Final version
C84

Discarding data may improve the parameter estimation accuracy in system identification.[*]

P. Carrette[†] Y. Genin, G. Bastin and M. Gevers

CESAME, Université Catholique de Louvain
Bâtiment Euler, Avenue G. Lemaître 4-6, B-1348 Louvain-la-Neuve, Belgium

E-mail : carrette@auto.ucl.ac.be

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 techniques, we show that in noise undermodeling situations, the remaining data may introduce large bias on the model parameters with a possible increase of the total mean square error. A data selection criterion is then proposed to discard such poorly informative data so as to increase the accuracy of the transfer function estimate.

Keywords : identification, persistence of excitation, singular value decomposition.

1 Introduction

The aim of this paper is to analyse the accuracy of the prediction error method [5] 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 while the model structure is chosen as a SISO ARX model whose input to output dynamics is able to represent that of the true system exactly. To motivate the present study, consider the following ARMAX system

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

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

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

on the basis of a finite number, N, of input-output (I/O) data so as to obtain the best approximation of the actual system in the least squares (LS) prediction error sense [5]. In (1) and (2), $u(t)$ and $y(t)$ stand, respectively, for the input and output signals while $e(t)$ denotes a white noise disturbance and $\varepsilon(t)$ is a modeling error.

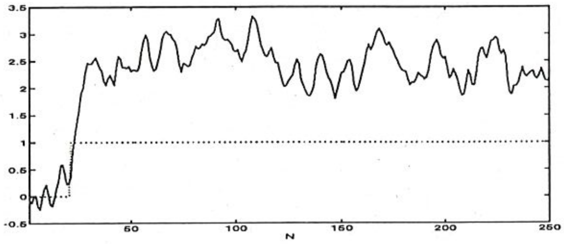

Figure 1: Input ($\cdots$) and output (—) signals of the ARMAX system.

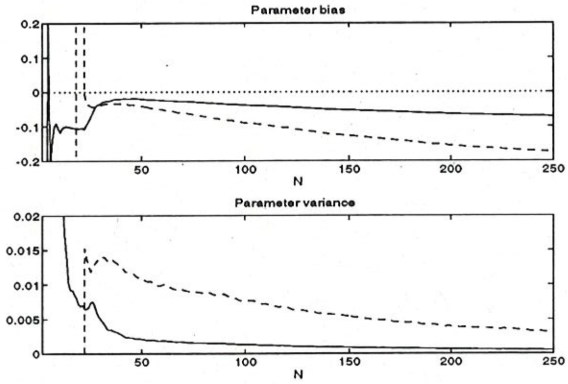

Figure 2: Bias and variance of the parameters $\hat{\theta}_1$ (—) and $\hat{\theta}_2$ (−−) as a function of N.

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.

The applied step input and the resulting system output signal are displayed in Figure 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 Figure 2. This fig-

[*] 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.

[†] Supported by a grant from IRSIA

ure shows that the variance decreases monotonically with the data length, while 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. 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 so as 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 are not persistently exciting, 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.

The paper is organized as follows. In Section 2, we present the system and the model structure as well as the data characteristics considered in the paper. In Section 3, we introduce the parameter estimation method based on the minimisation of the model prediction errors and we solve this parameter estimation problem using the singular value decomposition of the model regressor matrix. This decomposition actually splits the estimation accuracy into well and poorly estimated eigen-parameters. The statistical behaviour of these parameters is analysed in Section 4 and linked to that of the original model parameters. In Section 5, we describe in detail simulations of the parameter estimation procedure applied to the motivating system example. Finally, a data selection criterion based on the data excitation capabilities along the data set length is proposed in Section 6 in order to improve the accuracy of the estimated parameters. The efficiency of this selection criterion is illustrated on the same example with input signals exhibiting step-like behaviour.

