Proceedings of the 17th World Congress
The International Federation of Automatic Control
Seoul, Korea, July 6-11, 2008

Finite Sample Confidence Regions for Parameters in Prediction Error Identification using Output Error Models

Arnold J. den Dekker * **Xavier Bombois** *
Paul M.J. Van den Hof *

* *Delft Center for Systems and Control, Delft University of Technology, Delft, The Netherlands*
(Tel: +31 15 2781823; e-mail: a.j.dendekker@tudelft.nl).

Abstract: The purpose of this paper is to evaluate the reliability in finite samples of different methods for constructing probabilistic parameter confidence regions in prediction error identification using Output Error (OE) models. The paper presents alternatives to the "classical method" of constructing asymptotically valid confidence regions, which is based on the asymptotic statistical properties of the parameter estimator. It is shown that if alternative test statistics are used, more reliable confidence regions for finite samples can be obtained. Particularly, it is demonstrated that the use of a test statistic based on the Fisher score allows the construction of exact confidence regions for finite samples.

Keywords: Prediction error methods; uncertainty; system identification; statistical inference; parameter identification; maximum likelihood estimators; parameter estimation.

1. INTRODUCTION

Prediction error (PE) methods have become a wide-spread technique for system identification. In PE identification, the reliability of the parametric dynamic models identified on the basis of measurement data is generally limited due to noise disturbances and the finite length of the data. The need for quantifying model uncertainties is especially relevant when identified models are used as a basis for model based control, monitoring, simulation or any other model based decision-making. An indication of the reliability of the identified model is given by probabilistic confidence regions for its parameters. A confidence region is a region in the parameter space that attempts to "cover" the true but unknown parameter vector with a nominal probability. Confidence regions are exact if the coverage probability equals the nominal probability. The confidence regions that are most widely used in PE identification are derived from the asymptotic statistical properties of the parameter estimator. These regions generally have a simple ellipsoidal shape, but for finite data lengths their nominal level is often misleading. Finite-time analysis of parameter inference in PE identification is an important problem, however with few results so far. For some results see e.g. (Campi and Weyer, 2002; Weyer and Campi, 2002; Campi and Weyer, 2005). In this paper, we consider the construction of confidence regions for the parameters of Output Error (OE) models. The classical method as well as a recently proposed alternative (Douma and Van den Hof, 2005) are briefly reviewed. In addition, three new alternative methods (with finite time perspectives) are proposed, two of which are based on likelihood theory. The goal of the paper is to evaluate, validate and compare the reliability of different methods for constructing confidence regions for finite data lengths. This is done by means of Monte-Carlo simulation experiments. It is shown that although all methods considered are equivalent in very large samples , for finite data lengths the methods show different reliability. Results of a similar study for the case of ARX (Auto Regression with eXogenous inputs) models were presented earlier (den Dekker et al., 2007).

2. STATISTICAL INFERENCE IN PREDICTION ERROR IDENTIFICATION

We will consider dynamical data generating systems of the form

$$y(t) = G_0(q)u(t) + H_0(q)e(t) \tag{1}$$

with q the standard shift operator, $y(t)$ the stochastic (measurable) output signal, $u(t)$ the deterministic (measurable) input signal and $e(t)$ (non-measurable) zero-mean Gaussian white noise. In (1), $G_0(z)$ and $H_0(z)$ are proper rational transfer functions that have no poles in $|z| \geq 1$, which means that the system is stable. In addition, $H_0(z)$ will be restricted to be monic and minimum-phase. The one-step ahead predictor of $y(t)$, given the system (1) and given the observations $\{(y(s), u(s)), s \leq t-1\}$, is given by

$$\hat{y}(t|t-1) = H_0^{-1}(q)G_0(q)u(t) + [1 - H_0^{-1}(q)]y(t), \tag{2}$$

which can be rewritten as

$$y(t) = \hat{y}(t|t-1) + e(t). \tag{3}$$

The one-step ahead predictor (2) is the best one-step ahead predictor in the sense of the conditional expectation (Ljung, 1999). In reality, the true system $(G_0(z), H_0(z))$ is generally unknown, and predictor models determined by a collection of two rational transfer functions $(G(z), H(z))$ are considered instead. A predictor model set $\mathcal{M}$ is defined as any collection of predictor models:

