PERGAMON

Automatica 37 (2001) 99–106

automatica

www.elsevier.com/locate/automatica

Brief Paper

An estimator of the inverse covariance matrix and its application to ML parameter estimation in dynamical systems☆

B. David*, G. Bastin

Center for Systems Engineering and Applied Mechanics, Université catholique de Louvain, Av. G. Lemaitre 4, B1348 Louvain-La-Neuve, Belgium

Received 19 November 1998; revised 4 February 2000; received in final form 29 May 2000

Abstract

An exact formula of the inverse covariance matrix of an autoregressive stochastic process is obtained using the Gohberg–Semencul explicit inverse of the Toeplitz matrix. This formula is used to build an estimator of the inverse covariance matrix of a stochastic process based on a single realization. In this paper, we show that this estimator can be conveniently applied to maximum likelihood parameter estimation in nonlinear dynamical system with correlated measurement noise. The efficiency of the estimation scheme is illustrated via Monte-Carlo simulations. It is shown that the statistical properties of the estimated parameters are largely improved using the proposed inverse covariance matrix estimator in comparison to the classical variance estimator. © 2000 Elsevier Science Ltd. All rights reserved.

Keywords: Parameter estimation; Maximum likelihood; Covariance matrix; Correlated residuals; Nonlinear system

1. Introduction

This paper deals with parameter estimation in a dynamical system. The system under consideration is described by a phenomenological model which is based on the prior knowledge of the physical phenomena that are supposed to take place in the system. The model is written under the form of a differential parametric deterministic state-space representation:

$$\dot{x} = f(x, \theta, u), \quad x \in \mathbb{R}^n, \ u \in \mathbb{R}^m, \ \theta \in \mathbb{R}^p, \tag{1}$$

where $x = [x_1, \ldots, x_n]^T$ is the state vector, $u = [u_1, \ldots, u_m]^T$ the input vector and $\theta = [\theta_1, \ldots, \theta_p]^T$ the vector of parameters. The parameter estimation problem is to estimate the parameter values from input and state data in the presence of additive correlated noise on the state measurements.

During the last 30 years, a number of publications were devoted to maximum likelihood (ML) estimation of system parameters. We should especially mention the work of Schoukens, Pintelon and coworkers who have deeply analyzed many aspects of ML estimation and in particular the role of the noise model. Most of their works deal with linear systems, represented by their transfer function, or particular forms of nonlinear systems, typically represented by Volterra series (Schoukens, Pintelon & Renneboog 1988). For instance, the influence of estimating the covariance matrix of the noise by its sample value is studied in Schoukens, Pintelon, Vandersteen and Guillaume (1997) and it has been recently shown (Schoukens, Pintelon & Rolain, 1999) that in certain case, the full covariance matrix can be replaced by its main diagonal. Even though the frequency domain approach allows an elegant treatment of the noise model in ML parameter estimation, it is obviously not applicable to estimate the parameters of a model that does not have a parametrized frequency domain representation which is the case of (1).

In this paper, we address the problem of estimating the parameters of (1) tacking into account the correlation that the output measurement noise may exhibit, using the ML framework in the time domain.

The problem is stated as follows. We assume that an experiment has been performed with a known input signal (often it is a piecewise constant signal) and that

☆This paper was not presented at any IFAC meeting. This paper was recommended for publication in revised form by Associate Editor S. Sagara under the direction of Editor Torsten Söderström.

* Corresponding author. Tel.: +32-10-472-382; fax: +32-10-472-180.

E-mail address: david@auto.ucl.ac.be (B. David).

PII: S0005-1098(00)00127-8

measurements of the state have been recorded at evenly distributed time instants $t_1, \dots, t_N$. The measurements are denoted $y(t_1), \dots, y(t_N)$ with $y(t_j) = [y_1(t_j), \dots, y_n(t_j)]^T$. For simplicity, let us assume that the initial value of the state $x(t_0)$ is exactly known to the user without being measured. Although this assumption is quite restrictive, it may be realistic in situations where the initial state can be physically constrained to be zero without any measurement being necessary (for instance zero initial concentrations in a chemical reactor or zero initial velocities for a mechanical device). In other situations, this restriction can be easily relaxed by incorporating the unknown initial state in the parameter vector.

For a given input signal $u(t)$ and a given initial state $x(t_0)$, the solution of the differential system (1) is parametrized by the parameter vector θ and denoted $x(t, \theta)$. There is always a deviation between the state $x(t_j, \theta)$ computed with the model and the measurement $y(t_j)$. The origin of this deviation may be multiple: modeling errors, input or process noise and measurement noise. It is usually called an output error and we denote it as

$$w(t_j, \theta) = y(t_j) - x(t_j, \theta), \quad j = 1, \dots, N.$$

The ML method relies on a representation of the sequence of output errors $w(t_j, \theta)$ at the sampling instants as a realization of a stochastic process. The optimal parameter estimate $\hat{\theta}$ is then defined as the parameter vector which maximizes the *likelihood* that $y(t_j)$ be a measurement of $x(t_j, \theta)$ in a technical sense that will be explained later. In this setup, the estimate $\hat{\theta}$ is viewed as a realization of a random vector. This allows to define the covariance matrix of the estimator, $C_{\hat{\theta}} = E\{(\hat{\theta} - E(\hat{\theta}))(\hat{\theta} - E(\hat{\theta}))^T\}$, which can be used to compute confidence intervals of the parameter estimates. Moreover, if a *true* physical parameter value θ^* is postulated, the bias of the estimator $b_{\hat{\theta}} = \theta^* - E(\hat{\theta})$ can also be used as a conceptual tool to assess the validity and quality of the model.