2 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 true system is a scalar stable SISO ARMAX system written as

$$A_0(z)y(t) = B_0(z)u(t) + C_0(z)e(t) \tag{3}$$

where $(u(t), y(t))$ is the scalar system I/O data pair, $e(t)$ is a Gaussian white noise (i.e. $N(0,\sigma^2)$) and $(A_0, B_0, C_0)(z)$ 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

$$A(z)y(t) = B(z)u(t) + \varepsilon(t) \tag{4}$$

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 estimate 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 (4) by the relation

$$\hat{y}(t) = B(z)u(t) - (A(z) - 1)y(t) \tag{5}$$

With the help of the regressor vector $\phi(t) = [-y(t-1), \ldots, -y(t-n_a), u(t-1), \ldots, u(t-n_b)]^T$, 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 $\theta = [a_1, \ldots, a_{n_a}, b_1, \ldots, b_{n_b}]^T$ 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. In the sequel, all the regressor vectors $\phi(t)$ (with $t = 1 \cdots N$) will be assumed to be known in full so as to ignore the initialization transient phase. In vector form, the equation (6) can be reformulated as :

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

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

Regarding the data set, we assume that the input signal $u(t)$ is taken from a piecewise constant and persistently exciting signal of order n_p (i.e. $PE(n_p)$). This means that, for all t, there exists m such that (see [7, 5])

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

for some positive α, β with $\psi(k) = [u(k-1), \cdots, u(k-n)]^T$ and I_{n_p} the identity matrix of order n_p.

However, 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 length of $N > n_p$ data are available such 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.

3 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

$$C(\theta, N) := \frac{\|\varepsilon(\theta)\|_2^2}{2N} \tag{9}$$

where $\|.\|_2$ 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

$$\hat{\theta}(N) := \arg \min_{\theta \in \mathcal{R}^{n_p}} \{C(\theta, N)\} \tag{10}$$

The solution of this LS problem can be written in terms of the pseudo-inverse Φ^+ (see [8, chapter 3]) of the regressor matrix Φ as $\hat{\theta} = \Phi^+ y$. Using the singular value decomposition techniques (SVD), we can split the Φ matrix into

$$\begin{array}{ccccc} \Phi & = & U & \Sigma & V^T \\ (N, n_p) & & (N, r) & (r, r) & (r, n_p) \end{array} \tag{11}$$

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 respectively called the right and left singular vector matrices of Φ. The pseudo-inverse 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 2), $r = n_p$ in (11).

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

$$\begin{aligned} y &= \Phi_V \theta_{0V} + (\mu + e) \\ \hat{y}(\theta) &= \Phi_V \theta_V \end{aligned} \tag{12}$$

where $\Phi_V := \Phi V = U\Sigma$ is the (N, n_p) eigen-regressor matrix, θ_V is the $(n_p, 1)$ eigen-parameter vector and $\theta_{0V} := V^T\theta_0$ is the corresponding true eigen-parameter vector. Note that each column of Φ_V is orthogonal to all the others, for U and Σ are, respectively, left-orthogonal and diagonal matrices. The optimal estimate can then be expressed either in terms of the eigen-parameter vector $\hat{\theta}_V$ (i.e. $= \Phi_V^+ y$)

$$\hat{\theta}_V = \theta_{0V} + \Sigma^{-1}U^T(\mu + e) \tag{13}$$

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 eigen-parameter vector $\hat{\theta}_V$.

4 Statistical analysis of the parameters

In this section, we derive asymptotic expressions for the first two probability moments of the eigen-parameter and of the parameter vectors, respectively.

4.1 Excitation assumption

Recall from (11) 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 :

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

such that $\text{diag}(\sigma_i^2) = V^T(\Phi^T\Phi)V$ with V the right singular vector matrix of Φ. Note that each σ_i^2 is monotonically increasing with N (see [8, Corollary 4.9]). Similarly, the eigenvalues of the matrix $E\{\Phi^T\Phi\}$, with $E\{.\}$ the expectation operator over the noise characteristics, are denoted

$$s_i^2 := \lambda_i(E\{\Phi^T\Phi\}) \qquad i = 1 \cdots n_p \tag{15}$$

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

The excitation assumption then stipulates that :

$$\Phi \textit{ is such that } \sigma^2 \ll \sigma_i^2 \textit{ for } i = 1 \cdots n_p \tag{16}$$

where σ^2 is the variance of the white noise in (3). This excitation assumption imposes that each eigen-subspace energy σ_i^2 is much larger than the system noise power σ^2; in other words, the input-induced energy dominates the noise power in each eigen-subspace. Under this excitation assumption, it is shown in [1] that

$$V^{(i)T}\mathcal{V}^{(i)} \approx_p 1 \;\text{ and }\; \frac{\sigma_i^2}{s_i^2} \approx_p 1 \qquad i = 1 \cdots n_p \tag{17}$$

where $\mathcal{V}^{(i)}$ is the i-th 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 as an eigenvector of $\Phi^T\Phi$ with s_i^2 as eigenvalue. The expressions (17) 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 $i = 1 \cdots n_p$:

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

In other words, it turns out that, under this excitation assumption, $V^{(i)}$ and σ_i^2 are approximately deterministic. This allows us to indistinctly use, in the sequel of the paper, σ_i^2 for s_i^2 and $V^{(i)}$ for $\mathcal{V}^{(i)}$.