$$\mathcal{M} := \{(G(q,\theta), H(q,\theta))|\theta \in \Theta \subset \mathbb{R}^n\} \tag{4}$$

with θ a real valued parameter vector ranging over a subset of $\mathbb{R}^n$. It is assumed that this model set is composed of predictor models (i.e., transfer functions) that satisfy the same conditions of properness, stability and monicity as the transfer functions $H_0(z)$ and $G_0(z)$ described above. Underlying the set of models, there is a parameterization that determines the specific relation between a parameter $\theta \in \Theta$ and a model M within $\mathcal{M}$. If we assume that the data generating system $\mathcal{S}$ belongs to the model set ($\mathcal{S} \in \mathcal{M}$), there exists an exact parameter θ_0 reflecting the transfer functions G_0 and H_0 and one may rewrite (3) as

$$y(t) = \hat{y}(t|t-1;\theta_0) + e(t), \tag{5}$$

with

$$\hat{y}(t|t-1;\theta) = H^{-1}(q,\theta)G(q,\theta)u(t)[1-H^{-1}(q,\theta)]y(t). \tag{6}$$

The model of the observations is given by

$$y(t) = \hat{y}(t|t-1;\theta) + \varepsilon(t,\theta), \tag{7}$$

with $\varepsilon(t,\theta)$ the prediction errors. Since $\mathcal{S} \in \mathcal{M}$, the prediction errors evaluated at θ_0 are equal to $e(t)$ and thus zero mean, independent, Gaussian distributed, with probability density function (PDF)

$$f_\varepsilon(\varepsilon(t,\theta_0);\theta_0) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}\varepsilon^2(t,\theta_0)\right] \tag{8}$$

with σ^2 the variance of $e(t)$. The joint PDF of the observations $y^N = \{y(t)\}_{t=1,\cdots,N}$ (conditioned on the given deterministic input sequence u^N) is given by:

$$f_y(y^N;\theta_0) = \prod_{t=1}^{N} f_\varepsilon(y(t)-\hat{y}(t|t-1;\theta_0)) = \prod_{t=1}^{N} f_\varepsilon(\varepsilon(t,\theta_0);\theta_0). \tag{9}$$

Taking the logarithm yields:

$$\log f_y(y^N;\theta_0) = \sum_{t=1}^{N} \log f_\varepsilon(\varepsilon(t,\theta_0);\theta_0), \tag{10}$$

which can be written as

$$\log f_y(y^N;\theta_0) = -\frac{N}{2}\log(2\pi) - N\log\sigma - \frac{N}{2\sigma^2}V_N(\theta_0) \tag{11}$$

with

$$V_N(\theta_0) = \frac{1}{N}\sum_{t=1}^{N}\varepsilon(t,\theta_0)^2. \tag{12}$$

2.1 The Fisher score

The Fisher score $S(\theta)$ is defined as

$$S(\theta) = \frac{\partial \log f_y(y^N;\theta)}{\partial\theta}. \tag{13}$$

It can be shown that the Fisher score (13) evaluated at the true value θ_0 of θ has mean zero:

$$\mathbb{E}[S(\theta_0)] = 0. \tag{14}$$

It follows from (13) and (11) that the Fisher score can be written as

$$S(\theta) = \frac{-N}{2\sigma^2}V'(\theta) \tag{15}$$

with $V'_N(\theta)$ the first derivative of $V_N(\theta)$ with respect to θ.