In practice, the output error vectors $w(t_j, \theta)$ may be highly correlated. There exist quite popular tools for the linear discrete time case (see e.g. Ljung, 1987 or Söderström & Stoica, 1989) that take into account a correlation of the output errors by estimating a noise filter along with the model of the system. But it is fair to say that most engineering applications of parameter estimation techniques to *physical* nonlinear state-space models do not take into account that correlation.

In principle, it is an evidence that the ML approach should allow, also for nonlinear systems, to account for the correlation of the output errors. However, a basic apparent difficulty lies in the need for the computation and the numerical inversion of a very big covariance matrix. The main contribution of this paper is to show that this matrix inversion may be completely avoided and replaced by an appropriate filtering of the output errors. The method relies on the Gohberg–Semencul formula for the inverse of a Toeplitz matrix. It is furthermore illustrated with Monte-Carlo simulations that this method widely improves the statistical properties of the parameter estimates even when the true system does not belong to the model set.

The paper is organized as follows. The next section briefly summarizes the ML parameter estimation technique and its statistical properties. The explicit formula of the inverse covariance matrix (ICM) of an autoregressive (AR) process is given in Section 3 and an ICM estimate is subsequently proposed. Its application to ML parameter estimation leads to a two-step estimation algorithm that is described in Section 4. Two Monte-Carlo simulations illustrate the efficiency of the estimation scheme in Section 5.

2. Maximum likelihood parameter estimation

The vector of the parameter-dependent state variable $x_i(t,\theta)$ evaluated at the sampling instants $t_j = t_1, \dots, t_N$ is denoted $x_i(\theta) = [x_i(t_1, \theta), \dots, x_i(t_N, \theta)]^T$, $i = 1, \dots, n$. The vector of the measurements of the state variable x_i at the sampling instants is denoted $y_i = [y_i(t_1), \dots, y_i(t_N)]^T$. Similarly, the vector of output errors $w_i(\theta)$ is defined by $w_i(\theta) = y_i - x_i(\theta)$.

The estimation of the parameter vector θ will result from the maximization of a function describing the *likelihood* that the n vectors y_i are measurements of the n vectors $x_i(\theta)$. In order to build this likelihood function, it is assumed that $w_i(\theta^*)$ is a random vector that takes values in $\mathbb{R}^N$ with a zero mean normal probability density function (PDF) given by (see e.g. Brown, 1983)

$$f_{w_i} = \frac{1}{\sqrt{(2\pi)^N |\Sigma_i|}} e^{-(1/2) w_i(\theta^*)^T \Sigma_i^{-1} w_i(\theta^*)}, \tag{2}$$

where Σ_i is the covariance matrix of the random vector $w_i(\theta^*)$:

$$\Sigma_i = E\{(w_i - E(w_i))(w_i - E(w_i))^T\}.$$

Since θ^* is unknown, (2) is considered, in the ML framework (Ljung, 1987), as a function of the parameter θ to be estimated.

We make the additional realistic assumption that all w_i vectors are independent of each other, i.e.

$$E\{w_{i_1}(t_{j_1}) w_{i_2}(t_{j_2})\} = 0 \quad \forall j_1, j_2 \text{ and } \forall i_1 \neq i_2.$$

This means that the measurements of each state variable are supposed to be corrupted by different and independent stochastic processes. Still each process may be highly correlated in time, this correlation being nested in the Σ_i matrices. It is essential to emphasize here that we are considering time auto-correlation of each w_i but no cross-correlation between the w_i vectors.

This assumption allows us to compute readily the joint probability density for all y_i vectors, as a function

of θ:

$$f_{y_1,\dots,y_n}(\theta)=\prod_{i=1}^{n}\frac{1}{\sqrt{(2\pi)^N|\Sigma_i|}}\mathrm{e}^{-1/2(y_i-x_i(\theta))^{\mathrm{T}}\Sigma_i^{-1}(y_i-x_i(\theta))}. \quad (3)$$

This deterministic function of θ is called the *likelihood function* and is proportional to the probability that the n vectors y_i are measurements of the n vectors $x_i(\theta)$. The estimation of the parameter vector θ results from the maximization of this function. Considering its logarithm, we finally end up to the following equivalent minimization problem:

$$\hat{\theta}^{ML}=\arg\min_{\theta} J(\theta), \quad (4)$$

$$J(\theta)=\sum_{i=1}^{n}(y_i-x_i(\theta))^{\mathrm{T}}\Sigma_i^{-1}(y_i-x_i(\theta)). \quad (5)$$

Under mild assumptions (Walter & Pronzato, 1997), the maximum likelihood estimate has the following appealing asymptotic properties:

- it is *unbiased*: $E(\hat{\theta})=\theta^*$;
- it is *consistent*: $\forall\delta>0$, $\mathrm{prob}(\|\hat{\theta}-\theta^*\|\geq\delta)\to 0$ as $N\to\infty$;
- it is *asymptotically efficient*: there is no consistent unbiased estimator with a smaller covariance as $N\to\infty$;
- it is *asymptotically Gaussian*: the distribution of $\hat{\theta}$ converges to a normal distribution with a mean equal to θ^* and a covariance matrix given by the Cramér–Rao bound that is also the inverse of the Fisher information matrix $C_{\theta}^{N\to\infty}=M^{-1}$ with

$$M=-E\frac{\partial^2}{\partial\theta^2}\ln f_{y_1,\dots,y_n}(\theta)\Big|_{\theta=\theta^*}. \quad (6)$$

One can easily verify that the gradient of the cost function $J(\theta)$ and the Cramér–Rao bound are given by

$$\frac{\partial J(\theta)}{\partial\theta}=-2\sum_{i=1}^{n}(y_i-x_i(\theta))^{\mathrm{T}}\Sigma_i^{-1}G_i, \quad (7)$$

$$M=\sum_{i=1}^{n}G_i^{\mathrm{T}}\Sigma_i^{-1}G_i\Big|_{\theta=\theta^*}, \quad (8)$$

where G_i are the state sensitivity matrices:

$$G_i=\frac{\partial x_i(\theta)}{\partial\theta}\in\mathbb{R}^{N\times p}$$

which can be computed by the integration of the following matrix differential equation (Walter & Pronzato, 1997):

$$\frac{\mathrm{d}}{\mathrm{d}t}G_i(t)=\frac{\partial f(x,\theta,u)}{\partial x}G_i(t)+\frac{\partial f(x,\theta,u)}{\partial\theta},\quad G_i(0)=0.$$

Let us remark at this stage that all expressions involving the ICM have the same generic form

$$L^{\mathrm{T}}\Sigma_i^{-1}R, \quad (9)$$

where $L=w_i$ in (5) and (7), $L=G_i$ in (8) and $R=w_i$ in (5), and $R=G_i$ in (7) and (8).

3. Estimator of the inverse covariance matrix

Let us consider a stable, stationary and scalar AR stochastic process $v(j)$ described by

$$v(j)=e(j)-a_1v(j-1)-\cdots-a_dv(j-d),$$

where $e(j)$ are independent identically distributed normal random variables with zero mean and variance σ^2. Let v be the random vector formed by N consecutive values of $v(j)$: $v=[v(1),\dots,v(N)]^{\mathrm{T}}$, and $\Sigma_v=E\{vv^{\mathrm{T}}\}$ its covariance matrix.

A classical way to obtain the covariance matrix would be to use the biased sample covariances to estimate the moments (which are the entries of the matrix), as proposed in Porat (1994):

$$\hat{m}_k=\frac{1}{N}\sum_{j=1}^{N-k}v(j)v(j+k).$$

Besides the need for inverting this matrix, a major drawback of that method comes from the fact that the estimation of the moments becomes very rough as k increases because of the decreasing number of terms in the average. Actually, one may reasonably use this only for $k\ll N$. Yet all moments are required to build the covariance matrix we are interested in. This limitation can be overcome with the ICM estimate we are now going to detail.

This estimate is based upon the explicit formula of the ICM of an AR stochastic process. This formula directly follows from a more general result on Toeplitz matrices obtained by Gohberg and Semencul (1972).

Lemma 1. *The ICM of the random vector v is given by*

$$\Sigma_v^{-1}=\frac{1}{\sigma^2}[U^{\mathrm{T}}U-V^{\mathrm{T}}V], \quad (10)$$

where U and V are N^2 Toeplitz lower triangular matrices of the form

$$U=\begin{bmatrix}1&&&&&&\\a_1&\ddots&&&&&\\&\ddots&\ddots&&&&\\a_d&&\ddots&\ddots&&&\\0&\ddots&&\ddots&\ddots&&\\&\ddots&\ddots&&\ddots&\ddots&\\0&&0&a_d&&a_1&1\end{bmatrix},$$

$$V=\begin{bmatrix}0&&&&&&\\&\ddots&&&&&\\0&&\ddots&&&&\\a_d&\ddots&&\ddots&&&\\&\ddots&\ddots&&\ddots&&\\&&\ddots&\ddots&&\ddots&\\a_1&&&a_d&0&&0\end{bmatrix}. \quad (11)$$

Proof. Σ_v is a Toeplitz matrix whose kth diagonal element is equal to the second-order moment of $v(j)$ at lag k, denoted m_k:

$$m_k = E\{v(j)v(j+k)\}.$$

According to the stationarity hypothesis, $E\{v(j)\} = 0\ \forall j$ and m_k does not depend on j. Substituting $v(j+k)$ by $e(j+k) - \sum_{l=1}^{d} a_l v(j+k-l)$ in m_k, we easily see that the set of the N second-order moments m_k is the solution of the following system:

$$m_k = \delta(k)\sigma^2 - \sum_{l=1}^{d} a_l m_{|k-l|}, \quad k = 0, \ldots, N-1,$$

where $\delta(k)$ is the Kronecker operator ($\delta(0) = 1$, $\delta(k) = 0\ \forall k \neq 0$). These equations are called the Yule–Walker equations[1] and can also be written under the following matrix form:

$$\Sigma_v \begin{bmatrix} 1 \\ \alpha_1 \\ \vdots \\ \alpha_{N-1} \end{bmatrix} = \begin{bmatrix} \sigma^2 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad \begin{cases} \alpha_k = a_k, & k \leq d, \\ \alpha_k = 0, & k > d. \end{cases}$$

