Proceedings of the 38th
Conference on Decision & Control
Phoenix, Arizona USA • December 1999

A Maximum Likelihood Parameter Estimation Method for Nonlinear 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.
Phone: +32-10-472382, Fax: + 32-10-472180,
Email: david@auto.ucl.ac.be

Abstract— **This paper presents an original method for maximum likelihood parameter estimation in nonlinear dynamical systems with highly correlated residuals. The method relies on an autoregressive representation of the residuals to build an estimate of the inverse of its covariance matrix. Theoretical concepts are developed and we provides a successful application of the method on a two-parameters estimation problem with data collected on a real plant. This experimental study shows that the statistical properties of the estimated parameters are significantly improved with our method in comparison to classical estimation techniques that usually rely on an uncorrelated representation of the residuals. In addition, a far better estimation of the confidence region around the parameter vector is obtained.**

Keywords— **Parameter Estimation; Nonlinear System; Maximum Likelihood; Correlated Residuals; Autoregressive.**

Notations— **PDF:** *Probability Density Function*, **WLS:** *Weighted Least Squares*, **MLH:** *Maximum Likelihood*, **ICM:** *Inverse Covariance Matrix*, **AR:** *Auto-Regressive.*

I. Introduction

The model of the system under consideration is written under the form of a differential parametric deterministic state-space representation of the form:

$$\begin{aligned} \dot{x}(t) &= f(x(t), \theta, u(t)) \\ x(t) \in I\!R^n, u(t) &\in I\!R^m, \theta \in I\!R^p \end{aligned} \tag{1}$$

where $x(t)=[x_1(t), x_2(t), \ldots, x_n(t)]^T$ is the state vector, $u(t)=[u_1(t), \ldots, u_m(t)]^T$ 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.

We assume that an experiment has been performed with a known input signal $u(t)$ and that measurements of the state $x(t)$ have been recorded at evenly distributed time instants $t_1, t_2, \ldots, t_N$. The assumption that all state variables are measured could be easily removed, it has been used here for simplicity. The measurements are denoted: $y(t_1), \ldots, y(t_N)$ with $y(t_j)=[y_1(t_j), \ldots, y_n(t_j)]^T$.

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 $x(t, \theta)$.

The inherent uncertainty of the model is reflected by the fact that, whatever the quality of the model, there is always a deviation between the state $x(t_j, \theta)$ computed with the model and the measurement $y(t_j)$. This deviation is called the modeling uncertainty and denoted:

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

A natural approach to solve the parameter estimation problem is evidently to select the parameter values in order to make the modeling uncertainty as small as possible. This may be achieved by minimizing a cost function which represents the average magnitude of the deviations $w(t_j, \theta)$ in a compact way. The most classical cost function in engineering studies is the least squares criterion:

$$J(\theta) = \sum_{j=1}^{N} \|y(t_j) - x(t_j, \theta)\|^2.$$

The best (or optimal) parameter estimate $\hat{\theta}$ is then defined as the value of θ which minimizes the cost function:

$$\hat{\theta} = \arg\min_{\theta} J(\theta).$$

This cost minimization approach has an important drawback: it does not allow to assess the quality of the model in any way. This drawback is removed with the maximum likelihood method which relies on a representation of the sequence of model uncertainty vectors $w(t_j, \theta)$ at the sampling instants as a realization of a stochastic process.

The situation is depicted as shown in Fig.1. The system under consideration is now represented by the additive combination of the deterministic state space model (1) and a stochastic uncertainty model. The optimal

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.