2.2 The Fisher information matrix

The covariance matrix of the Fisher score $S(\theta_0)$ is described by

$$J(\theta_0) = \mathbb{E}\left[S(\theta_0)S^T(\theta_0)\right] \tag{16}$$

which is known as the Fisher information matrix (Fisher, 1922). It can be shown that $J(\theta_0)$ may alternatively be written as

$$J(\theta_0) = -\mathbb{E}\left[\left.\frac{\partial^2 \log f_y(y^N;\theta)}{\partial\theta^2}\right|_{\theta=\theta_0}\right]. \tag{17}$$

It follows from (11), (15), (16) and (17) that the Fisher information matrix is given by:

$$J(\theta_0) = \frac{N^2}{4\sigma^4}\mathbb{E}\left[V'(\theta_0)V'^T(\theta_0)\right], \tag{18}$$

or, alternatively, by

$$J(\theta_0) = \frac{N}{2\sigma^2}\mathbb{E}\left[V''(\theta_0)\right], \tag{19}$$

with $V''(\theta)$ the second derivative of $V(\theta)$ with respect to θ. Furthermore, by the multivariate central limit theorem, it is generally derived that, for $N \to \infty$,

$$S(\theta_0) \to \mathcal{N}(0, J(\theta_0)), \tag{20}$$

that is, the Fisher score is asymptotically normally distributed with expectation value zero and covariance matrix $J(\theta_0)$ (Wilks, 1962).

2.3 The likelihood function and the maximum likelihood estimator

By substituting the available observations y^N for the corresponding indeterminate variables in (9) and regarding the resulting expression as a function of the parameter vector θ for fixed observations y^N, the likelihood function, written as $f_y(\theta;y^N)$, is obtained. The maximum likelihood estimator (MLE) of θ_0 is given by

$$\hat{\theta}_N = \arg\max_\theta f_y(\theta;y^N) = \arg\min_\theta V_N(\theta). \tag{21}$$

Fisher (1922) has shown that, for $N \to \infty$,

$$\hat{\theta}_N \to \mathcal{N}(\theta_0, J^{-1}(\theta_0)). \tag{22}$$

Furthermore, Wald (1949) has shown that, under very general conditions, the MLE $\hat{\theta}_N$ is a consistent estimator. Finally, it can be shown that the MLE of σ^2 is given by

$$\hat{\sigma}^2_{ML} = V_N(\hat{\theta}_N), \tag{23}$$

whereas a (slightly) more accurate estimator of σ^2 is given by

$$\hat{\sigma^2} = \frac{N}{N-n}V_N(\hat{\theta}_N). \tag{24}$$

3. OUTPUT ERROR MODELLING

The OE model structure describes the input-output relationship of a linear dynamical system as in (1) with

$$G(q,\theta) = \frac{q^{-n_k}B(q^{-1},\theta)}{F(q^{-1},\theta)}, \quad H(q,\theta) = 1 \tag{25}$$

with n_k the delay and

$$B(q^{-1},\theta) = \left(b_0 + b_1q^{-1} + \cdots + b_{n_b-1}q^{-n_b+1}\right), \tag{26}$$

$$F(q^{-1},\theta) = 1 + f_1q^{-1} + \cdots + f_{n_f}q^{-n_f}, \tag{27}$$

with $\theta^T = [b_0, b_1, \cdots b_{n_b-1}, f_1, \cdots, f_{n_f}]$. In an output error model structure we consider the one-step ahead predictor

$$\hat{y}(t|t-1;\theta) = \frac{B(q,\theta)}{F(q,\theta)}u(t) \tag{28}$$

and we denote the predictor derivative:

$$\psi(t,\theta) = \frac{\partial}{\partial\theta}\hat{y}(t|t-1;\theta). \tag{29}$$

Furthermore, define

$$\Psi(\theta) = \begin{pmatrix} \psi^T(1,\theta) \\ \vdots \\ \psi^T(N,\theta) \end{pmatrix}. \tag{30}$$

Suppose that the data generating system belongs to the model class ($\mathcal{S} \in \mathcal{M}$). The MLE $\hat{\theta}_N$ of θ_0 is then given by (21), where

$$V_N(\theta) = \frac{1}{N}\sum_{t=1}^{N}(y(t) - \hat{y}(t|t-1;\theta))^2 \tag{31}$$

and the Fisher score is equal to

$$S(\theta) = \frac{-N}{2\sigma^2}V_N'(\theta) = \frac{1}{\sigma^2}\sum_{t=1}^{N}\psi(t,\theta)\left[y(t) - \hat{y}(t|t-1;\theta)\right]$$
$$= \frac{1}{\sigma^2}\Psi^T(\theta)\left(\mathbf{y} - \hat{\mathbf{y}}(\theta)\right), \tag{32}$$

with

$$\mathbf{y} = [y(1), y(2), \ldots, y(N)]^T, \tag{33}$$

and

$$\hat{\mathbf{y}}(\theta) = [\hat{y}(1|0;\theta), \hat{y}(2|1;\theta), \ldots, \hat{y}(N|N-1;\theta)]^T. \tag{34}$$