This particular form allows to apply the following matrix equality that directly provides the ICM (see Gohberg & Semencul, 1972 or Kailath, Vietra & Mort, 1978):

$$\sigma^2 \Sigma_v^{-1} = U^{\mathrm{T}}U - V^{\mathrm{T}}V$$

where U and V are given by (11). □

This result allows to compute analytically the ICM of an AR process from the knowledge of its parameter vector $a = [a_1, \ldots, a_d]^{\mathrm{T}}$ and σ^2. The resulting ICM is the same as the inverse of the matrix one would obtain by solving the Yule–Walker equations for m_k unknown, but it is worth noting that with (10) we do not have to compute the inverse of a N^2 matrix. This is of great interest here since the size of the ICM could be very large in practice.

Suppose now we wish to compute Σ_v^{-1} from the knowledge of one single realization of v, the parameters a and σ^2 being unknown. The idea is to compute first an estimate of a, σ^2 from the realization of the process using, for instance, the very classical minimum variance estimate (see e.g., Söderström & Stoica, 1989 or Rosenblatt, 1985), and to plug it into (10) to compute the ICM estimate. With a slight abuse of notation, this estimator may be written as

$$\hat{\Sigma}_v^{-1} = \Sigma_v^{-1}(\hat{a}, \hat{\sigma}^2) = \frac{1}{\hat{\sigma}^2}[U(\hat{a})^{\mathrm{T}}U(\hat{a}) - V(\hat{a})^{\mathrm{T}}V(\hat{a})]. \tag{12}$$

In order to implement the estimator in (12), the order d of the AR model must also be estimated. The standard model order selection strategy based upon the Rissanen's minimum description length criterion (also known as the Akaike's Bayesian information criterion), has been used here. Other strategies could be investigated as well (see e.g. Porat, 1994), and their effect on the quality of the estimated ICM could be studied, but this is beyond the scope of this paper. More extensive analyses related to the particular case of AR process model order selection can be found in Dickie and Nandi (1994).

Let us now look deeper into the particular form of (12). As already mentioned, the ICM appears at several levels of the ML estimation procedure. It appears in the cost function to be minimized, in the gradient of that cost function which has to be evaluated if a gradient search algorithm is used for the optimization and finally in the Cramér–Rao bound that is often used to compute a confidence region around the estimated parameter. In each case, the expression involving the ICM takes the general form (9) that can be rewritten using (12):

$$\frac{1}{\hat{\sigma}_i^2}[L_U^{\mathrm{T}}R_U - L_V^{\mathrm{T}}R_V], \tag{13}$$

where $L_U = U(\hat{a}_i)L$, $R_U = U(\hat{a}_i)R$, $L_V = V(\hat{a}_i)L$ and $R_V = V(\hat{a}_i)R$. From (11) it follows that L_U and R_U are made of the columns of L and R passed through the filter $A_i(z^{-1})$ while L_V and R_V are made of the first d rows of the columns of L and R passed through the filter $\tilde{A}_i(z^{-1})$, with

$$A_i(z^{-1}) = 1 + \hat{a}_{i,1}z^{-1} + \cdots + \hat{a}_{i,d}z^{-d},$$

$$\tilde{A}_i(z^{-1}) = \hat{a}_{i,d} + \hat{a}_{i,d-1}z^{-1} + \cdots + \hat{a}_{i,1}z^{-d+1}.$$

In practice, in order to evaluate (9), it appears that we have just to apply appropriate filters, $A_i(z^{-1})$ and $\tilde{A}_i(z^{-1})$, to the columns of L and R and compute (13). Hence, only the parameters of the AR filters, $\hat{a}_i$ and $\hat{\sigma}_i^2$ have to be carried along, the big ICM does not need to be formed explicitly. An estimate of the ICM such as (12) would therefore be conveniently used for ML parameter estimation. This leads us to propose a two-step ML estimation scheme that will be detailed in the next section.

4. Estimation algorithm

The idea is to first estimate AR models for the residual vectors w_i obtained in a preliminary step from a rough

[1] The Yule–Walker equations are commonly used to estimate the parameters a_l from an estimate of a portion (m_k: $k = 0, \ldots, d$) of the covariance matrix. Here, the parameters a_l are supposed to be known and the Yule–Walker equations are used to provide the relationship between Σ_v and a_l.

estimate of θ. To achieve this first step, one could for instance use a weighted-least-squares (WLS) estimate that does not require the covariance matrices to be known:

$$\hat{\theta}^{\mathrm{WLS}} = \arg\min_{\theta} \sum_{i=1}^{n} \zeta_i \|y_i - x_i(\theta)\|^2. \tag{14}$$

The weights ζ_i are used here to normalize the residuals in order to balance the individual contribution of each state variable in (14). These weights can be determined from a preliminary data analysis. For instance, one could take them inversely proportional to the variances of the state measurements.