0-7803-5250-5/99/$10.00 © 1999 IEEE

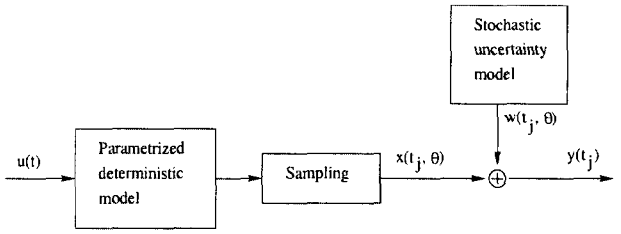
Stochastic uncertainty model
w(t_j, θ)
u(t)
Parametrized deterministic model
Sampling
x(t_j, θ)
y(t_j)

Fig. 1. Representation of the model uncertainty as a stochastic process

parameter estimate $\hat{\theta}$ is then defined as the parameter vector which maximizes the *likelihood* that $y(t_j)$ is a measurement of $x(t_j, \theta)$ in a technical sense that will be explained later on. In this set-up, the estimate $\hat{\theta}$ is viewed as a realization of a random vector. This allows us 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.

In practice, the uncertainty sequences $w(t_j, \theta)$ may be highly correlated. There exist quite popular tools for the linear discrete time case that take into account a correlation of the uncertainty part of the model, see e.g. [1] or [2]. However most of the parameter estimation techniques used in engineering for nonlinear continuous time system do not care about that correlation at all. In this paper we describe an estimation algorithm based on the maximum likelihood framework that allows us to take into account the way the model uncertainty vectors are correlated in time. An empirical Monte-Carlo study will show that the algorithm manages to reduce the variance of the estimated parameters and provides a better estimate of the confidence region around them.

The efficiency of the algorithm has already been revealed in [3] from a simulation case study, it is reinforced here from experimental results.

The paper is organized as follows. Section II briefly outlines the maximum likelihood parameter estimation technique and its statistical properties. The estimation algorithm developed in [3] and the estimate of the inverse covariance matrix the algorithm is based on are summarized in Section III. Section IV illustrates, from experimental data, the performances of the algorithm. Section V deals with confidence ellipsoid estimation and finally, conclusions are drawn in Section VI.

II. Maximum Likelihood Parameter Estimation

The vector of the parameter dependent state variables $x_i(t, \theta)$ evaluated at the sampling instants $t_j{=}t_1, \ldots, t_N$ is denoted $x_i(\theta){=}[x_i(t_1, \theta), \ldots, x_i(t_N, \theta)]^T$, $i{=}1, \ldots, n$. The vector of the measurements of the state variable $x_i(\theta)$ at the sampling instants is denoted $y_i = [y_i(t_1), \ldots, y_i(t_N)]^T$. Similarly the vector $w_i(\theta)$ 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, we assume that w_i is a random vector that takes values in $I\!R^N$ with a zero mean normal probability density function (PDF) given by, see e.g. [4]:

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

where Σ_i is the covariance matrix of the random vector w_i:

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

The random vector y_i has also a normal PDF with covariance matrix Σ_i and mean $x_i(\theta)$. We make the additional often realistic assumption that all w_i vectors are independent from each other, i.e.

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

That means that the measurements of each state variable are supposed to be corrupted by 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 ($i{=}1, \ldots, n$) to be observed, as a function of θ:

$$\begin{array}{l} f_{y_1,\ldots,y_n}(\theta) = \prod_{i=1}^n f_{w_i = y_i - x_i(\theta)} \\ = \prod_{i=1}^n \frac{1}{\sqrt{(2\pi)^N |\Sigma_i|}} e^{-\frac{1}{2}(y_i - x_i(\theta))^T \Sigma_i^{-1} (y_i - x_i(\theta))}. \end{array} \tag{3}$$

This function of θ is called the *likelihood function.* We seek the parameter vector $\hat{\theta}$ that maximizes this function. It is equivalent to maximize the *log-likelihood function* given by:

$$\begin{array}{l} \ln f_{y_1,\ldots,y_n}(\theta) = -\frac{nN}{2}\ln(2\pi) - \frac{1}{2}\sum_{i=1}^n \ln|\Sigma_i| \\ -\frac{1}{2}\sum_{i=1}^n (y_i - x_i(\theta))^T \Sigma_i^{-1} (y_i - x_i(\theta)). \end{array} \tag{4}$$