Evaluating (32) at $\theta = \theta_0$ yields:

$$S(\theta_0) = \frac{1}{\sigma^2}\Psi^T(\theta_0)\boldsymbol{e}, \tag{35}$$

with $\boldsymbol{e} = [e(1)\cdots e(N)]^T$. It is very important to note that for OE models, the Fisher score (35) is exactly normally distributed (due to the fact that the term $\Psi(\theta_0)$ is deterministic) and the asymptotic result (20) holds for finite data as well:

$$S(\theta_0) \sim \mathcal{N}(0, J(\theta_0)), \forall N, \tag{36}$$

where the Fisher information matrix $J(\theta_0)$ is given by:

$$J(\theta_0) = \mathbb{E}\left[S(\theta_0)S^T(\theta_0)\right] = \frac{1}{\sigma^2}\Psi^T(\theta_0)\Psi(\theta_0). \tag{37}$$

As we will see later, this result allows us to construct confidence regions that are exact also for finite values of N, which is remarkable since in the classical approach, based on the statistical properties of the parameter estimator, exact confidence regions for finite samples are only obtained for linear regression models with deterministic regressors, such as Finite Impulse Response (Ljung, 1999).

4. CONFIDENCE REGIONS

Confidence regions can be interpreted as the result of hypothesis testing. If we want to test the null hypothesis

$$H_0 : \theta_0 = \theta \tag{38}$$

against the alternative hypothesis

$$H_1 : \theta_0 \neq \theta, \tag{39}$$

at significance level α, where the significance level is defined as the probability of rejecting H_0 when H_0 is true, we first have to construct a test statistic with known distribution under H_0. Such a test statistic will be a function of θ and y^N. The general test principle now states that the null hypothesis H_0 is rejected if the sample value of the test statistic used is larger than some user specified threshold. Knowledge of the PDF of the test statistic under H_0 allows one to compose tests (i.e., set thresholds) with a desired significance level. This principle can now be used to compose confidence regions for the parameters θ_0. This is done as follows. First, select a test statistic for testing the null hypothesis $\theta_0 = \theta$ against the alternative $\theta_0 \neq \theta$, at significance level α. A $100(1-\alpha)\%$ confidence region for θ_0 is then constituted by the set of all values θ for which the null hypothesis $\theta_0 = \theta$ would be accepted.

4.1 The classical approach

In the classical approach to constructing confidence regions for the parameters of OE models, the starting point is a first order Taylor expansion:

$$(\hat{\theta}_N - \theta_0) \approx -[V''(\theta_0)]^{-1}[V_N'(\theta_0)], \tag{40}$$

with $V''(\theta_0) = \frac{2}{N}\Psi^T(\theta_0)\Psi(\theta_0)$ (assuming a deterministic input sequence u^N) (Ljung, 1999). It can be shown that

$$V_N'(\theta_0) \sim \mathcal{N}(0, \frac{2\sigma^2}{N}V''(\theta_0)). \tag{41}$$

Hence, in the first order Taylor approximation, an expression for the covariance matrix of $(\hat{\theta}_N - \theta_0)$ is given by

$$P = [V''(\theta_0)]^{-1}\frac{2\sigma^2}{N}V''(\theta_0)[V''(\theta_0)]^{-1} = \frac{2\sigma^2}{N}[V''(\theta_0)]^{-1}. \tag{42}$$

This covariance matrix is approximated using an estimate of $V''(\theta_0)$ given by the term $\frac{2}{N}\Psi^T(\hat{\theta}_N)\Psi(\hat{\theta}_N)$ to arrive at the test statistic

$$\frac{1}{\sigma^2}(\hat{\theta}_N - \theta)^T\Psi^T(\hat{\theta}_N)\Psi(\hat{\theta}_N)(\hat{\theta}_N - \theta). \tag{43}$$