The proposed algorithm for ML parameter estimation in dynamical systems is then as follows:

(1) obtain ζ_i from data analysis,
compute a preliminary WLS estimate of θ with (14),
(2) compute the residuals: $w_i(\hat{\theta}^{\mathrm{WLS}}) = y_i - x_i(\hat{\theta}^{\mathrm{WLS}})$,
estimate an AR model for each w_i: $\hat{a}_i$, $\hat{\sigma}_i^2$,
compute the ML estimate $\hat{\theta}^{\mathrm{ML}}$ with (4).

The method of using an AR model of the residuals from a preliminary estimate came out firstly in nonlinear regression for static function fitting (see Gallant & Goebel, 1976; Glasbey, 1979, 1980 or Seber & Wild, 1989). Our algorithm can be seen as an extension of this method to parameter estimation in dynamical systems with a particular ICM estimate that fits nicely in the ML framework.

One might go further and suggest an iterative algorithm. That is to iterate several times step (2) with $\hat{\theta}^{\mathrm{WLS}}$ replaced by $\hat{\theta}^{\mathrm{ML}}$ of the previous iteration. However, there is no guarantee of convergence of such an algorithm in general. Moreover, some experimental tests have shown that the improvement by an additional iteration (i.e. a third step) is not very significant in practice. The extra computational cost induced by each new iteration is therefore not worthwhile. In Gallant and Goebel (1976) and Seber and Wild (1989) a third step is nevertheless suggested for small data sets.

5. Simulation results

In this section, we illustrate the benefits of the proposed algorithm on the statistical properties of the estimator. A linear case is studied in the first part of this section while the second part deals with a nonlinear case. In both parts, we have applied the algorithm on a two-parameter estimation problem. For comparison purpose, we have also recorded the parameters estimated by a modified two-step algorithm where our full ICM estimate is replaced by the classical sample variance estimate. The modified algorithm has the same preliminary step while the second step is as follows:

(2) compute the residuals: $w_i(\hat{\theta}^{\mathrm{WLS}}) = y_i - x_i(\hat{\theta}^{\mathrm{WLS}})$,
compute the sample variance of each w_i: $\hat{\sigma}_i^2$,
compute the refined WLS estimate $\hat{\theta}^{\mathrm{WLS}}$ with (14) where $\zeta_i = 1/\hat{\sigma}_i^2$.

5.1. Linear case

The considered dynamical model is a two-dimensional linear system with one input:

$$\dot{x}_1 = f_1(x, \theta, u) = -\theta_1 x_1 + \theta_2 u,$$
$$\dot{x}_2 = f_2(x, \theta, u) = x_1 - \theta_1 x_2. \tag{15}$$

The trajectories of this dynamical system have been computed for a *true* parameter θ^* equal to $[0.2, 1]^{\mathrm{T}}$ and zero initial states. The input u is a piecewise constant signal defined over a period of 100 s. The trajectories have been sampled at 1 Hz ($t_j = 0, \ldots, 99$; $N = 100$) to form the x_i vectors. Note the rather small value of N: the purpose is to show that the method is efficient even when only few measurements are available.

A bunch of realizations of stochastic processes that we are going to detail in the next paragraphs have been generated and added to x_i to form the measurements y_i. Two cases of stochastic uncertainty processes have been investigated. The objective of the first case is to illustrate the behavior of the algorithm when the stochastic part of the system belongs to the model set while the second case will illustrate the robustness of the algorithm when the stochastic part is not in the model set. In both cases, there are no structural errors on the deterministic part of the system.

(a) AR *stochastic processes*: The noises $w_i^a(t_j)$ corrupting the state measurements are generated by AR processes with identically distributed normal random innovations $e_i^a(j)$. The complete system (deterministic and stochastic parts) can be said to belong to the model set scanned by the algorithm. The noise filters are described by

$$[1 - 0.5z^{-1} - 0.45z^{-2}]w_1^a(t_j) = e_1^a(j),$$
$$[1 - 0.35z^{-1} - 0.6z^{-2}]w_2^a(t_j) = e_2^a(j). \tag{16}$$

(b) FIR *stochastic processes*: In the second case, denoted $w_i^b(t_j)$, the noises are generated by finite impulse response (FIR) lowpass filters applied to uncorrelated random sequences $e_i^b(j)$ with uniform distribution. The stochastic part of this second system does not belong to the model set. The filter structure and the distribution of the input noise both differ from the framework described in Section 3. The exact form of the noise filters is

$$w_1^b(t_j) = B_1(z^{-1})e_1^b(j),$$
$$w_2^b(t_j) = B_2(z^{-1})e_2^b(j), \tag{17}$$

where $B_1(z^{-1})$ and $B_2(z^{-1})$ are 10th-order lowpass FIR digital filters using Hamming window and with cutoff frequencies equal to 0.1 and 0.08, respectively.

The variances of the uncorrelated random variables $e_i^{a,b}(j)$ are scaled in order to have signal-to-noise ratios equal to 15 dB. The input signal, the trajectories and one example of corrupted measurements under both cases of stochastic processes are illustrated in Fig. 1.