Maximizing the log-likelihood function amounts to minimizing the last term of (4) (with a reversed sign) since the other terms do not depend on θ. The estimation problem is therefore formulated as follows:

$$\hat{\theta} = \arg\min_{\theta} \sum_{i=1}^{n} (y_i - x_i(\theta))^T \Sigma_i^{-1} (y_i - x_i(\theta)). \tag{5}$$

Under mild assumptions, the maximum likelihood estimate has the following appealing asymptotic properties [5]: it is *asymptotically unbiased* ($E(\hat{\theta}){=}\theta^*$ where θ^* denotes the *true* value of θ), *consistent*, *asymptotically efficient* and *asymptotically Gaussian*. The latter implies that the distribution of $\hat{\theta}$ converges to a normal distribution with a covariance matrix given by the Cramér-Rao bound that is also the inverse of the Fisher information matrix: $C_{\hat{\theta}}^{N \to \infty} = M^{-1}$ with

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

Note that the inverse of the covariance matrix Σ_i^{-1} of each w_i vector has to be known to compute the estimate $\hat{\theta}$ in (5) and an asymptotic estimate of its covariance matrix $C_{\hat{\theta}}$ with (6). In practice, these matrices are not known and have to be estimated in one way or another. Often they are replaced by identity matrices which amounts to considering uncorrelated w_i vectors. The next Section will present an efficient way to build an estimator of the inverse covariance matrix Σ_i^{-1} (ICM) directly from a linear autoregressive (AR) model of the uncertainty process $w_i(t_j)$. The estimation algorithm based upon this ICM estimate will be described subsequently.

III. Estimation Algorithm

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

$$v(j) = e(j) - a_1 v(j-1) - \dots - a_d v(j-d),$$

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 $v(j)$: $v = [v(1), \dots, v(N)]^T$, and $\Sigma_v = E((v - E(v))(v - E(v))^T)$ its covariance matrix.

The following result, proof is given in [3], provides an elegant analytical expression of the ICM of the random vector v. It is a straightforward application of the Gohberg-Semencul's matrix equality, see [6] or [7].

$$\Sigma_v^{-1} = \frac{1}{\sigma^2} \left[U^T U - V^T V \right] \qquad (7)$$

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

$$U = \begin{bmatrix} 1 & & & & & \\ a_1 & \ddots & & & & \\ \vdots & \ddots & \ddots & & & \\ a_d & & \ddots & \ddots & & \\ 0 & \ddots & & \ddots & \ddots & \\ \vdots & \ddots & \ddots & & \ddots & \ddots \\ 0 & \cdots & 0 & a_d & \cdots & a_1 \;\; 1 \end{bmatrix}, \quad V = \begin{bmatrix} 0 & & & & \\ \vdots & \ddots & & & \\ 0 & & \ddots & & \\ a_d & \ddots & & \ddots & \\ \vdots & \ddots & \ddots & & \ddots \\ a_1 & \cdots & a_d & 0 & \cdots \;\; 0 \end{bmatrix}. \qquad (8)$$

This is used to compute the ICM of an AR process from the knowledge of its parameter vector $a = [a_1, \dots, a_d]^T$ and its innovation variance σ^2.

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 first compute estimates of a, σ^2 and use them to compute the ICM. Let us use the very classical minimum variance estimate of a, σ^2 given by, see e.g. [1] or [8]:

$$\hat{a} = R^{-1} F, \quad \hat{\sigma}^2 = \frac{[\Psi \hat{a} - b]^T [\Psi \hat{a} - b]}{N - d}$$

with $R = \Psi^T \Psi$, $F = \Psi^T b$ and

$$\Psi = \begin{bmatrix} v(d) & \dots & v(1) \\ \vdots & & \vdots \\ v(N-1) & \dots & v(N-d) \end{bmatrix}, b = \begin{bmatrix} -v(d+1) \\ \vdots \\ -v(N) \end{bmatrix}.$$

The proposed estimator of the ICM of v is therefore given by (7) where σ^2 is replaced by $\hat{\sigma}^2$ and U and V are computed for $\hat{a}$. With a slight abuse of notation, this estimator may be written:

$$\hat{\Sigma}_v^{-1} = \Sigma_v^{-1}(\hat{a}, \hat{\sigma}^2) \qquad (9)$$

In order to implement the estimator in (9), 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, BIC) has been used here. Other strategies could be investigated as well, see e.g. [9], but this is beyond the scope of this paper.