Finally, let us give a close expression to the eigenvalues s_i^2 of the matrix $E\{\Phi^T\Phi\}$. Considering the eigen-regressor column $\Phi_{\mathcal{V}}^{(i)} = \Phi\mathcal{V}^{(i)}$ as a signal over time (i.e. $\Phi_{\mathcal{V}}^{(i)}(t)$), we have the following relations :

$$\begin{aligned} \Phi_{\mathcal{V}}^{(i)}(t) &= \mathcal{V}_u^{(i)}(z)u(t) - \mathcal{V}_y^{(i)}(z)y(t) \\ &= G_i(z)u(t) + H_i(z)e(t) \end{aligned}$$

where 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 :

$$\Phi_{\mathcal{V}}^{(i)} = G_i u + H_i e \tag{19}$$

[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 $E\{(x/x_0 - 1)^2\} \ll 1$.

[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 $k \geq N_f$.

with $G_i, H_i \in \mathcal{R}^{N\times N}$, the Tœplitz matrix representation of the filters $G_i(z)$ and $H_i(z)$, respectively (see [1]). So, we have :

$$\begin{aligned} s_i^2 &= E\{\|\Phi_V^{(i)}\|_2^2\} \\ &= \|G_i u\|_2^2 + \eta_i N\sigma^2 \end{aligned} \qquad (20)$$

with $\eta_i := \mathrm{Tr}\{H_i^T H_i\}/N$, where Tr{} is the Trace operator (see [8]). Note the respective contributions of the input and noise signals to these eigenvalues : s_i^2 increases with N due to the stationary noise power (i.e. $\sim \sigma^2$) but also with $\|G_i u\|_2^2$ (which is input-signal dependent). Actually, this latter term brings high contributions at the time instants for which the input signal excites the i-th eigen-subspace determined *via* the corresponding eigenvector $\mathcal{V}^{(i)}$. This is in contrast with classical excitations (i.e. pseudo-random input signals [3, 4]) for which s_i^2 can be expressed as Ns_{0i}^2 with s_{0i}^2 constant, for the $\Phi_V^{(i)}$'s are stationary over the data set length N.

4.2 Eigen-parameters : $\hat{\theta}_V(N)$

Assuming that $e(t)$ is a Gaussian white noise (i.e. $\sim N(0,\sigma^2)$) and that the excitation assumption (16) holds, we can compute the first two probability moments of the eigen-parameter vector distribution $\hat{\theta}_V(N)$ (see (13)). Using the same decomposition of $\Phi_V^{(i)}$ as in (19), the transitivity relations (18) as well as the independence between u and $\mu + e$, and also between $H_i e$ and e, we obtain :

$$E\{\hat{\theta}_{Vi}(N)\} \approx_p \theta_{0Vi} + \alpha_i N \frac{\sigma^2}{\sigma_i^2(N)}, \qquad (21)$$

where $\alpha_i := \mathrm{Tr}\{H_i^T(C_0 - I)\}/N$ denotes a real constant with $C_0 \in \mathcal{R}^{N\times N}$, the Tœplitz matrix representation of $C_0(z)$. Similarly :

$$Cov\{\hat{\theta}_V(N)\}_{ij} \approx_p \frac{[u^T G_i^T C_0 C_0^T G_j u + \kappa_{ij} N\sigma^2]}{\sigma_i^2(N)\sigma_j^2(N)}\sigma^2 \qquad (22)$$

where $\kappa_{ij} := \mathrm{Tr}\{H_i^T C_0 C_0^T H_j + H_i^T(C_0 - I)H_j^T(C_0 - I)\}/N$ is an appropriate constant (see [1] for more details).