For each case, 10 000 independent realizations of the stochastic processes have been generated and added to x_i to form the measurements $y_i^{a,b}$. For each realization, the two-step WLS and ML estimates of θ have been computed. The very classical simplex method (Nelder & Mead, 1965) has been used for the minimization. This choice was motivated by its relative robustness for optimization in the presence of noise and its efficiency for small dimension problems, as stated in Parkinson and Hutchinson (1972). Tests have been done to ensure that it did not reach local minima. The huge number of realizations will allow us to represent accurately the empirical statistical distribution of the parameter estimates.

We will compare those distributions, in particular their variance, with the distribution of the *best possible* unbiased estimator, i.e. the one with the smaller covariance matrix $C_{\hat{\theta}}$. That is the estimator that would have been obtained if the covariance matrices Σ_i were exactly known. The Cramér–Rao lower bound allows us to compute the covariance matrix of this *best* estimator.

We have done the computation of the Cramér–Rao bounds M_a and M_b for (15) combined with the two cases of stochastic processes (16) and (17). The square root of the diagonal elements of M_a^{-1} and M_b^{-1} gives us the minimum standard deviations σ_{CR} of the estimated parameters that may be achieved.

The standard deviations of the 10 000 independent WLS and ML estimates of θ obtained for each case are given in Table 1, together with the minimum possible standard deviations. The quality of the ML estimator is measured with a function δ that indicates how its standard deviation is situated between the two-step WLS and minimum ones: $\delta = 100(\sigma_{WLS} - \sigma_{ML})/(\sigma_{WLS} - \sigma_{CR})$.

The distributions of the estimated parameters are illustrated in Fig. 2 where normal distributions with the minimum variances are also plotted. The empirical biases of the estimated parameters from both methods and for both cases of noise structure are almost equal to zero. The 10 000 identified orders are almost all equal to 2 for both filters of the AR case while they are distributed between 3 and 7 for both filters of the FIR case.

We observe that when the noise process is in the model set, the performances of the algorithm are excellent. The standard deviation of the ML estimator is very close to the *best possible* estimator. This proves the coherence of the method. In this ideal case, we should asymptotically converge to the Cramér–Rao bound. We see that nearly 90% of this bound is already achieved for N equal to 100 only! In the other case, supposed to represent a more practical situation (the actual noise structure is not in the model set), the estimator performances are still quite good as brought out by the δ indicators.

5.2. Nonlinear case

The considered model is inspired by biotechnology. It describes a microbial (x_2) growth reaction with substrate

Table 1
Standard deviations of $\hat{\theta}$ and δ indicators: linear case

	σ_{WLS}	σ_{ML}	σ_{CR}	δ (%)
$\hat{\theta}_1^{AR}$	0.0107	0.0062	0.0055	87
$\hat{\theta}_2^{AR}$	0.0803	0.041	0.036	89
$\hat{\theta}_1^{FIR}$	0.0055	0.0032	0.0024	77
$\hat{\theta}_2^{FIR}$	0.0434	0.0188	0.0095	73

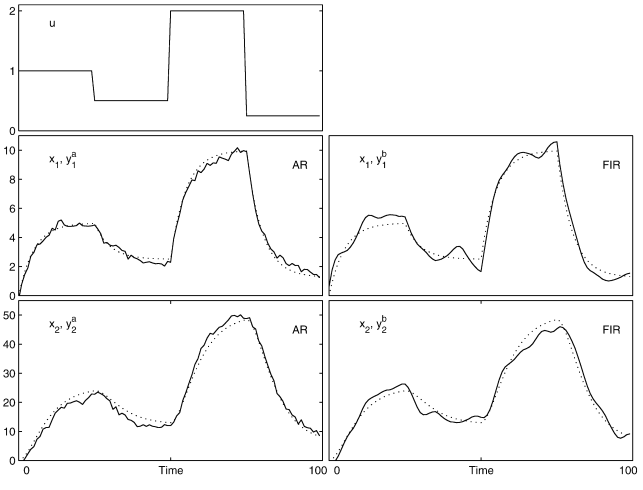

Fig. 1. Linear system: input signal u, trajectories x_i in dotted line and measurements y_i in plain line.

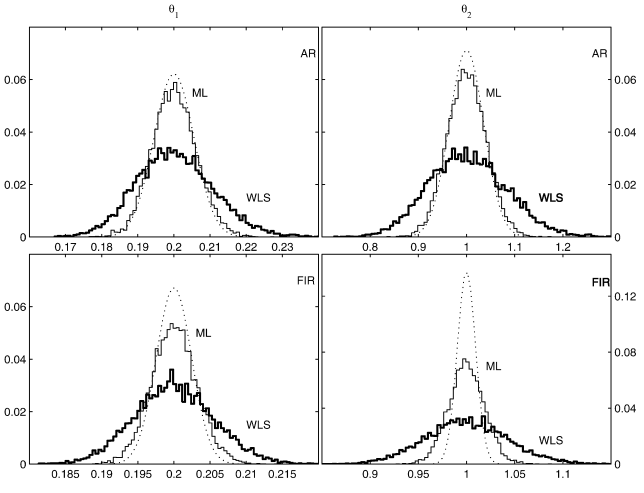

Fig. 2. Linear system: empirical distribution of $\hat{\theta}$ obtained from 10 000 simulations. Dotted lines represent the normal distributions corresponding to Cramér–Rao bounds.