As already mentioned, the computation of the MLH estimator of θ requires the n matrices Σ_i^{-1} to be known. The idea is thus to estimate first an AR model for each w_i sequence and then to compute an estimate of its ICM with (9). Obviously the n vectors w_i have to be known in order to compute the AR models. Since we do not know θ^* which is precisely the parameter vector we are looking for, w_i can only be obtained from a preliminary rough estimate $\hat{\theta}$. To achieve this one could for instance use a weighted least squares (WLS) estimator that does not require the covariance matrices to be known:

$$\hat{\theta}^{WLS} = \arg\min_{\theta} \sum_{i=1}^{n} \zeta_i \| y_i - x_i(\theta) \|^2 . \qquad (10)$$

The weights ζ_i are used here to normalize the residuals so that the influences of each state variable in (10) are balanced. Those weights can be determined from a preliminary data analysis. For instance, one could take them proportional to the variances of the state measurements.

From a computational point of view, it is not necessary to build the matrix $\hat{\Sigma}_i^{-1}$ explicitly because the scalar $w_i(\theta)^T \hat{\Sigma}_i^{-1} w_i(\theta)$ in (5) can be obtained easily by appropriate filterings of $w_i(\theta)$. Indeed, combining (7) and (9), $w_i(\theta)^T \hat{\Sigma}_i^{-1} w_i(\theta)$ can be written:

$$\begin{aligned} \tfrac{1}{\hat{\sigma}_i^2} [\; & (U(\hat{a}_i) w_i(\theta))^T (U(\hat{a}_i) w_i(\theta)) \\ - \; & (V(\hat{a}_i) w_i(\theta))^T (V(\hat{a}_i) w_i(\theta))]. \end{aligned} \qquad (11)$$

Let us define $\tilde{w}_i(j, \theta) = w_i(j, \theta)$ for $j \leq d$ (0 otherwise) and the following discrete filters:

$$\begin{aligned} A_i(z^{-1}) &= 1 + \hat{a}_{i,1} z^{-1} + \dots + \hat{a}_{i,d} z^{-d}, \\ \tilde{A}_i(z^{-1}) &= \hat{a}_{i,d} + \hat{a}_{i,d-1} z^{-1} + \dots + \hat{a}_{i,1} z^{-d+1}. \end{aligned}$$

Considering the following filtered sequences and their associated vector:

$$\begin{aligned} r_i(j,\theta) &= A_i(z^{-1})w_i(j,\theta),\ r_i(\theta) = [r_i(1,\theta),\ldots,r_i(N,\theta)]^T, \\ \tilde{r}_i(j,\theta) &= \tilde{A}_i(z^{-1})\tilde{w}_i(j,\theta),\ \tilde{r}_i(\theta) = [\tilde{r}_i(1,\theta),\ldots,\tilde{r}_i(d,\theta)]^T, \end{aligned}$$

it becomes trivial to see from (8) that (11) is equivalent to:

$$\frac{1}{\hat{\sigma}_i^2}[r_i^T(\theta)r_i(\theta) - \tilde{r}_i^T(\theta)\tilde{r}_i(\theta)].$$

This allows us to rewrite (5) in the following elegant way:

$$\hat{\theta}^{MLH} = \arg\min_{\theta} \sum_{i=1}^{n} \frac{1}{\hat{\sigma}_i^2}(r_i(\theta)^T r_i(\theta) - \tilde{r}_i(\theta)^T \tilde{r}_i(\theta)) \quad (12)$$

where r_i is the a posteriori innovation of the AR process and $\tilde{r}_i^T\tilde{r}_i$ can be seen as a transient correction term.