Under H_0, the test statistic (43) has approximately a χ^2_n distribution, i.e., a chi-square distribution with n degrees of freedom. An approximately valid $100(1-\alpha)\%$ confidence region for θ_0 is then given by

$$\left\{\theta | \frac{1}{\sigma^2}(\hat{\theta}_N - \theta)^T\Psi^T(\hat{\theta}_N)\Psi(\hat{\theta}_N)(\hat{\theta}_N - \theta) \leq \chi^2_{n,1-\alpha}\right\}, \tag{44}$$

where $\chi^2_{n,1-\alpha}$ is the $1-\alpha$ quantile of the chi-square distribution with n degrees of freedom (cfr. Mood et al. (1974)). This confidence region corresponds with the one implemented in the Matlab System Identification Toolbox (Ljung, 2003) and has also been derived by Douma and Van den Hof (2006) using an alternative paradigm for probabilistic uncertainty bounding.

Alternatively, (44) can be derived starting from the (asymptotic) statistical properties of the MLE described in section 2. It follows from (22), and the consistency property of the MLE that the quadratic form

$$(\hat{\theta}_N - \theta_0)^T J(\hat{\theta}_N)(\hat{\theta}_N - \theta_0). \tag{45}$$

has asymptotically (i.e., for $N \to \infty$) a χ^2_n distribution. (Kay, 1998). Then, if we want to test the null hypothesis (38) against the alternative hypothesis (39) the test statistic

$$T_W = (\hat{\theta}_N - \theta)^T J(\hat{\theta}_N)(\hat{\theta}_N - \theta), \tag{46}$$

which is known as the Wald test statistic, may be used. Asymptotically, (46) has a χ^2_n distribution under H_0. It follows from section 3 that for the case of normally distributed data y^N and an OE model structure, the test statistic T_W can be written as:

$$T_W = \frac{1}{\sigma^2}(\hat{\theta}_N - \theta)^T \Psi^T(\hat{\theta}_N)\Psi(\hat{\theta}_N)(\hat{\theta}_N - \theta), \quad (47)$$

which equals (43). Finally, it can be shown that if σ^2 is replaced by $\hat{\sigma^2}$, the distribution of (43) is better approximated by an F distribution and the term $\chi^2_{n,1-\alpha}$ in (44) is to be replaced by $nF_{n,N-n,1-\alpha}$, with $F_{n,N-n,1-\alpha}$ the $1-\alpha$ quantile of the F distribution with n and N degrees of freedom. This leads to the following asymptotically valid $100(1-\alpha)\%$ confidence region for θ_0:

$$\left\{\theta | \frac{1}{n}\frac{(\hat{\theta}_N - \theta)^T \Psi^T(\hat{\theta}_N)\Psi(\hat{\theta}_N)(\hat{\theta}_N - \theta)}{\hat{\sigma^2}} \leq F_{n,N-n,1-\alpha}\right\} \quad (48)$$

4.2 An alternative approach without Taylor approximation

An alternative approach, without Taylor approximation, was recently proposed by Douma and Van den Hof (2005). Since the MLE $\hat{\theta}_N$ of θ_0 is given by Eq.(21), it must satisfy $V'_N(\hat{\theta}_N) = 0$. By defining

$$y_F(t) = F(q,\hat{\theta}_N)^{-1}y(t); \quad u_F(t) = F(q,\hat{\theta}_N)^{-1}u(t) \quad (49)$$

the equation $V'_N(\hat{\theta}_N) = 0$ can be rewritten as

$$\frac{1}{N}\sum_{t=1}^{N}\left[F(q,\hat{\theta}_N)y_F(t) - z^{-n_k}B(q,\hat{\theta}_N)u_F(t)\right]\cdot\psi(t,\hat{\theta}_N) = 0. \quad (50)$$

