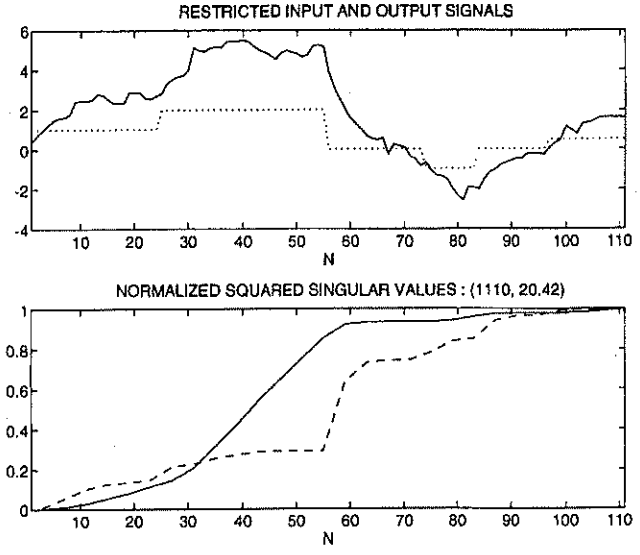


Fig. 9. Restricted input $(\cdots)$ and output (——) signals and associated normalized eigenregressor energies $\sigma_i^2(N)$.

From a practical point of view, the discarding criterion focuses only on the smallest singular value of the $\Phi(N)$ matrix, i.e., $\sigma_{i_{\min}}(N)$. Therefore, we suggest that we track only this singular value and the corresponding singular subspace $V^{(i_{\min})}(N)$ instead of computing the whole SVD for each N; a procedure for doing this can be found, e.g., in [4]. This makes the computational burden of the discarding procedure of the same order as the original LS estimation (i.e., Nn_p^2).

We have tested this selection criterion on simulated examples using the system and model description of Section I and an input signal made of several successive steps with $\eta_c = 0.035$. The original data set and the normalized squared singular values associated with its eigenregressors are shown in Fig. 8. Fig. 9 displays the same variables after removal from the original data set of all the data satisfying (34). On the first hand, it is seen that the surviving data subset is four times shorter than the original one and that the index $i_{\min}$ of the smallest singular value does not change with N in both the original and the restricted data set: $i_{\min} = 2$ $(- -)$. On the other hand, it can be noted that the qualitative behavior of the singular values are in both cases quite similar.

The bias and the variance of the estimated parameters are shown in Figs. 10 and 11 for the full data set and the restrictive data set, respectively. We observe that the restricted data set yields smaller parameter biases, whereas the variances have similar values. For what regards the overall parameter estimation accuracy, Fig. 12 shows the MSE of the parameters estimated from the original data set (——) and the associated restricted set $(\cdots)$. Actually, the dotted line represents the evolution of the MSE obtained from the restricted data set but drawn as a function of the actual data set length N; this allows a realistic comparison of the evolution of the MS errors as a function of time. We see that the proposed removal of data leads to a generally decreasing MSE with N, whereas the whole set obviously does not.

Hence, the proposed criterion properly selects the high signal-to-noise ratio samples from data sets exhibiting problems of excitation. Of course, the criterion threshold η_c has to be chosen to keep the really informative samples (η_c not too big) and to discard as efficiently as possible the uninteresting ones (η_c not too small). The average and the standard deviation values (using Monte-Carlo simulations) of the restricted data set length are displayed in Fig. 13 as functions of the criterion threshold η_c for the same system and input as above. We observe that the sensitivity of the restricted data set length to the threshold value η_c is much larger for small η_c (i.e., $\eta_c < 0.03$) than for large η_c. This is because, for small η_c, the removed samples are essentially due to the noise and its contribution due to the smallest singular value of the regressor matrix. This contribution is considerably less important than that of the remaining interesting data samples, which are relatively few. This reflects that the excitation capabilities of the input signal are really local in time. Moreover, this explains why $\eta_c = 0.035$ was a rather good discriminating threshold.

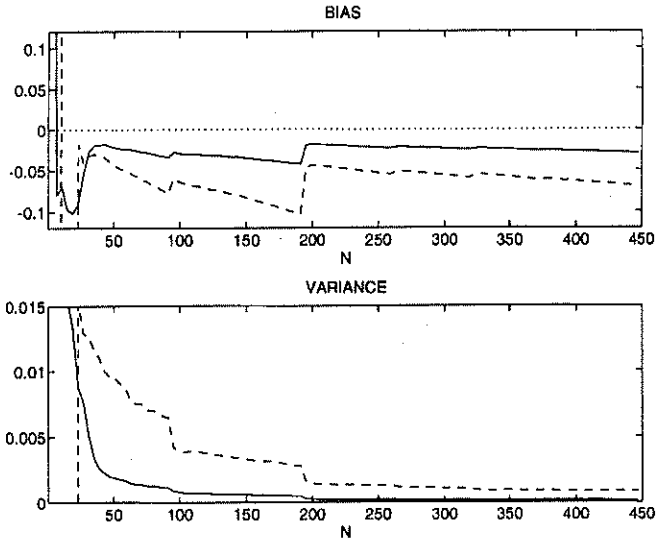


Fig. 10. Bias and variance of the $\hat{\theta}_1$ (——) and $\hat{\theta}_2$ $(- -)$ parameters with N using the full data set.