Hence, the proposed algorithm for MLH parameter estimation in dynamical systems is as follows:

1. obtain ζ_i from data analysis,
2. compute a preliminary WLS estimate of θ with (10),
3. compute the residuals: $w_i(\hat{\theta}^{WLS}) = y_i - x_i(\hat{\theta}^{WLS})$,
4. estimate an AR model for each w_i: $\hat{a}_i$, $\hat{\sigma}_i^2$,
5. compute the MLH estimate $\hat{\theta}^{MLH}$ with (12).

This may be viewed as an extension to dynamical system of the algorithm suggested in [10] for static nonlinear regression.

One might go further and suggest an iterative algorithm. That is to plug $\hat{\theta}^{MLH}$ from step 5 into step 3 and iterate several times. 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 is not very significant in practice.

IV. Experimental Results

In this Section we illustrate with an empirical Monte-Carlo study, the benefits of the proposed algorithm on the statistical properties of the estimator. A two parameters estimation problem for a second order nonlinear dynamical model is treated. The data are collected on a real (laboratory) plant.

The plant, depicted in Fig.2, is a two-tank system consisting of two vertical plexiglas cylinders with similar cross sections of 154 cm^2. The two cylinders are connected together from the bottom by a cylindrical pipe. The first tank possess also an outflow pipe at its bottom and is fed with water from its top. The level measurements of the two tanks are carried out by piezo-resistive difference pressure sensors and are recorded in cm. These levels are the state variables while the input flow rate, in cm^3/s, is the command signal. The dynamical model of the system is as follows:

$$\begin{aligned} 154\,\dot{x}_1 &= u - q_1(x_1,\theta_1) - q_2(x_1 - x_2,\theta_2) \\ 154\,\dot{x}_2 &= q_2(x_1 - x_2,\theta_2) \end{aligned} \quad (13)$$

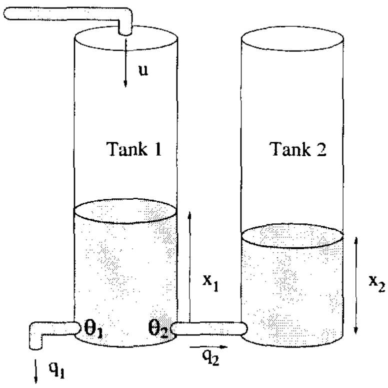

Fig. 2. Two-tank system. u: inlet flow rate, x_i: water levels, q_1: outlet flow rate, q_2: connection flow rate, θ_1: outlet flow characteristic and θ_2: connection flow characteristic.

where the outflow rate q_1 and the connection flow rate q_2 are modeled using Torricelli's rule:

$$\begin{aligned} q_1(x_1,\theta_1) &= \alpha_1 s_1\sqrt{2gx_1} \\ &= \theta_1\sqrt{x_1} \\ q_2(x_1 - x_2,\theta_2) &= \alpha_2 s_2 \mathrm{sgn}(x_1 - x_2)\sqrt{2g|x_1 - x_2|} \\ &= \theta_2 \mathrm{sgn}(x_1 - x_2)\sqrt{|x_1 - x_2|}, \end{aligned}$$

where g is the earth acceleration, s_1 and s_2 are the sections of the outflow and connection pipes, α_1 and α_2 are unknown correction dimensionless factors depending upon the flow properties. Those parameters are grouped together to form the parameters subject to identification: θ_1 and θ_2. In some sense, the parameters θ_1 and θ_2 describe the characteristic of the outflow and the connection pipes respectively.