The parameter estimate $\hat{\theta}_N$ satisfying these equations can now be written in a linear regression-type equation:

$$\hat{\theta}_N = \left(\Psi^T\Phi\right)^{-1}\Psi^T\mathbf{y}_F \quad (51)$$

with $\Psi = \Psi(\hat{\theta}_N)$, $\mathbf{y}_F = [y_F(1)\cdots y_F(N)]^T$,

$$\Phi^T = \left[\varphi_F(1,\hat{\theta}_N),\cdots,\varphi_F(N,\hat{\theta}_N)\right], \quad (52)$$

and

$$\varphi_F^T(t,\hat{\theta}_N) = [u_F(t-n_k)\cdots u_F(t-n_k-n_b+1) \\ -y_F(t-1)\cdots - y_F(t-n_f)]. \quad (53)$$

The system's relations

$$y(t) = \frac{q^{-n_k}B_0(q)}{F_0(q)}u(t) + e(t) \quad (54)$$

can then be rewritten as:

$$F_0(q)y_F(t) = q^{-n_k}B_0(q)u_F(t) + \frac{F_0(q)}{F(q,\hat{\theta}_N)}e(t), \quad (55)$$

which can be rewritten in the regression form:

$$\mathbf{y}_F = \Phi\theta_0 + \boldsymbol{e}_F, \quad (56)$$

where $\boldsymbol{e}_F = \frac{F_0(q)}{F(q,\hat{\theta}_N)}[e(1)\cdots e(N)]^T$. Substituting (56) in (51) now delivers:

$$\left(\Psi^T\Phi\right)(\hat{\theta}_N - \theta_0) = \Psi^T\boldsymbol{e}_F. \quad (57)$$

Unlike the Taylor approximation (40), (57) is an exact result. Unfortunately, the statistical distribution of the random variable $\Psi^T\boldsymbol{e}_F$ is unknown for finite values of N. It can be shown, however, that asymptotically $\Psi^T\boldsymbol{e}_F$ is normally distributed with zero mean and covariance matrix $Q = \sigma^2\Psi(\theta_0)\Psi^T(\theta_0)$. Therefore, the test statistic $(\hat{\theta}_N - \theta)^T P_D^{-1}(\hat{\theta}_N - \theta)$ with

$$P_D = \left(\Psi^T\Phi\right)^{-1}\cdot Q\cdot\left(\Phi^T\Psi\right)^{-1} \quad (58)$$

is asymptotically χ^2_n distributed under H_0. By replacing the term $\Psi(\theta_0)\Psi(\theta_0)^T$ by the estimate $\Psi\Psi^T$, and σ^2 by $\hat{\sigma^2}$, Douma and Van den Hof (2005) arrive at the following asymptotically valid $100(1-\alpha)\%$ confidence region for θ_0:

$$\left\{\theta | (\hat{\theta}_N - \theta)^T P_s^{-1}(\hat{\theta}_N - \theta) \leq nF_{n,N-n,1-\alpha}\right\}, \quad (59)$$

with

$$P_s = \left(\Psi^T\Phi\right)^{-1}\hat{\sigma^2}\Psi^T\Psi\left(\Phi^T\Psi\right)^{-1}. \quad (60)$$

4.3 A new approach without Taylor approximation

A new alternative approach starts again from the equation $V'(\hat{\theta}_N) = 0$, or equivalently,

$$\frac{1}{N}\sum_{t=1}^{N}\psi(t,\hat{\theta}_N)\varepsilon(t,\hat{\theta}_N) = 0. \quad (61)$$

It can be shown that, in the case of OE systems,

$$\varepsilon(t,\hat{\theta}_N) = e(t) - \varphi_{oe}^T(t,\theta_0)(\hat{\theta}_N - \theta_0), \quad (62)$$

with

$$\varphi_{oe}^T(t,\theta_0) = [u_F(t-n_k)\cdots u_F(t-n_k-n_b+1) \\ -G(q,\theta_0)u_F(t-1)\cdots - G(q,\theta_0)u_F(t-n_f)] \quad (63)$$

being a vector with dimension $n = n_b + n_f$. Substituting (62) for $\varepsilon(t,\hat{\theta}_N)$ in (61) and using a matrix notation it follows

$$\Psi^T\Phi_{oe}(\theta_0)(\hat{\theta}_N - \theta_0) = \Psi^T\boldsymbol{e}, \quad (64)$$