Let us point out that :

- the mean of the eigen-parameter $\hat{\theta}_{Vi}$ is independent of the other $\hat{\theta}_{Vj}$ (for $j \neq i$). However, the noise undermodeling term introduces some correlation between the $\hat{\theta}_{Vi}$'s. Actually, if the true system is ARX (i.e. $C_0 = I$), the eigen-parameter $\hat{\theta}_{Vi}$ is unbiased and uncorrelated with the others : $E\{\hat{\theta}_{Vi}(N)\} \approx 0$ and $Cov\{\hat{\theta}_V(N)\}_{ij} \approx_p \delta_{ij}\sigma^2/\sigma_i^2(N)$.
- 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 $\eta_i N\sigma^2$ (see (20)), i.e. when there is no significant input energy in the i-th right singular subspace associated to $V^{(i)}$ (i.e. $\sigma_i^2(N) \sim \eta_i N\sigma^2$, see (20)).
- the variance of $\hat{\theta}_{Vi}(N)$ monotonically decreases with N because of the monotonic increase of $\sigma_i^2(N)$ of Φ.
- in the presence of a singular value $\sigma_{i_{\min}}$ significantly smaller than the others (i.e. $\sigma_{i_{\min}} \ll \sigma_i$ with $i \neq i_{\min}$), the corresponding eigen-parameter, denoted by $\hat{\theta}_{Vi_{\min}}$, is the *most poorly* estimated one for 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.

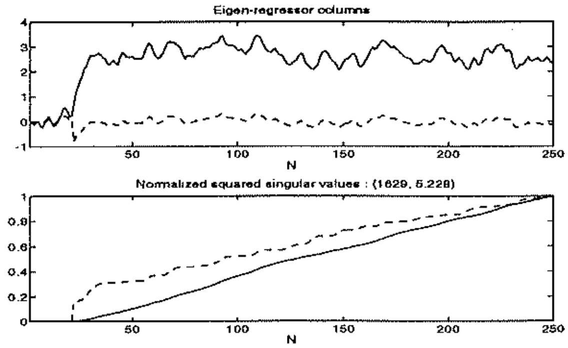

Figure 3: Eigen-regressors $\Phi_V^{(i)}$ and associated normalized energies $\sigma_i^2(N)$.

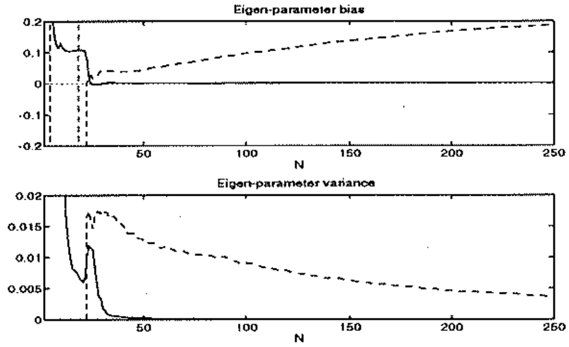

Figure 4: Bias and variance of the $\hat{\theta}_{V1}$ (—) and $\hat{\theta}_{V2}$ (−−) eigen-parameters with N.

4.3 Parameters : $\hat{\theta}(N)$

Using the excitation assumption (leading to (18)) and the first two probability moments of the eigen-parameters computed above, we can evaluate the first two probability moments of the original parameters $\hat{\theta} = V\hat{\theta}_V$. In case of $\sigma_{i_{\min}} \ll \sigma_i$ with $i \neq i_{\min}$, we can write :

$$\begin{aligned} E\{\hat{\theta}_i - \theta_{0i}\} &\approx_p (V^{(i_{\min})})_i \frac{\alpha_{i_{\min}} N\sigma^2}{\sigma_{i_{\min}}^2(N)} \\ Var\{\hat{\theta}_i\} &\approx_p (V^{(i_{\min})})_i^2 Var\{\hat{\theta}_{Vi_{\min}}\} \end{aligned} \qquad (23)$$

where $(V^{(i_{\min})})_i$ denotes the i-th component of the $i_{\min}$-th right singular vector associated to $\sigma_{i_{\min}}$, and $Var\{\hat{\theta}_{Vi_{\min}}\}$ is obtained from (22) with $i = j = i_{\min}$. Thus, the probability moments of the most poorly estimated eigen-parameter $\hat{\theta}_{Vi_{\min}}$ determine the accuracy of every parameter $\hat{\theta}_i$ and, moreover, the lowest singular value $\sigma_{i_{\min}}$ is the dominant factor of the estimation quality.