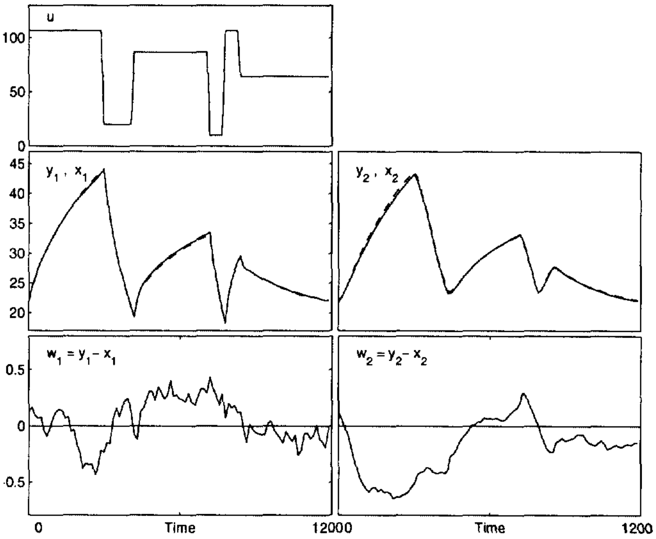

Fig. 3. Two-tank system. Top: inlet flow rate u $[cm^3/s]$; middle: measured levels y_i $[cm]$ in plain line and modeled trajectories x_i $[cm]$ in dashed line; bottom: correlated residuals w_i $[cm]$.

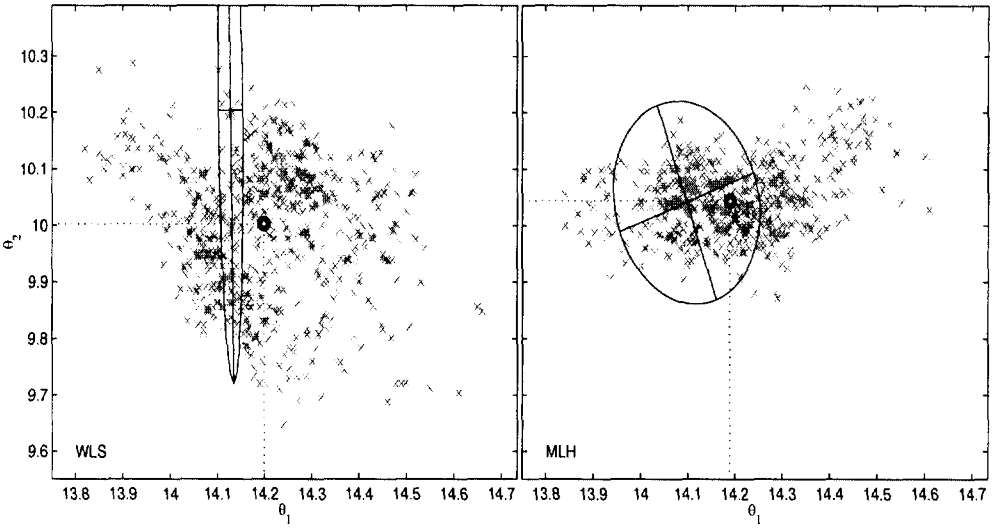

Fig. 4. WLS and MLH parameter estimates. The bold circles point the average values μ_{WLS} and μ_{MLH}. The confidence ellipses and their axes are drawn for one estimate chosen arbitrarily.

A series of 1000 similar experiments have been carried out on this system. The purpose is to perform a Monte-Carlo experimental case study. For each experiment, the system is excited with the same input signal u and the resulting level trajectories are recorded. The sampling period is 12 s and the duration of one experiment is 1200 s such that 100 data samples are available per experiment. The input signal u is designed to be sufficiently exciting for identification purpose. The initial state is about $[22, 22]^T$ for each experiment.

One estimate of θ has been computed for each experiment, using the algorithm described in Section III. The parameters estimated by WLS have been recorded as well for comparison purpose. Exhaustive search has been used for the optimization in order to ensure global minimization of the cost within a given bounded grid.

The input signal, one example of the measured trajectories and the corresponding estimated model trajectories $x_i(\hat{\theta}^{MLH})$ are illustrated in Fig.3. The resulting residuals w_i are represented as well. Let us notice that they exhibit a severe auto-correlation. Therefore, we may legitimately expect an improvement with MLH estimation method with respect to the WLS one.

The 1000 parameter estimates obtained by WLS and MLH are represented in Fig.4 while Table I gives their mean and standard deviation. The orders of the identified AR models are distributed between 1 and 5 for w_1 and between 2 and 6 for w_2 with the higher occurrence of order 3 for both variables.