(x_1) consumption in a continuous stirred tank, with Monod-like monotonic growth kinetics $k(x_1,\theta_1)$. The input is the dilution rate and the parameters to be estimated are the half saturation constant θ_1 and the inlet substrate concentration θ_2:

$$\begin{aligned}\dot{x}_1 &= f_1(x,\theta,u) = u(\theta_2 - x_1) - 10k(x_1,\theta_1)x_2,\\ \dot{x}_2 &= f_2(x,\theta,u) = -ux_2 + k(x_1,\theta_1)x_2,\end{aligned} \tag{18}$$

where

$$k(x_1,\theta_1) = \frac{x_1}{\theta_1 + x_1}.$$

The trajectories have been computed for a *true* parameter θ^* equal to $[5,10]^T$ and an initial state of $[0,0.1]^T$.

The same signal-to-noise ratio and exactly the same uncertainty structures (16) and (17) we have described for the linear case have been used here. Only 2000 Monte-Carlo simulations have been done due to the heavy computational cost required by the simulation of a nonlinear system. The input signal, the model trajectories and one example of measured trajectories are shown in Fig. 3.

Table 2 gives the standard deviations of the 2000 independent two-step WLS and ML estimates of θ, their corresponding minimum possible standard deviations and the δ indicators. The parameter distributions are shown on Fig. 4 along with the Cramér–Rao bounds. Because of the lesser number of realizations, the resolution of those distribution curves is lower than the one obtained for the linear case. However, they are still informative enough. Here again, the empirical biases are not significant and the identified orders are almost all equal to 2 in the AR case while they range between 3 and 6 for both filters of the FIR case.

We still observe good performances of the algorithm in this nonlinear case. When the stochastic part is in the model set, the estimate reaches about 75% of the bound which is quite satisfactory although a bit lower than the 90% we had for the linear case. When it is not in the model set, the algorithm manages to reduce the distance to the *best* estimate by more that an half which is still valuable. It is not easy to interpret the improvement for the FIR case based on Fig. 4. Let us remind that the relative improvement, i.e. with respect to the best possible situation, must be considered.

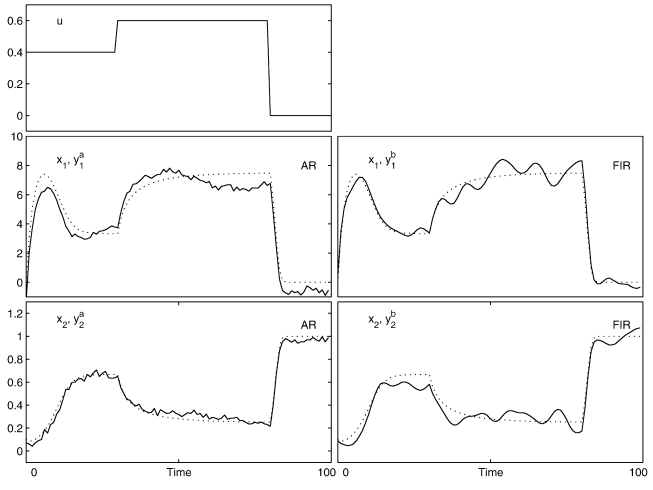

Fig. 3. Nonlinear system: input signal u, trajectories x_i in dotted line and measurements y_i in plain line.

6. Conclusion

An estimator of the inverse covariance matrix of an AR stochastic process has been proposed. It follows from the explicit formula that requires the polynomial coefficients of the AR filter and the innovation variance to be known. A minimum variance estimate of these parameters is used to build the ICM estimator from only one single realization of the process.

This ICM estimate has been applied to maximum likelihood parameter estimation in dynamical systems. Its efficiency has been emphasized from a case study involving linear and nonlinear systems. This study showed that when both deterministic and stochastic parts of the system belong to the model set, the statistical properties of the estimated parameters converge to the Cramér–Rao lower bound. It also appeared that the

Table 2
Standard deviations of $\hat{\theta}$ and δ indicators: nonlinear case

	σ_{WLS}	σ_{ML}	σ_{CR}	δ (%)
$\hat{\theta}_1^{AR}$	0.2231	0.1176	0.0897	79
$\hat{\theta}_2^{AR}$	0.3649	0.2275	0.1572	66
$\hat{\theta}_1^{FIR}$	0.1066	0.0686	0.0456	62
$\hat{\theta}_2^{FIR}$	0.1672	0.1243	0.0814	50