5 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 Figure 1. The columns of the associated regressor matrix Φ are made up of the vectors $-y$ and u, respectively. The singular value decomposition of the regressor matrix Φ is performed for each value of the data set length N. The columns $\Phi_V^{(i)}$ of the eigen-regressor matrix are made up of two signals represented in the

upper part of Figure 3 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 (—), denoted $\Phi_V^{(1)}$, is similar to the step input signal; the other one (−−), $\Phi_V^{(2)}$, is significantly non-zero only at the jumping part of the data set. From an energy viewpoint, we remark that the $\sigma_i^2(N)$'s have specific behaviours : $\sigma_1^2(N)$, associated with $\Phi_V^{(1)}$, increases linearly with N, while $\sigma_2^2(N)$ 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 $\eta_i N\sigma^2$ term in (20)). Finally, note the difference in the final energy of the eigen-regressors, 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 (16)) even for a small number of data.

Since the system is not in the model set, we have seen in Section 4 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 behaviour of the estimated parameters. Therefore, we make Monte-Carlo simulations over 200 experiments to estimate the means and the variances of the eigen-parameters and parameters, $\hat{\theta}_{Vi}$ and $\hat{\theta}_i$, respectively. The bias and the variance of the estimated parameters, computed by such Monte-Carlo simulations, are shown in Figures 4 (for the eigen-parameters) and 2 (for the actual parameters). It can be seen that the statistical behaviour of the eigen-parameters are quite different : $\hat{\theta}_{V1}(N)$, associated to the highest singular value $\sigma_1(N)$, has insignificant bias and very low variance while $\hat{\theta}_{V2}(N)$ (associated to $\sigma_2(N)$) 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 eigen-parameter (i.e. $i_{\min} = 2$). Moreover, the actual parameters $\hat{\theta}_i(N)$ become strongly biased with slowly decreasing variances for $N > 50$ (see Figure 2), for they both depend on the most poorly estimated eigen-parameter $\hat{\theta}_{Vi_{\min}} := \hat{\theta}_{V2}$. Note also that these observations are fully consistent with our remarks of Sections 4.2 and 4.3. For what concerns the total mean square error (MSE) of the estimated parameters, it can be written with the help of the estimated eigen-parameters $\hat{\theta}_{Vi}(N)$ as :

$$MSE(N) = \sum_{i=1}^{2}\{E\{\hat{\theta}_{Vi}(N) - \theta_{0Vi}\}^2 + Var(\hat{\theta}_{Vi}(N))\} \quad (24)$$

With the bias and the variances computed by the Monte-Carlo simulations, we see in Figure 5 that a minimum of $MSE(N)$ (—) is reached just after the step instant. Moreover, $MSE_{V2}(N)$ (−−), that is the MSE computed for the most poorly estimated eigen-parameter $\hat{\theta}_{V2}(N)$ (i.e. $i = 2$ in (24) 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 $MSE_{V2}(N)$.

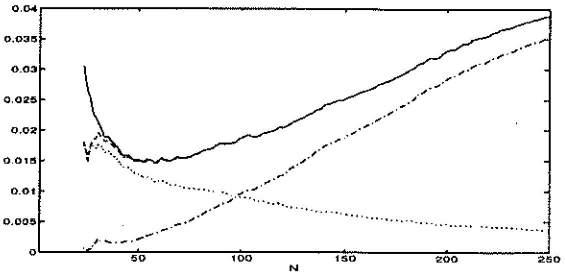

Figure 5: Mean square error (—) of the parameters of the ARX model : see text for details.

As mentioned in Section 1, 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.

6 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, it appears that the interesting time intervals are those for which the information carried by the input data is much larger than that coming from the noise. And 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 eigen-regressor columns). From (20), the input and noise signals are seen to contribute to their values in a very different manner (linear with N for the noise). So, 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 eigen-parameter $\hat{\theta}_{Vi_{\min}}$, 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 :

$$\Delta\sigma_{i_{\min}}^2(N) := \sigma_{i_{\min}}^2(N) - \sigma_{i_{\min}}^2(N-1) < \eta_c \quad (25)$$

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 (20), η_c should be chosen at least of order $\eta_{i_{\min}}\sigma^2$.

