PERGAMON

Automatica 38 (2002) 81–90

automatica

www.elsevier.com/locate/automatica

Parameter estimation in nonlinear systems with auto and crosscorrelated noise☆

B. David*, G. Bastin

Center for Systems Engineering and Applied Mechanics, CESAME, Université catholique de Louvain, Batiment Euler, 4-6, Avenue Georges Lemaitre, B1348 Louvain-La-Neuve, Belgium

Received 21 September 2000; revised 9 April 2001; received in final form 6 July 2001

Abstract

The Gohberg–Heinig explicit formula for the inversion of a block-Toeplitz matrix is used to build an estimator of the inverse of the covariance matrix of a multivariable autoregressive process. This estimator is then conveniently applied to maximum likelihood parameter estimation in nonlinear dynamical systems with output measurements corrupted by additive auto and crosscorrelated noise. An appealing computational simplification is obtained due to the particular form taken by the Gohberg–Heinig formula. The efficiency of the obtained estimation scheme is illustrated via Monte-Carlo simulations and compared with an alternative that is obtained by extending a classical technique of linear system identification to the framework of this paper. These simulations show that the proposed method improves significantly the statistical properties of the estimator in comparison with classical methods. Finally, the ability of the method to provide, in a straightforward way, an accurate confidence region around the estimated parameters is also illustrated.

Keywords: Parameter estimation; Nonlinear system; Maximum likelihood; Correlated noise; Block-Toeplitz

1. Introduction

In David and Bastin (2001), a maximum likelihood (ML) parameter estimation method for nonlinear systems was proposed for the case of autocorrelated output noise. It was assumed that all state variables are measured and that there does not exist any crosscorrelation between the noise sequences corrupting each state measurement. Basically, the method consisted of computing a preliminary weighted least-squares (WLS) estimate of the parameter vector, estimating the inverse of the noise covariance matrix from the residuals of this preliminary estimate and using this inverse covariance matrix (ICM) to compute an ML estimate. The originality of the method was in the particular estimate of the ICM. This estimate was derived from an explicit inversion formula for a Toeplitz matrix (Gohberg & Semencul, 1972) and required to identify an autoregressive (AR) model of the residuals. In David and Bastin (1999) the method was successfully applied to a real life application.

In this paper the method is generalized by removing two restrictive assumptions. First, the measured output variables are not necessarily the state variables and second, both the auto and the crosscorrelation of the noise sequences are taken into account. This leads to a more complicated block-Toeplitz structure of the covariance matrix. The generalization of the Gohberg–Semencul inversion formula to a block-Toeplitz matrix is therefore required in order to follow the same approach as in David and Bastin (2001). This formula is obtained as a particular case of Gohberg and Heinig (1974) and will require two multivariable AR models of the residuals to be identified, a causal and an anticausal one.

2. Problem formulation

It is assumed that a phenomenological model of the system under consideration is available to the user. The model is given under the quite general form of a set of

☆This paper was originally presented at the IFAC meeting SYSID 2000. This paper was recommended for publication in revised form by Associate Editor Johan Schoukens under the direction of Editor Torsten Söderström.

* Corresponding author. Tel.: +32-10-472382; fax: +32-10-472180.

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

0005-1098/02/$ - see front matter
PII: S0005-1098(01)00183-2

deterministic differential nonlinear state-space equations combined with a set of static nonlinear output equations. Both dynamic and static parts may be parametrized, the set of parameters being grouped into a single row vector $\theta = [\theta_1, \ldots, \theta_p]$. The model structure is as follows:

$$\dot{x}(t,\theta) = f(x(t),\theta,u(t)),$$

$$y(t,\theta) = g(x(t),\theta,u(t)), \tag{1}$$

where $x(t) = [x_1(t), \ldots, x_n(t)]^{\mathrm{T}}$ is the state vector, $u(t) = [u_1(t), \ldots, u_m(t)]^{\mathrm{T}}$ is the input vector and $y(t) = [y_1(t), \ldots, y_q(t)]^{\mathrm{T}}$ is the output vector.

The parameter estimation problem is to estimate the parameter values from input and output data obtained from a single experiment carried out on the system in presence of output additive correlated noise. The experiment is performed with a known input signal $u(t)$ and a known initial state $x(t_0)$. The measurements of the output variables are recorded at N evenly distributed time instants $t_j = t_1, \ldots, t_N$ and are denoted by

$$z(j) = [z_1(j), \ldots, z_q(j)]^{\mathrm{T}}.$$

For a given input signal and a given initial state, the solution of the state and output equations in (1) is parametrized by θ and denoted $x(t,\theta)$ and $y(t,\theta)$, respectively. The state sensitivity matrix, $\partial x(t,\theta)/\partial\theta$, is obtained by integrating the following matrix differential equation along with the system state equations (see Walter & Pronzato, 1997):

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial x(t,\theta)}{\partial\theta} = \frac{\partial f}{\partial x^{\mathrm{T}}}\frac{\partial x(t,\theta)}{\partial\theta} + \frac{\partial f}{\partial\theta}. \tag{2}$$

The output sensitivity matrix is derived from the state sensitivity matrix using

$$G(t,\theta) = \frac{\partial y(t,\theta)}{\partial\theta} = \frac{\partial g}{\partial x^{\mathrm{T}}}\frac{\partial x(t,\theta)}{\partial\theta} + \frac{\partial g}{\partial\theta}. \tag{3}$$

There is always a deviation between the model output evaluated at the sampling instants, $y(t_j,\theta)$, and the measurements $z(j)$. The origin of this deviation may be multiple: modeling error, input or process noise and measurement noise. It is usually called the output error or the residual and is denoted by

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

The sequence of output error vectors $w(j,\theta)$ is then viewed as a realization of a stationary zero mean stochastic process with covariance matrix at lag k defined by

$$\Sigma_k = E\{w(j)w(j-k)^{\mathrm{T}}\} \in \mathbb{R}^{q\times q}. \tag{4}$$

In most applications, the residual sequences corresponding to each output variable are assumed to be independent, i.e. not crosscorrelated, which amounts to considering a diagonal covariance matrix. Often also, each residual sequence is assumed to be uncorrelated along the time, i.e. not autocorrelated, which amounts to considering only Σ_0. In David and Bastin (2001), the case of autocorrelated residuals was treated only. Here the more general problem of auto and crosscorrelated residual sequences is addressed.