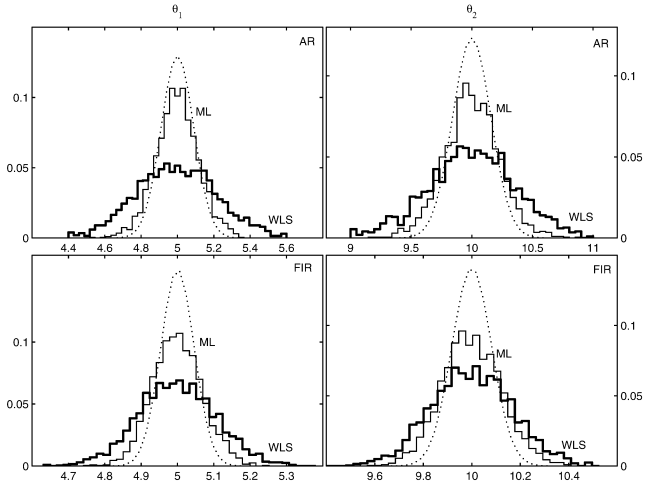

Fig. 4. Nonlinear system: empirical distribution of $\hat{\theta}$ obtained from 2000 simulations. Dotted lines represent the normal distributions corresponding to Cramér–Rao bounds.

algorithm is still valuable when the stochastic part of the system does not belong to the model set, i.e. when uncertainties cannot be exactly described by an AR process. For this part of the study, an uniformly distributed random noise passed through a FIR lowpass filter was used as measurement errors. The ICM estimated from AR models of the residuals yielded in this case an estimate of θ that was far better in term of its variance than the one obtained from classical diagonal covariance matrix estimate.

This seems to indicate that the ICM estimated from an AR representation of the uncertainty could be an efficient way to estimate the ICM of highly correlated stochastic process of other structures, at least for ML parameter estimation purpose. Combined with an appropriate AR model order selection strategy, the proposed estimation method should allow to improve the classical WLS one in most cases.

Acknowledgements

The authors would like to thank Yves Genin from the *CESAME* for his fruitful advice.

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

References

Brown, R. G. (1983). *Introduction to random signal analysis and Kalman filtering*. New York: Wiley.

Dickie, J. R., & Nandi, A. K. (1994). A comparative study of AR order selection methods. *Signal Processing*, *40*(2–3), 239–255.

Gallant, A., & Goebel, J. (1976). Nonlinear regression with autocorrelated errors. *Journal of American Statistical Association*, *71*(356), 961–967.

Glasbey, C. (1979). Correlated residuals in non-linear regression applied to growth data. *Applied Statistics*, *28*(3), 251–259.

Glasbey, C. (1980). Nonlinear regression with autoregressive times series errors. *Biometrics*, *36*, 135–140.

Gohberg, I. C., & Semencul, A. A. (1972). On inversion of finite-section Toeplitz matrices and their continuous analogues. *Matem. Issled.*, *7*, 201–224. (in Russian).

Kailath, T., Vieira, A., & Mort, M. (1978). Inverse of Toeplitz operators, innovations and orthogonal polynomials. *SIAM Review*, *20*, 106–119.

Ljung, L. (1987). *System identification: Theory for the user*. Englewood Cliffs, NJ: Prentice-Hall.

Nelder, J. A., & Mead, R. (1965). A simplex method for function minimization. *Computer Journal*, *7*, 308–313.

Parkinson, J. M., & Hutchinson, D. (1972). An investigation into the efficiency of variants on the simplex method. In F. A. Lootsma (Ed.), *Numerical methods for non-linear optimization* (pp. 115–135). New York: Academic Press.

Porat, B. (1994). *Digital processing of random signals: Theory and methods*. Englewood Cliffs, NJ: Prentice-Hall.

Rosenblatt, M. (1985). *Stationary sequences and random fields*. Basel: Birkhäuser.

Schoukens, J., Pintelon, R., & Renneboog, J. (1988). A maximum likelihood estimator for linear and nonlinear systems — a practical application of estimation techniques in measurement problems. *IEEE Transactions on Instrumentation and Measurement*, *37*, 10–17.

Schoukens, J., Pintelon, R., & Rolain, Y. (1999). Study of conditional ML estimators in time and frequency-domain system identification. *Automatica*, *35*, 91–100.

Schoukens, J., Pintelon, R., Vandersteen, G., & Guillaume, P. (1997). Frequency-domain system identification using non-parametric noise models estimated from a small number of data sets. *Automatica*, *33*(6), 1073–1086.

Seber, G., & Wild, C. (1989). *Nonlinear regression*. New York: Wiley.

Söderström, T., & Stoica, P. (1989). *System identification*. Englewood Cliffs, NJ: Prentice-Hall.

Walter, E., & Pronzato, L. (1997). *Identification of parametric models from experimental data*. Berlin: Springer.

Benoit David was born in Leuven, Belgium, in 1971. He received the electrical engineering degree in July 1996 from the Université Catholique de Louvain (UCL), Louvain-la-Neuve, Belgium. He is presently Ph.D. student in the field of parameter estimation at the Center for System Engineering and Applied Mechanics (CESAME), Belgium. His research interests are in nonlinear system identification and industrial application of control theory.

Georges Bastin received the electrical engineering degree and the Ph.D. degree, both from Université Catholique de Louvain, Louvain-la-Neuve, Belgium. He is presently Professor in the Center for Systems Engineering and Applied Mechanics (CESAME) at the Université Catholique de Louvain and Head of the Department of Mathematical Engineering. His main research interests are in system identification, nonlinear control theory, adaptive systems and random fields with applications to mechanical systems and robotics, biological and chemical processes, and environmental problems.