We have tested this selection criterion on the simulated example using the system and model description of Section 1 and an input signal made of successive steps with $\eta_c = 0.035$. The original data set and the normalized squared singular values associated with its eigen-regressors are shown in Figure 6. Figure 7 displays the same variables after removal from the original data set of all the data satisfying (25). On the first hand, it is seen that the surviving data subset is 4 times shorter than the original one and that the index $i_{\min}$ of the

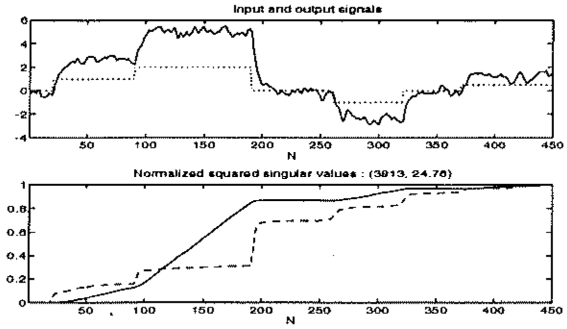

Figure 6: Input $(\cdots)$ and output $(—)$ signals and associated normalized eigen-regressor energies $\sigma_i^2(N)$.

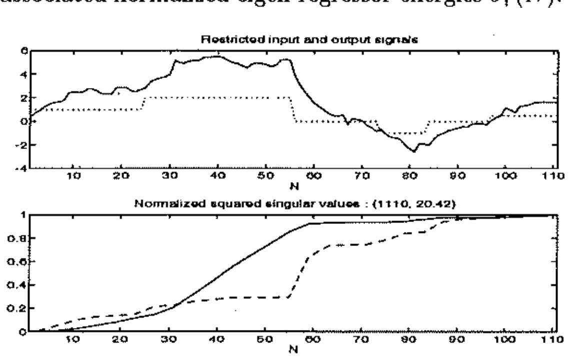

Figure 7: Restricted input $(\cdots)$ and output $(—)$ signals and associated normalized eigen-regressor energies $\sigma_i^2(N)$.

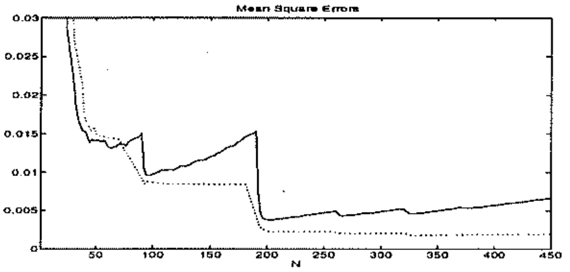

Figure 8: Mean Square Errors as a function of N using the full $(—)$ and the restricted $(\cdots)$ data set.

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 behaviour of the singular values are in both cases quite similar.
The bias and the variance of the estimated parameters have also been computed and are shown with further details in [1]. Regarding the overall parameter estimation accuracy, Figure 8 shows the MSE (24) 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 data (η_c not too big) and to discard as efficiently as possible the uninteresting ones (η_c not too small). The selection of a threshold η_c that is determined by the data themselves and that is robust with respect to prior assumption is the object of continuing researches [2]. Finally, from a practical viewpoint, this removal criterion is applied on-line along the identification data and provides an efficient way to focus on the interesting parts of the data set. This is in contrast with the classical forgetting factor introduced in recursive identification algorithms (see e.g. [6]) in order to enhance the contribution of the most recent data whatever their impact is on the global estimated parameter MSE. So, replacing this forgetting factor by our data selection criterion would lead to an improvement of the quality of the estimates of the I/O system dynamics in recursive identification.

7 Conclusions

We have exhibited a situation where the use of insufficiently 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. 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 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 data discarding procedure that eliminates data that cause a deterioration of the total MSE.

References

[1] P. Carrette, G. Bastin, Y. Genin and M. Gevers, *Discarding data may help in system identification!*, submitted to IEEE Transactions on Signal Processing.

[2] P. Carrette, G. Bastin, Y. Genin and M. Gevers, *Decrease of the MSE of estimated parameters in system identification via data selection*, submitted to the 34-th CDC, New Orleans, 1995.

[3] M. Gevers and L. Ljung, *Optimal Experiment Designs with Respect to the Intended Model Application*, Automatica, Vol. 22, N^0 5, 1986, pp 543-554.

[4] G.C. Goodwin and R.L. Payne, *Dynamic System Identification : Experiment Design and Data Analysis*, Mathematics in Science and Engineering, Academic Press, Vol. 136, 1977, pp 124-207.

[5] L. Ljung, *System Identification : theory for the user*, Prentice Hall, 1987.

[6] L. Ljung and T. Söderström, *Theory and practice of the recursive identification*, MIT Press, 1983.

[7] T. Söderström and P. Stoica, *System Identification*, Prentice Hall, 1990.

[8] G.W. Stewart and J.G. Sun, *Matrix Perturbation Theory*, Computer science and Scientific computing, Academic Press, Inc. San Diego, 1990.