Let us define the following compact notations:

$$\mathbf{x}(\theta) = [x(t_1,\theta)^{\mathrm{T}}, \ldots, x(t_N,\theta)^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{R}^{nN},$$

$$\mathbf{y}(\theta) = [y(t_1,\theta)^{\mathrm{T}}, \ldots, y(t_N,\theta)^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{R}^{qN},$$

$$\mathbf{G}(\theta) = [G(t_1,\theta)^{\mathrm{T}}, \ldots, G(t_N,\theta)^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{R}^{qN\times p},$$

$$\mathbf{z} = [z(1)^{\mathrm{T}}, \ldots, z(N)^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{R}^{qN},$$

$$\mathbf{w}(\theta) = [w(1,\theta)^{\mathrm{T}}, \ldots, w(N,\theta)^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{R}^{qN},$$

$$= \mathbf{z} - \mathbf{y}(\theta).$$

The full covariance matrix, $\mathbf{\Sigma} = E\{\mathbf{w}(\theta)\mathbf{w}(\theta)^{\mathrm{T}}\}$, of the random vector $\mathbf{w}(\theta)$ is then a very large $(qN \times qN)$ block-Toeplitz matrix of the form

$$\mathbf{\Sigma} = \begin{bmatrix} \Sigma_0 & \Sigma_{-1} & & & \Sigma_{1-N} \\ \Sigma_1 & \cdot & \cdot & & \\ & \cdot & \cdot & \cdot & \\ & & \cdot & \cdot & \Sigma_{-1} \\ \Sigma_{N-1} & & & \Sigma_1 & \Sigma_0 \end{bmatrix}. \tag{5}$$

Assuming furthermore, a normal probability density function for $\mathbf{w}(\theta)$, the negative log-likelihood function takes the following form:

$$\frac{qN}{2}\log(2\pi) + \frac{1}{2}\log|\mathbf{\Sigma}| + \frac{1}{2}\mathbf{w}(\theta)^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{w}(\theta). \tag{6}$$

If $\mathbf{\Sigma}$ is known, only the last term of (6) depends on θ. The ML estimate of θ is given by maximizing the likelihood function which is equivalent to minimizing the last term of (6) (see e.g. Seber & Wild, 1989):

$$\hat{\theta}^{\mathrm{ML}} = \arg\min_{\theta} J(\theta), \tag{7}$$

where the scalar cost function is given by

$$J(\theta) = \mathbf{w}(\theta)^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{w}(\theta). \tag{8}$$

The gradient of the cost function and the Cramér–Rao (CR) lower bound on the covariance matrix of the estimator are given, respectively, by

$$\frac{\partial J(\theta)}{\partial\theta} = -2\mathbf{w}(\theta)^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{G}(\theta), \tag{9}$$

$$\Sigma_{\hat{\theta}^{\mathrm{ML}}} \geqslant [\mathbf{G}(\theta)^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{G}(\theta)]^{-1}. \tag{10}$$

It turns out that the ICM, $\mathbf{\Sigma}^{-1}$, is required at several levels of the computation of the ML estimate of θ. It appears in

the cost function (8) that has to be minimized, in gradient (9) of the cost function that is needed if a gradient search method is used to solve the nonlinear minimization problem (7) and finally, in the CR lower bound (10) which is commonly used to build a confidence region around the estimated parameter. This ICM is unknown, in practice, and has therefore to be estimated. An original solution to this problem is given in the next section.

3. Gohberg–Heinig inverse

In Gohberg and Semencul (1972) an explicit formula was derived for the inversion of a finite Toeplitz matrix. This result was generalized two years later (Gohberg & Heinig, 1974) for a Toeplitz matrix with entries belonging to a noncommutative algebra, the block-Toeplitz matrix being a particular case. Other formulas exist for this inversion problem (see e.g. Iohvidov, 1982), and several numerical methods to solve it have been developed and are still under investigation. It will be shown in this section that the Gohberg–Heinig formula can be very conveniently used in the ML parameter estimation framework.

Let us postulate a causal and an anticausal multidimensional AR representation of the residuals, defined by the following relations:

$$w(j) + \sum_{r=1}^{d} A_r^+ w(j-r) = \varepsilon^+(j),$$

$$w(j) + \sum_{r=1}^{d} A_r^- w(j+r) = \varepsilon^-(j), \tag{11}$$

where A_r^+ and A_r^- are q by q matrices containing the coefficients of the causal and anticausal AR models, $\varepsilon^+(j)$ and $\varepsilon^-(j)$ are zero mean i.i.d. random vectors with covariance matrix equal, respectively, to $\Sigma_{\varepsilon^+} = E\{\varepsilon^+(j)\varepsilon^+(j)^{\mathrm{T}}\}$ and $\Sigma_{\varepsilon^-} = E\{\varepsilon^-(j)\varepsilon^-(j)^{\mathrm{T}}\}$ while d is the order of the AR models with $d < N$.

The Yule–Walker equations generalized for the multidimensional case can be explicitly derived by substituting $w(j)$ in (4) by its expression coming from (11), taking into account the stationarity assumption. This provides the following sets of relations, for $k = 0, \ldots, N-1$:

$$\sum_{r=0}^{d} A_r^+ \Sigma_{k-r} = \delta_{0,k}\Sigma_{\varepsilon^+}, \quad A_0^+ = I_q,$$

$$\sum_{r=0}^{d} A_r^- \Sigma_{r-k} = \delta_{0,k}\Sigma_{\varepsilon^-}, \quad A_0^- = I_q. \tag{12}$$