with $\Psi = \Psi(\hat{\theta}_N)$ and $\Phi_{oe}^T(\theta_0) = [\varphi_{oe}(1,\theta_0),\cdots,\varphi_{oe}(N,\theta_0)]$. Pursuing a similar line of reasoning as in subsection 4.2, we can conclude that the test statistic $(\hat{\theta}_N - \theta)^T P_B^{-1}(\theta)(\hat{\theta}_N - \theta)$, with

$$P_B(\theta) = \left(\Psi^T\Phi_{oe}(\theta)\right)^{-1}\sigma^2\Psi^T\Psi\left(\Phi_{oe}^T(\theta)\Psi\right)^{-1}, \quad (65)$$

is asymptotically χ^2_n distributed under H_0. An asymptotically valid $100(1-\alpha)\%$ confidence region for θ_0 can then be formulated as

$$\left\{\theta | (\hat{\theta}_N - \theta)^T P_{oe}^{-1}(\theta)(\hat{\theta}_N - \theta) \leq nF_{n,N-n,1-\alpha}\right\}, \quad (66)$$

with

$$P_{oe}(\theta) = \left(\Psi^T\Phi_{oe}(\theta)\right)^{-1}\hat{\sigma^2}\Psi^T\Psi\left(\Phi_{oe}^T(\theta)\Psi\right)^{-1}. \quad (67)$$

Note that unlike P_s in (59) and $\Psi^T\Psi$ in (48), the term P_{oe} in (66) depends on θ. Therefore, the construction of (66) is generally computationally expensive, requiring the evaluation of $P_{oe}(\theta)$ at a sufficient number of points to produce contours. In addition, whereas (48) and (59) are ellipsoids, confidence region (66) generally is not.

4.4 An approach based on the likelihood ratio

A second alternative approach is to use a test statistic that is based on a comparison of (maximized) likelihood functions under the hypotheses H_0 and H_1. Since the models underlying these hypotheses are nested, the generalized likelihood ratio (LR)

$$L_G = \frac{f_y(\theta;y^N)}{\sup_\theta f_y(\theta;y^N)} = \frac{f_y(\theta;y^N)}{f_y(\hat{\theta}_N;y^N)} \quad (68)$$

is bound to be between 0 (likelihoods are non-negative) and 1. It has been shown that under certain regularity conditions, the test statistic

$$T_{LR} = -2\log L_G \quad (69)$$

has asymptotically a χ^2_n distribution under H_0 (Kay, 1998). It follows from section 3 that for the case of normally distributed data y^N and an OE model structure, the test statistics T_{LR} can be written as:

$$T_{LR} = \frac{N}{\sigma^2}\left(V_N(\theta) - V_N(\hat{\theta}_N)\right), \tag{70}$$

Based on this test statistic, the following (asymptotically valid) $100(1-\alpha)\%$ confidence region for θ_0 can be derived:

$$\left\{\theta | N\left(V_N(\theta) - V_N(\hat{\theta}_N)\right) \leq \sigma^2 \chi^2_{n,1-\alpha}\right\}, \tag{71}$$

If σ^2 is unknown and replaced by $\hat{\sigma^2}$, $\chi^2_{n,1-\alpha}$ is replaced by $nF(n, N-n, 1-\alpha)$ to arrive at

$$\left\{\theta | \frac{N}{n}\frac{\left(V_N(\theta) - V_N(\hat{\theta}_N)\right)}{\hat{\sigma^2}} \leq F_{n,N-n,1-\alpha}\right\} \tag{72}$$

Note that the construction of (71) and (72) is generally computationally expensive, requiring the evaluation of $V(\theta)$ at a sufficient number of points to produce contours. Likelihood ratio based confidence regions have also been discussed by Quinn et al. (2005).