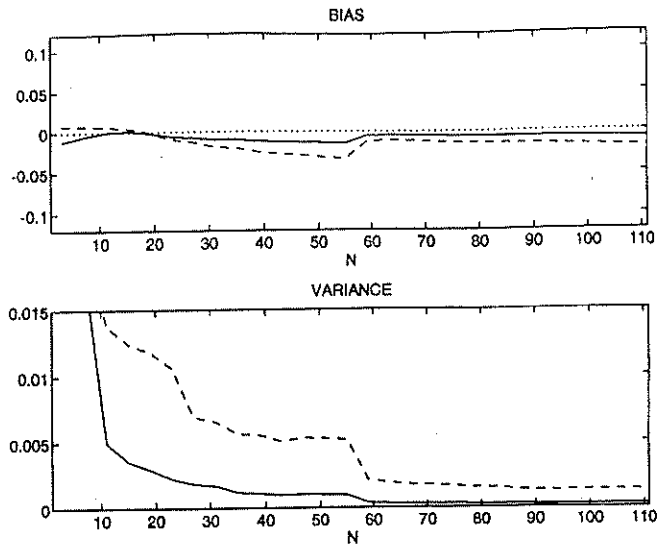


Fig. 11. Bias and variance of the $\hat{\theta}_1$ (——) and $\hat{\theta}_2$ (– –) parameters with N using the restricted data set.

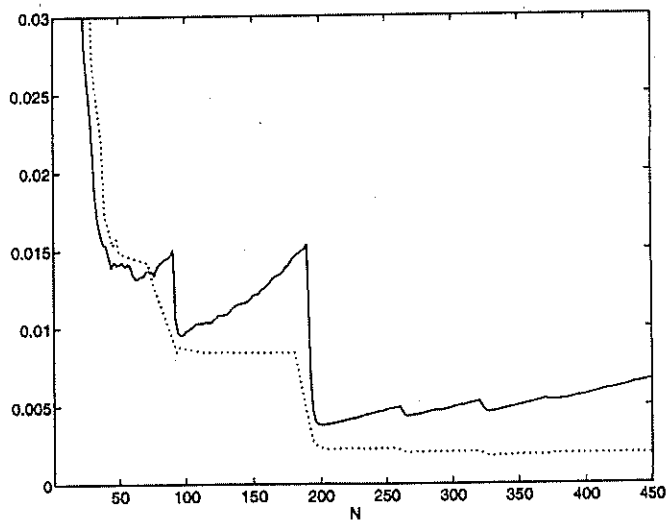


Fig. 12. Mean square errors as a function of N using the full and the restricted data set.

VII. Conclusions

We have exhibited a situation where the use of insufficently rich input data for system identification may deteriorate the accuracy of parameter estimates of the input-output model and, therefore, the accuracy of the transfer function estimate. Such insufficiently rich input data are typical of industrial process applications where the input is a reference signal that changes from time to time.

Our statistical analysis has been limited to the case where the "true system" is an ARMAX system, and where the model structure is of ARX type, in such a way that the I/O dynamics of the true system can be modeled exactly but not the noise dynamics. A characteristic feature of this setup is that the I/O and noise dynamics have a common denominator. We have shown that for such setup, the mean square error (MSE) of the estimated parameters may increase during time intervals in which the input signal is not exciting, and we have explained why. We have also proposed a discarding procedure that eliminates data that cause a deterioration of the total MSE. The phenomenon that we have highlighted can be generalized to all situations in which an equation error model structure is used but with undermodeling of the noise dynamics. However, the statistical analysis for model structures other than ARX will be more complicated.

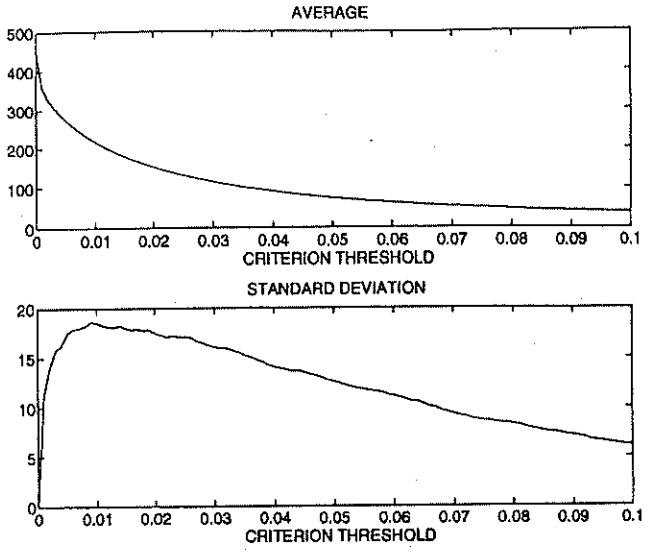


Fig. 13. Average and standard deviation of the restricted data set length as a function of η_c.