The use of the Gohberg–Heinig formula requires first to find the coefficients that verify the so-called *generating system*. The inverse matrix is then obtained explicitly from these coefficients. In the general case, this system is constituted of four sets of equations. In the particular case of symmetric block-Toeplitz matrix, i.e. with blocks satisfying $\Sigma_{-k} = \Sigma_k^{\mathrm{T}}$, the *generating system* is reduced to two sets of equations and corresponds exactly to (12). Hence, for the inversion of the covariance matrix of an AR process, the coefficients of the causal and anticausal AR models are also the generating coefficients. Therefore Gohberg and Heinig (1974, Corollary 1.1) directly gives the following explicit expression for the ICM of $\mathbf{w}(\theta)$:

$$\boldsymbol{\Sigma}^{-1} = \mathbf{U}^{\mathrm{T}}\boldsymbol{\Sigma}_{\varepsilon^+}^{-1}\mathbf{U} - \mathbf{V}^{\mathrm{T}}\boldsymbol{\Sigma}_{\varepsilon^-}^{-1}\mathbf{V}, \tag{13}$$

where $\mathbf{U}$ and $\mathbf{V}$ are $qN \times qN$ lower triangular block-Toeplitz matrices of the form

$$\mathbf{U} = \begin{bmatrix} I_q & & & & & & \\ A_1^+ & \ddots & & & & & \\ \vdots & \ddots & \ddots & & & & \\ A_d^+ & & \ddots & \ddots & & & \\ 0 & \ddots & & \ddots & \ddots & & \\ \vdots & \ddots & \ddots & & \ddots & \ddots & \\ 0 & & 0 & A_d^+ & & A_1^+ & I_q \end{bmatrix},$$

$$\mathbf{V} = \begin{bmatrix} 0 & & & & & & \\ \vdots & \ddots & & & & & \\ 0 & & \ddots & & & & \\ A_d^- & \ddots & & \ddots & & & \\ \vdots & \ddots & \ddots & & \ddots & & \\ A_1^- & & \ddots & A_d^- & 0 & & 0 \end{bmatrix},$$

while $\boldsymbol{\Sigma}_{\varepsilon^+}^{-1}$ and $\boldsymbol{\Sigma}_{\varepsilon^-}^{-1}$ are $qN \times qN$ block-diagonal matrices of the form:

$$\boldsymbol{\Sigma}_{\varepsilon^+}^{-1} = \begin{bmatrix} \Sigma_{\varepsilon^+}^{-1} & & \\ & \ddots & \\ & & \Sigma_{\varepsilon^+}^{-1} \end{bmatrix}, \quad \boldsymbol{\Sigma}_{\varepsilon^-}^{-1} = \begin{bmatrix} \Sigma_{\varepsilon^-}^{-1} & & \\ & \ddots & \\ & & \Sigma_{\varepsilon^-}^{-1} \end{bmatrix}.$$

Assuming that the coefficients of the causal and anticausal AR models of the residual sequence are available, (13) provides a straightforward way to obtain the ICM required for the computation of the ML parameter estimate. In David and Bastin (2001), only the causal scalar AR model of each independent output noise sequence was necessary to build the ICM.

Besides the evident advantage that this ICM formula overcomes matrix inversion, the particular form of (13) provides also an appealing computational simplification. Indeed, in order to compute expressions involving the ICM like (8)–(10) that are of the form $\mathbf{L}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{R}$, one has just

to apply appropriate filters to the columns of **L** and **R**. Here follows a short proof.

Let $L(j)$ and $R(j)$ be any sequences of column vectors (matrices)[1] for $j = 1, \dots, N$ and let **L** and **R** be their corresponding vertically stacked vectors (matrices)

$$\mathbf{L} = [L(1)^{\mathrm{T}}, \dots, L(N)^{\mathrm{T}}]^{\mathrm{T}},$$

$$\mathbf{R} = [R(1)^{\mathrm{T}}, \dots, R(N)^{\mathrm{T}}]^{\mathrm{T}}.$$

Let us also define the following multidimensional discrete filters, associated with the AR models (11):

$$\mathscr{A}^{+}(z^{-1}) = I_q + A_1^{+} z^{-1} + \cdots + A_d^{+} z^{-d}$$

$$\mathscr{A}^{-}(z^{-1}) = A_d^{-} + A_{d-1}^{-} z^{-1} + \cdots + A_1^{-} z^{-d+1}$$

and let T_{ε^+} and T_{ε^-} denote the square root of the innovation covariance matrices: $\Sigma_{\varepsilon^+} = T_{\varepsilon^+} T_{\varepsilon^+}$, $\Sigma_{\varepsilon^-} = T_{\varepsilon^-} T_{\varepsilon^-}$. Then, considering the filtered versions of $L(j)$ and $R(j)$ through $T_{\varepsilon^+}^{-1}\mathscr{A}^{+}(z^{-1})$ and $T_{\varepsilon^-}^{-1}\mathscr{A}^{-}(z^{-1})$:

$$L_U(j) = T_{\varepsilon^+}^{-1}\mathscr{A}^{+}(z^{-1})L(j),$$

$$R_U(j) = T_{\varepsilon^+}^{-1}\mathscr{A}^{+}(z^{-1})R(j),$$

$$L_V(j) = T_{\varepsilon^-}^{-1}\mathscr{A}^{-}(z^{-1})L(j),$$

$$R_V(j) = T_{\varepsilon^-}^{-1}\mathscr{A}^{-}(z^{-1})R(j)$$

and their corresponding stacked matrices:

$$\mathbf{L}_{\mathbf{U}} = [L_U(1)^{\mathrm{T}}, \dots, L_U(N)^{\mathrm{T}}]^{\mathrm{T}},$$

$$\mathbf{R}_{\mathbf{U}} = [R_U(1)^{\mathrm{T}}, \dots, R_U(N)^{\mathrm{T}}]^{\mathrm{T}},$$

$$\mathbf{L}_{\mathbf{V}} = [L_V(1)^{\mathrm{T}}, \dots, L_V(d)^{\mathrm{T}}]^{\mathrm{T}},$$

$$\mathbf{R}_{\mathbf{V}} = [R_V(1)^{\mathrm{T}}, \dots, R_V(d)^{\mathrm{T}}]^{\mathrm{T}},$$

it becomes trivial to see, using the particular form of **U** and **V**, that:

$$\mathbf{L}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{R} = \mathbf{L}_{\mathbf{U}}^{\mathrm{T}}\mathbf{R}_{\mathbf{U}} - \mathbf{L}_{\mathbf{V}}^{\mathrm{T}}\mathbf{R}_{\mathbf{V}}. \tag{14}$$

The practical consequence of (14) is that to compute expressions involving the ICM like (8)–(10), one has just to apply appropriate filters, $\mathscr{A}^{+}(z^{-1})$ and $\mathscr{A}^{-}(z^{-1})$, to the columns or part of the columns of **L** and **R** and normalize with T_{ε^+} and T_{ε^-}. Hence, only the coefficients of the AR filters, A_r^{+}, A_r^{-}, Σ_{ε^+} and Σ_{ε^-} have to be carried along. The large ICM does not need to be formed explicitly. This is a major advantage that leads us to propose the following two-step ML estimation algorithm.