TABLE I

Mean and standard deviations of $\hat{\theta}$.

	μ_{WLS}	μ_{MLH}	σ_{WLS}	σ_{MLH}
$\hat{\theta}_1$	14.20	14.19	0.1316	0.1201
$\hat{\theta}_2$	10.00	10.04	0.1116	0.0571

It clearly appears that the parameter estimates are much less scattered with the MLH estimator. The improvement may be quantified by computing the following ratio:

$$\frac{\sqrt{|C_{\hat{\theta}}^{MLH}|}}{\sqrt{|C_{\hat{\theta}}^{WLS}|}} = 0.45,$$

where $|C_{\hat{\theta}}^{WLS}|$ and $|C_{\hat{\theta}}^{MLH}|$ are the determinants of the empirical covariance matrices of the WLS and MLH estimates, i.e. computed from the 1000 realizations. This scalar indicator basically tells that the area of the cloud of points is approximately twice smaller in the MLH case.

V. Confidence Region

It is evident that the same experiment is rarely repeated 1000 times in order to have an idea of the parameter dispersion. Most often, only one experiment is carried out and a confidence region around the identified parameter vector has to be evaluated. The covariance matrix of one identified parameter vector, $C_{\hat{\theta}}$, is commonly estimated using its Cramér-Rao bound computed for $\hat{\theta}$ and $\hat{\Sigma}_i^{-1}$. Combining (6) and (3) one can easily verify that:

$$\hat{C}_{\hat{\theta}}^{-1} = \hat{M} = \sum_{i=1}^{n} \hat{G}_i^T \hat{\Sigma}_i^{-1} \hat{G}_i$$

where

$$\hat{G}_i = \left. \frac{\partial x_i(\theta)}{\partial \theta} \right|_{\theta=\hat{\theta}}$$

and $\hat{\Sigma}_i^{-1}$ is obtained by (9) in the MLH case and by $I_d/\hat{\sigma}_i^2$ in the WLS case, where $\hat{\sigma}_i^2$ is the a posteriori unbiased estimate of the noise variance:

$$\hat{\sigma}_i^2 = \frac{(y_i - x_i(\hat{\theta}^{WLS}))^T(y_i - x_i(\hat{\theta}^{WLS}))}{N-p}.$$

Let $S(t)$ be the state sensitivity matrix, defined as follows:

$$S(t) = \frac{\partial x(t,\theta)}{\partial \theta}.$$

$S(t)$ is a n by p time-dependent matrix that can be computed by the integration of the following matrix differential equation [5]:

$$\frac{d}{dt}S(t) = \frac{\partial f(x,\theta,u)}{\partial x}S(t) + \frac{\partial f(x,\theta,u)}{\partial \theta},\ S(0) = 0.$$

All G_i matrices can then be obtained from $S(t)$ evaluated at the sampling instants: $G_i = [s_i(t_1)^T, \ldots, s_i(t_N)^T]^T$ where $s_i(t_j)$ is the i'th row of $S(t_j)$.

The square-root of k'th diagonal element of $\hat{C}_{\hat{\theta}}$ gives an estimate of the standard deviation of $\hat{\theta}_k$, the k'th component of $\hat{\theta}$. Assuming a normal distribution of $\hat{\theta}$, an approximate $100(1-\alpha)\%$ confidence region for the parameter vector $\hat{\theta}$ is given by, see [11] or [1]:

$$\{\theta : (\theta - \hat{\theta})^T \hat{C}_{\hat{\theta}}^{-1} (\theta - \hat{\theta}) \leq p\mathcal{F}^{\alpha}_{p,N-p}\}, \qquad (14)$$

where $\mathcal{F}^{\alpha}_{p,N-p}$ is the upper α critical value of the $\mathcal{F}_{p,N-p}$ distribution. For α=0.01, p=2 and N=100, $\mathcal{F}^{\alpha}_{p,N-p} \approx 4.82$. The set described by (14) is the inner space of an ellipsoid centered on $\hat{\theta}$. The directions of its axes are the eigenvectors of $\hat{C}_{\hat{\theta}}^{-1}$ and the axes lengths are proportional to the eigenvalues.