Acknowledgment

This paper presents research results of the Belgian Program on Interuniversity Poles of Attraction, initiated by the Belgian State, Prime Minister's Office for Science, Technology, and Culture. The scientific responsability rests with its authors.

References

[1] M. Gevers and L. Ljung, "Optimal experiment designs with respect to the intended model application", *Automatica*, vol. 22, no. 5, pp. 543–554, 1986.

[2] G. C. Goodwin and R. L. Payne, *Dynamic System Identification: Experiment Design and Data Analysis*. New York: Academic, 1977, series in Mathematics in Science and Engineering.

[3] L. Ljung, *System Identification: Theory for the User*. Englewood Cliffs, NJ: Prentice-Hall, 1987.

[4] M. Moonen, P. Van Dooren, and J. Vandewalle, "A systolic array for SVD updating," *SIAM J. Matrix Anal. Appl.*, vol. 14, no. 2, pp. 353–371, 1993.

[5] T. Soderstrom and P. Stoica, *System Identification*. Englewood Cliffs, NJ: Prentice-Hall Int., 1990.

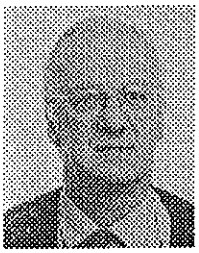
[6] G. W. Stewart and J. G. Sun, *Matrix Perturbation Theory*. San Diego: Academic, 1990, Series in Computer Science and Scientific Computing.

Pierre Carrette was born in Namur, Belgium, in 1967. He received the electrical engineering degree in 1989 and the physicist degree in 1990 from the Université Catholique de Louvain, Belgium. Since 1992, he has been a Ph.D. student in system identification at the Center for Systems Engineering and Applied Mechanics (CESAME), Belgium.

His research interests are in system identification, stochastic processes, wavelet theory, and its related applications.

Georges Bastin received the electrical engineering degree and the Ph.D. degree, both from Université Catholique de Louvain, Louvain-la-Neuve, Belgium.

He is currently Professor in the Center for Systems Engineering and Applied Mechanics (CESAME) at the Université Catholique de Louvain, Belgium. His main research interests are in system identification, nonlinear control theory, adaptive systems and random fields with applications to mechanical systems and robotics, biological processes, transportation systems, and environmental problems.



Yves V. Genin (M'76–SM'82–F'84) received the B.S. degree in electrical engineering from the University of Louvain, Belgium, and the Doctoral degree in applied mathematics from the University of Liege, Belgium, in 1962 and 1969, respectively.

From 1963 to 1991, he was with Philips Research Laboratory, Belgium, where he was head of the Applied Mathematics Department. From 1974 to 1976 and from 1984 to 1985, he was a visiting professor at the Department of Mathematics of the Facultes Universitaires de Namur, Belgium, and from 1979 to 1980, he was a consulting professor with the Electrical Engineering Department, Stanford University, Stanford, CA. He is presently a part-time professor at the Applied Sciences Department of the University of Louvain, Belgium. His main interests lie in the field of fast numerical algorithms for digital signal processing applications.

Dr. Genin received the 1974 IEEE Guillemin–Cauer Award. He currently serves as an Associate Editor for the *SIAM Journal on Matrix Analysis and Applications* and *Multidimensional Systems and Signal Processing*. He is a former Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS, the *Philips Journal of Research*, and *Mathematics of Control, Signals, and Systems*."

Michel Gevers (F'90) was born in Antwerp, Belgium, in 1945. He received the electrical engineering degree from the Université Catholique de Louvain, Louvain la Neuve, Belgium, in 1968, and the Ph.D. degree from Stanford University, Stanford, CA, in 1972.

He is now Professor and Head of the Laboratoire d'Automatique, Dynamique et Analyze des Systèmes and one of the Co-Directors of the Centre for Engineering Systems and Applied Mechanics (CESAME) at the Université Catholique de Louvain. He has spent long-term visits at the University of Newcastle, Australia, the Technical University of Vienna, Austria, and the Australian National University, Canberra. His research interests are in system identification and its connection with robust control design, adaptive estimation and control, multivariable system theory, optimal control and filtering, and the numerical aspects of filter and controller design. He has published two books and close to 150 papers and conference papers in these areas.

Prof. Gevers has been Associate Editor of *Automatica* and of the IEEE TRANSACTIONS ON AUTOMATIC CONTROL. He is presently Associate Editor of *Mathematics of Control, Signals, and Systems* and Associate Editor at Large of the *European Journal of Control*. He is a co-author, with R. R. Bitmead and V. Wertz of *Adaptive Optimal Control—The Thinking Man's GPC* (Englewod Cliffs, NJ: Prentice-Hall, 1990) and of *Parametrizations in Control, Estimation and Filtering Problems: Accuracy Aspects* (New York: Springer-Verlag Communications and Control Engineering Series, 1993) with G. Li. He is a Fellow of the Belgian American Educational Foundation and is presently member of the Board of Governors of the IEEE Control System Society.