[1] $L(j) = w(j)$ in (8,9) and $L(j) = G(t_j)$ in (10), $R(j) = w(j)$ in (8) and $R(j) = G(t_j)$ in (9,10).

4. Estimation algorithm

The idea is to obtain a rough WLS estimate of θ in a first step and identify the causal and anticausal AR models on the residuals of this preliminary estimate. The anticausal model is simply obtained by presenting the residual vector in the reversed order to the AR model identification procedure. Then, these AR models are used to estimate the ICM needed to compute the ML estimate in a second step.

If we denote the estimates of the AR matrix coefficients as $\hat{A}^{+} = [\hat{A}_1^{+}, \dots, \hat{A}_d^{+}]$, $\hat{A}^{-} = [\hat{A}_1^{-}, \dots, \hat{A}_d^{-}]$, $\hat{\Sigma}_{\varepsilon^+}$ and $\hat{\Sigma}_{\varepsilon^-}$, then the ICM estimator may be written, according to (13), as

$$\hat{\boldsymbol{\Sigma}}^{-1} = \mathbf{U}(\hat{A}^{+})^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}_{\varepsilon^+}^{-1}\mathbf{U}(\hat{A}^{+}) - \mathbf{V}(\hat{A}^{-})^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}_{\varepsilon^-}^{-1}\mathbf{V}(\hat{A}^{-}),$$

while the cost function (8), can be rewritten using the computational simplification (14) as

$$\mathbf{w}(\theta)^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{w}(\theta) = \sum_{j=1}^{N} e^{+}(j,\theta)^{\mathrm{T}}\hat{\Sigma}_{\varepsilon^+}^{-1}e^{+}(j,\theta) - \sum_{j=1}^{d} e^{-}(j,\theta)^{\mathrm{T}}\hat{\Sigma}_{\varepsilon^-}^{-1}e^{-}(j,\theta),$$

where $e^{+}(j,\theta)$ and $e^{-}(j,\theta)$ are the residuals of the AR models, or say the posterior innovations, defined by

$$e^{+}(j,\theta) = \hat{\mathscr{A}}^{+}(z^{-1})(z(j) - y(t_j,\theta)),$$

$$e^{-}(j,\theta) = \hat{\mathscr{A}}^{-}(z^{-1})(z(j) - y(t_j,\theta)).$$

The cost function is thus made up of two terms, the first one is related to the normalized variance of the residuals of the causal AR model and the second, involving the anticausal model, may be viewed as a correction term that is necessary to make the ICM the inverse of a Toeplitz matrix, that is to say to consider a stationary noise sequence. Asymptotically, actually for $N \gg d$, the second term can be neglected. For a rather small number of data though, this correction term plays an important role as it will show up in the sequel.

The WLS preliminary estimate is defined by

$$\hat{\theta}^{\mathrm{WLS}} = \arg\min_{\theta} \sum_{j=1}^{N} w(j,\theta)^{\mathrm{T}}\zeta w(j,\theta), \tag{15}$$

where ζ is a q by q diagonal weighting matrix used to normalize the residuals in order to balance the individual contribution of each output variable. The weights can be determined from a preliminary data analysis. For instance, one could take them inversely proportional to the variance of the output measurements.

The proposed algorithm for ML parameter estimation in dynamical systems with auto and crosscorrelated output noise is then as follows:

1. • obtain ζ from data analysis,
   • compute a preliminary WLS estimate,
2. • compute the residuals $\mathbf{w}(\hat{\theta}^{\mathrm{WLS}})$,
   • estimate the AR models of $\mathbf{w}(\hat{\theta}^{\mathrm{WLS}})$,
   • compute the ML estimate.

Let us notice that the estimate outcoming from this algorithm is not strictly speaking the ML one since the true ICM is replaced by an estimated one. Nevertheless, for simplicity, it is still denoted ML in the following. This estimation algorithm may be extended to an iterative one by repeating several times step 2 with $\hat{\theta}^{\mathrm{WLS}}$ replaced by $\hat{\theta}^{\mathrm{ML}}$ of the previous iteration.

Naturally, the first thing to do in the computation of the ICM estimate is to select an appropriate order d for the AR models. A number of methods have been developed in the last three decades for that purpose (see e.g. Ljung, 1999). The two most common being the strategies based upon the Akaike's information criterion (Akaike, 1981) and the Rissanen's minimum description length (MDL) criterion (Rissanen, 1985). It is worth noting that several alternatives are extensively analyzed in Dickie and Nandi (1994) from simulation studies for the particular case of the AR model order selection. It is out of the scope of this paper to give a detailed description of each of the possible strategies and to analyze their effect on the estimate of the ICM.

5. An alternative method

By extending directly a method coming from the field of discrete time identification for linear systems (Ljung, 1999; Söderström & Stoica, 1989) to the framework of this paper, the following alternative approach can be considered. It consists of augmenting, from the beginning, the continuous time phenomenological model (1) with a discrete time AR noise model parametrized by ψ:

$$z(j) = y(t_j, \theta^*) + \frac{1}{\mathscr{A}(z^{-1}, \psi^*)}\varepsilon(j),$$

where θ^* and ψ^* denote the true parameter values and $\varepsilon(j)$ are i.i.d. random normal vectors with a covariance matrix equals to $\Sigma_\varepsilon \in \mathbb{R}^{q\times q}$. The complete model is therefore a mixture of continuous time and discrete time parts.

The vector of parameters is then augmented with the parameters of the noise model, namely ψ, and the maximization of the log-likelihood function is done with respect to the augmented parameter vector. That is, the phenomenological and the stochastic models are optimized at the same time.