In Fig.4 we have represented the 99% confidence ellipses of one estimate chosen arbitrarily. We clearly see that the confidence region estimated by MLH is better than the one estimated by WLS. It includes the mean point which could be considered as representing the *true* parameter. The shape of the ellipse estimated by WLS is far from what we could expect. Actually, the variance on the first parameter is quite underevaluated while the variance on the second one is a bit overevaluated.

The computation of $\hat{C}_{\hat{\theta}}$ has been done systematically for each parameter vector estimated in the previous Section. If $\hat{\theta}$ had really been normally distributed and the estimate of each confidence region had been exact, 99% of those regions would have included the mean. Since those conditions are far to be verified, we find that only 7.6% of the confidence regions include the mean in the WLS case. This percentage becomes 43.8% in the MLH case! This provides a quantitative evaluation of the improvement obtained in the confidence region estimation.

VI. Conclusion

An analytical solution has been given to directly compute the inverse of the covariance matrix of a random vector formed by N consecutive observations of an autoregressive stationary stochastic process. The computation 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 an estimator of the ICM that only requires one single realization of the AR process. An algorithm using the proposed ICM estimate has been subsequently proposed for maximum likelihood estimation of the parameters of a dynamical system from state measurement data.

The efficiency of the algorithm has been emphasized from an experimental Monte-Carlo study involving a nonlinear two-tank system. This study shows that the extent of the estimate distribution can be significantly reduced using the proposed algorithm. In addition to the improvement on the parameter variance itself, the method produces also a far better estimate of the confidence region around the estimated parameter.

As already mentioned, the idea of using AR model of the residuals is already used in nonlinear regression for static function fitting, see [10], [12], [13] or [11]. It is also a basic idea in linear system identification. It seemed thus natural to extend this idea to parameter estimation for nonlinear dynamical system.

References

[1] L. Ljung, *System Identification: Theory for the User*, Prentice Hall, 1987.

[2] T. Söderström and P. Stoica, *System Identification*, Prentice Hall, 1989.

[3] B. David and G. Bastin, "An estimator of the inverse covariance matrix and its application to ML parameter estimation in dynamical systems," Submitted for publication.

[4] R. G. Brown, *Introduction to Random Signal Analysis and Kalman Filtering*, John Wiley & sons, 1983.

[5] E. Walter and L. Pronzato, *Identification of Parametric Models from Experimental Data*, Springer, 1997.

[6] I. C. Gohberg and A. A. Semencul, "On inversion of finite-section toeplitz matrices and their continuous analogues," *Matem. Issled.*, vol. 7, pp. 201–224, 1972, (in Russian).

[7] T. Kailath, A. Vieira, and M. Mort, "Inverse of toeplitz operators, innovations and orthogonal polynomials," *SIAM Review*, vol. 20, pp. 106–119, 1978.

[8] M. Rosenblatt, *Stationary Sequences and Random Fields*, Birkhäuser, 1985.

[9] J. R. Dickie and A. K. Nandi, "A comparative study of AR order selection methods," *Signal Processing*, vol. 40, no. 2-3, pp. 239–255, 1994.

[10] A.R. Gallant and J.J. Goebel, "Nonlinear regression with autocorrelated errors," *J. Am. Stat. Assoc.*, vol. 71, no. 356, pp. 961–967, Dec. 1976.

[11] G.A.F. Seber and C.J. Wild, *Nonlinear Regression*, John Wiley & Sons, 1989.

[12] C.A. Glasbey, "Correlated residuals in non-linear regression applied to growth data," *Appl. Statist.*, vol. 28, no. 3, pp. 251–259, 1979.

[13] C.A. Glasbey, "Nonlinear regression with autoregressive time series errors," *Biometrics*, vol. 36, pp. 135–140, 1980.