In this context, the residuals of the overall model are given by

$$e(j, \theta, \psi) = \mathscr{A}(z^{-1}, \psi)(z(j) - y(t_j, \theta))$$

and the negative log-likelihood function takes the following form:

$$\frac{qN}{2}\log(2\pi) + \frac{N}{2}\log|\Sigma_\varepsilon| + \frac{1}{2}\sum_{j=1}^{N} e(j, \theta, \psi)^{\mathrm{T}}\Sigma_\varepsilon^{-1}e(j, \theta, \psi), \tag{16}$$

where Σ_ε is obtained, if unknown, from the residuals by

$$\hat{\Sigma}_\varepsilon = \frac{1}{N}\sum_{j=1}^{N} e(j, \theta, \psi)e(j, \theta, \psi)^{\mathrm{T}}. \tag{17}$$

If (17) is introduced into (16), the last term of the log-likelihood function becomes constant and the estimate is given by the solution of the following minimization problem:

$$\hat{\theta}^{\mathrm{AML}} = \arg\min_{\theta,\psi} \log\left|\sum_{j=1}^{N} e(j, \theta, \psi)e(j,\theta,\psi)^{\mathrm{T}}\right|. \tag{18}$$

For a fixed θ, the minimization of (18) with respect to ψ is a linear problem whose explicit solution can be directly plugged into (18) in order to end up with a nonlinear optimization related to θ only. Therefore, the computation load of (18) is comparable to the one of (7) or (15).

This method will be denoted as alternative maximum likelihood (AML) in the following. The question that naturally arises now is: How does this method differ from the algorithm of the previous section? Although the two methods could look similar at a first glance since both whiten the residuals with an AR filter, there exists three main differences.

The first one is in the cost functions to be minimized. Basically, $\hat{\theta}^{\mathrm{ML}}$ minimizes the last term of the negative log-likelihood (6), the second one being constant, while $\hat{\theta}^{\mathrm{AML}}$ maximizes the second term of the log-likelihood (16), the last one being constant. Moreover, as we have already mentioned, the ML approach contains a correction term in the cost function while the AML one does not.

Second, in the ML method the covariance matrix is estimated only once and it is kept constant during the optimization with respect to θ while in the AML method, the noise model is updated at each step of the optimization process. As a consequence, the ML method requires to solve two successive nonlinear optimization problems (or more if the algorithm is further iterated) while only one nonlinear optimization pass is necessary for the AML method.

Finally, both the causal and the anticausal AR models are required for the ML estimator while only the causal one is used in the AML estimator.

It is obvious that from a computational point of view, the AML method is to be preferred to the ML one. However, it will be shown in the next section that the ML estimate clearly outperforms, in the considered example, the AML one from the point of view of their statistical properties and their robustness against noise model misspecification.

6. Monte-Carlo simulations

The purpose of this section is to illustrate, from Monte-Carlo simulations on a particular example, the statistical properties of the ML estimator obtained with the proposed two-step algorithm and to compare it with the alternative approach. The iterative version of the algorithm will be also briefly analyzed.

A three-parameter estimation problem on a nonlinear fourth order dynamical system with two output measurements is treated. The model is as follows:

$$\dot{x}_1 = u - \theta_1 x_1,$$

$$\dot{x}_2 = -2x_2 + h_1(x_3, x_4, \theta_2),$$

$$\dot{x}_3 = x_1 - x_3 - h_2(x_2, x_3, \theta_3),$$

$$\dot{x}_4 = -x_4 + h_2(x_2, x_3, \theta_3),$$

$$y_1 = x_4,$$

$$y_2 = h_1(x_3, x_4, \theta_2),$$

where

$$h_1(x_3, x_4, \theta_2) = \frac{x_4}{1 + x_3/\theta_2},$$

$$h_2(x_2, x_3, \theta_3) = 4x_3 \mathrm{e}^{-\theta_3 x_2}.$$

The output trajectories of this model are computed for a true parameter $\theta^* = [8,5,10]$ and zero initial states. The experiment lasts 20 s. The input $u(t)$ is a piecewise constant signal and the output trajectories are sampled with a frequency of 10 Hz to form the $\mathbf{y}(\theta^*)$ vector containing $N = 200$ data samples.

Two different types of noise generator are used to produce the true output error sequences. The first type is a first order bidimensional AR stochastic process with i.i.d. normal random innovation vectors. The objective of this first structure is to illustrate the behavior of the algorithm when the stochastic part of the system belongs to the model set scanned by the algorithm. The structure of the process is

$$w^{\mathrm{AR}}(j) + \begin{bmatrix} -0.65 & -1.80 \\ -0.07 & -0.5 \end{bmatrix} w^{\mathrm{AR}}(j-1) = \varepsilon^{\mathrm{AR}}(j),$$

$$\Sigma_{\varepsilon}^{\mathrm{AR}} = 10^{-5} \begin{bmatrix} 90 & 3 \\ 3 & 2.5 \end{bmatrix}.$$

The goal of the second type of noise generator is to evaluate the robustness of the algorithm when the stochastic part of the system is not in the model set. For that case, a bidimensional finite impulse response (FIR) filter of the following form is used:

$$w^{\mathrm{FIR}}(j) = \begin{bmatrix} 0.5 & 0.25 \\ 0.05 & 0.1 \end{bmatrix} \begin{bmatrix} B_1(z^{-1}) & \varepsilon_1^{\mathrm{FIR}}(j) \\ B_2(z^{-1}) & \varepsilon_2^{\mathrm{FIR}}(j) \end{bmatrix},$$

where $B_1(z^{-1})$ and $B_2(z^{-1})$ are scalar FIR lowpass discrete filters of order 10 with normalized cutoff frequencies equal to 0.1 and 0.08, respectively, and $\varepsilon_{1,2}^{\mathrm{FIR}}(j)$ are i.i.d. normal random variables with unit variance.

The structure of the two processes have been scaled in order to produce highly auto and crosscorrelated noise sequences with a similar signal-to-noise ratio of 20 dB on each measured output trajectory.

A bunch of independent realizations of the AR and FIR stochastic processes have been generated and added to the model output trajectory vector $y(t_j, \theta^*)$ to form two benchmarks of 5000 sets of measurements each. The input signal, the model output trajectories and one example of the measured output trajectories under both cases of stochastic process are illustrated in Fig. 1.

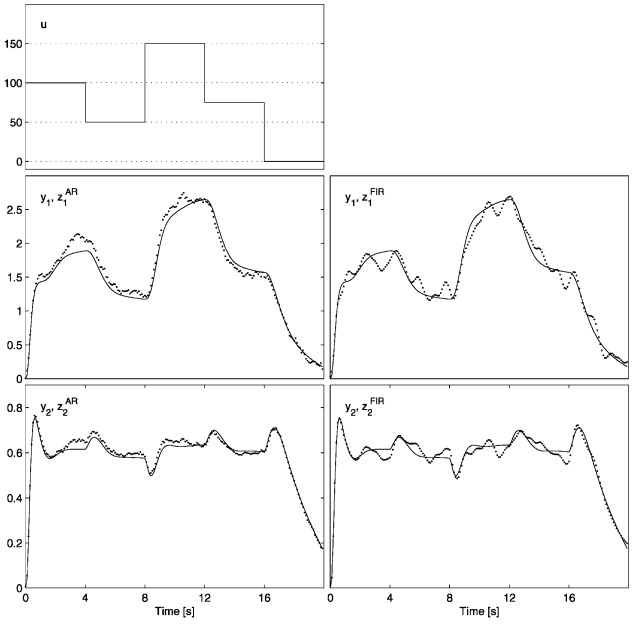

Fig. 1. Input signal u, output model trajectories y_i in plain line and output measurements z_i in dotted line, for the AR and FIR cases.

For each set of measurements, an ML estimate of the parameter vector has been computed with the algorithm. The WLS estimate obtained in the first step has been recorded as well, for comparison purpose. For a more appropriate comparison, the parameters estimated by the AML method are also computed. The order of the AR models used to estimate the ICM in the ML approach is 2 for the first type of noise generator and 8 for the second type. The same values are used for the order of the AR filter in the AML method.

The nonlinear minimization problems required to find the parameter estimates have been solved using the Nelder–Mead simplex method (Nelder & Mead, 1965). This choice was motivated by its relative robustness for optimization in presence of noise and its efficiency for small dimension problems, as stated in Parkinson and Hutchinson (1972, Chap. 8). Some basic tests have been done on a subset of 100 experiments out of the 5000 to verify that local minima were not reached. These tests consisted in repeating the optimization procedure with different starting values and checking that the same optimum is achieved.

Since the true parameter vector and the true noise structure are known, the CR lower bound on the covariance matrix of the parameter estimates can be computed exactly using (10). This bound provides the *minimum possible* standard deviation that can be achieved and will serve as a comparison basis.

The huge number of simulations will allow us to represent accurately several statistical properties of the estimators such as their empirical distribution and their standard deviation. To justify this large number of simulations, one has just to realize the way the empirical standard deviation of a parameter estimate evolves with the number of experiments. Indeed, as illustrated in Fig. 2, the convergence of this quantity is quite slow. On this typical example, it takes about 2500 experiments for the value to stabilize around five percent of the final value. A number of 5000 experiments should hopefully ensure a sufficient accuracy on the statistical properties we are now going to compare.

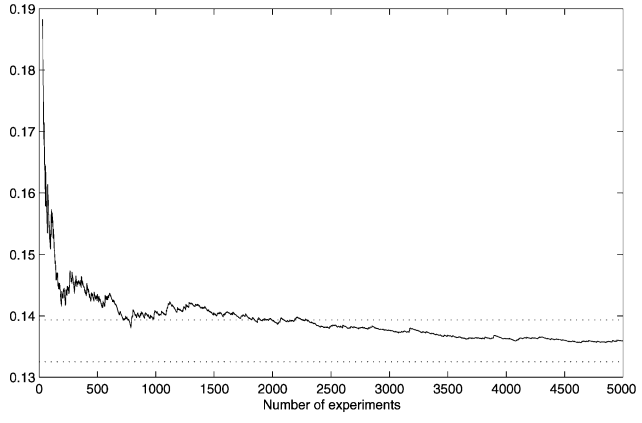

Fig. 2. Typical evolution of the empirical standard deviation of a parameter estimate with respect to the number of experiments. The dotted lines delimit the 5% interval around the final value.

Table 1
Standard deviations of $\hat{\theta}$ and CR bounds

	σ^{WLS}	σ^{AML}	σ^{ML}	σ^{CR}
$\hat{\theta}_1^{\text{AR}}$	0.535	0.132	0.136	0.108
$\hat{\theta}_2^{\text{AR}}$	0.400	0.100	0.103	0.085
$\hat{\theta}_3^{\text{AR}}$	0.286	0.089	0.093	0.069
$\hat{\theta}_1^{\text{FIR}}$	0.566	0.224	0.136	0.087
$\hat{\theta}_2^{\text{FIR}}$	0.426	0.166	0.101	0.064
$\hat{\theta}_3^{\text{FIR}}$	0.288	0.126	0.093	0.069

The standard deviation of the 5000 independent WLS, AML and ML estimates of θ obtained under each case of stochastic process are given in Table 1, along with the CR bounds. The empirical distributions of the estimated parameters are illustrated in Fig. 3, superimposed on normal distributions corresponding to the CR lower bound.

In the AR case, i.e. the stochastic part of the system is in the model set, the ML and AML estimates are not significantly different. The improvement of those estimators with respect to the single step WLS estimator is impressive. This clearly proves the necessity of taking into account the correlation in the noise.

In the FIR case a very high improvement with respect to the WLS estimator is still observed for the ML and AML ones. That reveals the ability of an AR filter to be used as a noise model even if the actual noise has a different structure. It is also observed that the improvement achieved by the ML estimator is significantly better than the one obtained with the AML estimator. This tends to illustrate that the ML method appears to be more robust against noise model misspecification. Hence, this suggests that the ML method is relevant in most practical applications where the noise structure is unknown, at least if the computational extra charge does not matter.

Let us now investigate the iterative version of the algorithm. Three additional iterations of step 2 of the algorithm have been computed, identifying each time new AR models on the residuals of the previous iteration. The standard deviation of the ML estimates obtained at each iteration is represented in Fig. 4. For comparison purpose, the CR bounds and the standard deviation of the AML estimates, taken from Table 1, are illustrated as well. In all cases, the biggest improvement is achieved at the first iteration (step 2). Yet a second iteration still brings a slight improvement in both AR and FIR cases while the subsequent iterations do not significantly reduce the standard deviation. In David and Bastin (2001) and David and Bastin (1999) it appeared that, in the scalar case, further iterations were not necessary. We see, from the considered example, that in the more general

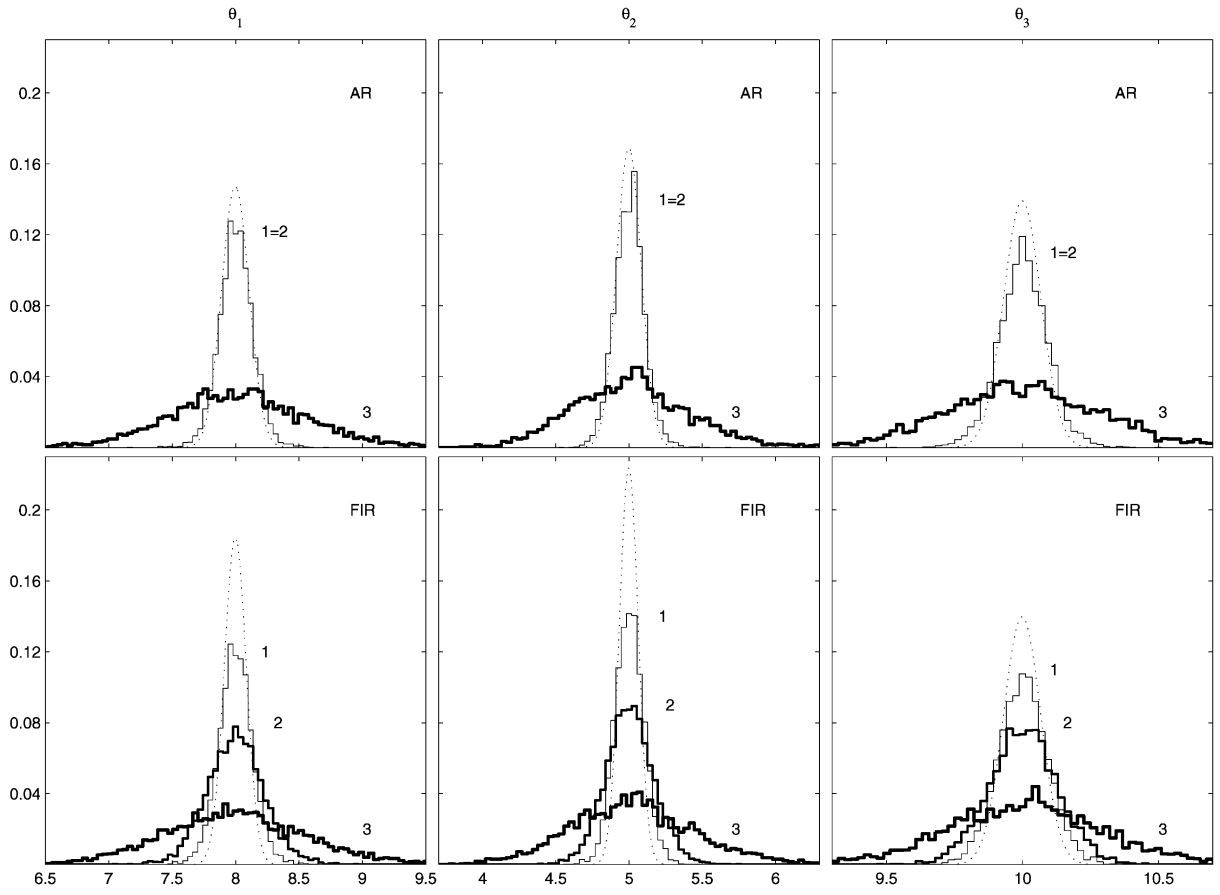

Fig. 3. The histograms are represented with an increasing degree of line thickness: 1 is for ML, 2 for AML and 3 for WLS. The dotted lines represent the normal distributions corresponding to CR bounds.

framework addressed in this paper, one additional iteration seems worthwhile.

Fig. 4 illustrates also the convergence behavior of the algorithm. It shows up that the standard deviation of the ML estimate converges very closely to the CR bound in the AR case while it seems to converge to a greater bound in the FIR case. This is normal since the ICM of the FIR process will obviously never be exactly identified from an AR model. If we compare the convergence limit of the ML estimator with the AML estimator, we can conclude that even in the AR case now, the iterated ML estimator outperforms the AML one. If we forget the second term in its cost function, the iterated ML approach would coincide with the AML one. The slight difference that we observe here comes therefore from this second term. Actually, without noise model misspecification, the AML method would yield the true ML estimate asymptotically. For a finite N though, the AML method should not be considered as the ideal one anymore since it is unable to account for unknown initial conditions of the AR filter. On the other hand, the ML approach does implicitly take into account these unknown initial conditions via the correction term. For that reason, it achieves the best performances on the condition that it is sufficiently iterated. Obviously, the difference between the ML and AML approaches should vanish as N increases, that is as the effect of the correction term becomes negligible.

7. Confidence region

The ICM estimate can still be advantageously applied to the construction of a confidence region around the estimated parameter vector. It is evident that the same experiment is rarely repeated 5000 times in order to have an idea of the parameter dispersion. Often, only one experiment is carried out and a confidence region around the identified parameter has to be evaluated. The covariance matrix of the parameter vector, $\Sigma_{\hat{\theta}}$, is commonly estimated using its CR bound (10) computed for $\hat{\theta}$ and $\hat{\mathbf{\Sigma}}^{-1}$. The large matrix $\hat{\mathbf{\Sigma}}^{-1}$ being estimated by inverting the sample covariance in the WLS case and with the full ICM estimate in the ML case. Assuming a normal distribution for $\hat{\theta}$, an approximate $100(1-\alpha)\%$ confidence region for the parameter vector $\hat{\theta}$ is given by (see e.g. Seber & Wild, 1989)

$$\{\theta\colon (\theta-\hat{\theta})\hat{\Sigma}_{\hat{\theta}}^{-1}(\theta-\hat{\theta})^{\mathrm{T}} \leqslant p\mathscr{F}_{p,N-p}^{\alpha}\}, \tag{19}$$

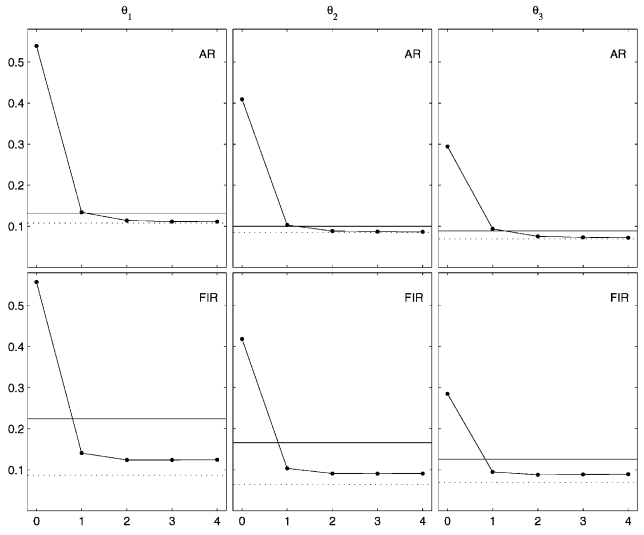

Fig. 4. Standard deviation of $\hat{\theta}^{\mathrm{ML}}$ with respect to the number of iterations. The initial point (iteration 0) corresponds to the preliminary WLS estimate (step 1). The horizontal dotted line indicates the CR bounds while the standard deviation corresponding to the AML estimates is represented by the horizontal plain line.

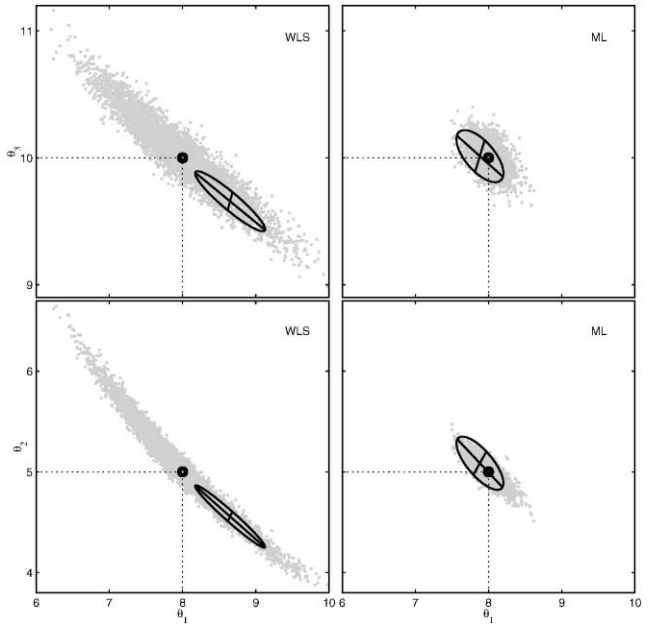

Fig. 5. Projections of the 5000 WLS and ML parameter estimates. The projections of the confidence ellipsoids are drawn for one estimate chosen arbitrarily. The bold circles point the true value of θ.

where $\mathcal{F}^{\alpha}_{i,j}$ is the upper α critical value of the $\mathcal{F}_{i,j}$ distribution. The set described by (19) is the inner space of an ellipsoid centered on $\hat{\theta}$.

Fig. 5 represents two projections of the 3D plot of the 5000 WLS and ML parameter estimates taken from the FIR set. Superimposed are the projections of the 99% confidence ellipsoids corresponding to one estimate chosen arbitrarily. It clearly appears that the confidence region estimated in the ML case, using the full ICM estimate, is far better than the one estimated in the WLS case. It includes the true parameter value and its shape better fits the actual parameters dispersion.

The computation of $\hat{\Sigma}_{\hat{\theta}}$ has been done systematically for each estimated parameter vector. If $\hat{\theta}$ was really normally distributed and if the estimate of each confidence region was exact, 99% of those regions would include the true parameter. Since those conditions are far to be met, it turns out that among the confidence regions computed with the WLS method, only 6% include the true parameter in the AR case and 21% in the FIR case. These percentages become 93% and 83%, respectively, with the ML method.

8. Conclusion

The ML parameter estimation method developed in David and Bastin (2001) has been generalized in this paper to the case of auto and crosscorrelated output noise. The efficiency of the estimation method has been emphasized, in comparison with classical ones, from Monte-Carlo simulations. A significant gain is achieved in the statistical properties of the estimator, namely a reduction of its standard deviation. In addition, the proposed ICM estimate provides an efficient way to compute accurately the confidence region around the estimated parameter.

The quite general form of the nonlinear deterministic part of model (1) and the ability of the proposed ICM estimate to be successfully applied also when the noise structure is not in the model set suggest a wide range of application of the proposed method.

Acknowledgements

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

Akaike, H. (1981). Modern development of statistical methods. In P. Eykhoff (Ed.), *Trends and progress in system identification*. New York: Pergamon Press.

David, B., & Bastin, G. (1999). A maximum likelihood parameter estimation method for nonlinear dynamical systems. In *38th IEEE conference on decision and control*, Phoenix, Arizona, USA. (pp. 612–617).

David, B., & Bastin, G. (2001). An estimator of the inverse covariance matrix and its application to ML parameter estimation in dynamical systems. *Automatica*, *37*, 99–106.

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

Gohberg, I. C., & Heinig, G. (1974). Inversion of finite Toeplitz matrices with entries being elements from a noncommutative algebra. *Revue Roumaine de Mathematiques Pures et appliques* XIX(5), 623–663 (in Russian).

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

Iohvidov, I. S. (1982). *Hankel and toeplitz matrices and forms*. Boston, USA: Birhhäuser (translated by G. Philip and A. Thijsse).

Ljung, L. (1999). *System identification: Theory for the user*. (2nd ed.). Englewood Cliffs, NJ, USA: 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). London: Academic Press.

Rissanen, J. (1985). Minimum description length principles. In S. Kotz & N.L. Johnson (Eds.), *Encyclopedia of statistical sciences* (Vol. 5). New York: Wiley.

Seber, G. A. F., & Wild, C. J. (1989). *Nonlinear regression*. New York, USA: 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 Systems 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.