4.5 An exact Fisher score based finite sample approach

A third new approach is based on the statistical properties of the Fisher score described in section 3. It follows from (36) that for OE models the quadratic form

$$S(\theta_0)^T J^{-1}(\theta_0) S(\theta_0) \tag{73}$$

has a χ^2_n distribution. Then, if we want to test the null hypothesis (38) against the alternative hypothesis (39) the test statistic

$$T_R = S(\theta)^T J^{-1}(\theta) S(\theta), \tag{74}$$

which is known as the Rao (or score) test statistic (Kay, 1998) may be used, since it is known to be exactly χ^2_n distributed under H_0 (for all N). It follows from section 3 that (74) can be written as:

$$T_R = \frac{1}{\sigma^2}(\mathbf{y} - \hat{\mathbf{y}}(\theta))^T P(\theta)(\mathbf{y} - \hat{\mathbf{y}}(\theta)), \tag{75}$$

with

$$P(\theta) = \Psi(\theta)(\Psi^T(\theta)\Psi(\theta))^{-1}\Psi^T(\theta). \tag{76}$$

Note that the $N \times N$ matrix (76) is an orthogonal projection matrix since it is symmetric and idempotent, i.e., $P^2 = P$. Based on this test statistic, the following exact $100(1-\alpha)\%$ confidence regions for θ_0 can be derived:

$$\left\{\theta | (\mathbf{y} - \hat{\mathbf{y}}(\theta))^T P(\theta)(\mathbf{y} - \hat{\mathbf{y}}(\theta)) \leq \sigma^2 \chi^2_{n,1-\alpha}\right\} \tag{77}$$

If σ^2 is unknown, it is known from nonlinear regression theory (Hamilton, 1986; Seber and Wild, 1989) that it is still possible to construct an exact confidence region based on the Rao test statistic. Such a region is obtained by using the following variant of the test statistic T_R:

$$T'_R = \frac{N-n}{n}\frac{(\mathbf{y} - \hat{\mathbf{y}}(\theta))^T P(\theta)(\mathbf{y} - \hat{\mathbf{y}}(\theta))}{(\mathbf{y} - \hat{\mathbf{y}}(\theta))^T [I - P(\theta)](\mathbf{y} - \hat{\mathbf{y}}(\theta))}, \tag{78}$$

with I the $N \times N$ identity matrix. It can be shown that the test statistic T'_R is the ratio of two independent χ^2 distributed random variables with n and $N-n$ degrees of freedom, respectively. Therefore, it is exactly $F_{n,N-n}$ distributed under H_0. This leads to the following exact $100(1-\alpha)\%$ confidence region:

$$\left\{\theta | \frac{N-n}{n}\frac{(\mathbf{y} - \hat{\mathbf{y}}(\theta))^T P(\theta)(\mathbf{y} - \hat{\mathbf{y}}(\theta))}{(\mathbf{y} - \hat{\mathbf{y}}(\theta))^T [I - P(\theta)](\mathbf{y} - \hat{\mathbf{y}}(\theta))} \leq F_{n,N-n,1-\alpha}\right\} \tag{79}$$

Note that the construction of (77) and (79) is generally computationally expensive, requiring the evaluation of $P(\theta)$ at a sufficient number of points to produce contours. The method of obtaining the exact confidence region (79) is often referred to as the lack-of-fit method (Donaldson and Schnabel, 1987; Gallant, 1987).

5. SIMULATION EXPERIMENT

In a MATLAB environment, a Monte Carlo simulation experiment was performed to evaluate and compare the methods for computing confidence regions described in the preceding sections. For different data lengths N, K data sets $(y^N, u^N) = \{y(t), u(t)\}_{t=1,\cdots,N}$ were generated using a data generating system $\mathcal{S}$ that is completely known and belongs to the OE model class:

$$G_0(q) = \frac{q^{-1}(b_0 + b_1 q^{-1})}{1 + f_1 q^{-1} + f_2 q^{-2}}, \quad H_0(q) = 1, \tag{80}$$

with $b_1 = 0.1047, b_2 = 0.0872, f_1 = -1.5578$ and $f_2 = 0.5769$. For each value of N, we used a fixed input sequence u^N, with u^N a realization of a zero mean, Gaussian distributed white noise process with variance $\sigma_u^2 = 1$ being uncorrelated with the zero mean, Gaussian distributed white noise process $\{e(t)\}$ with variance $\sigma^2 = 1$. For each value of N, K different data sets were obtained by repeating the same experiment K times, where each time only the noise realization e^N was different. From each data set, the model was identified using a model set $\mathcal{M}$ with the same OE structure as the data generating system:

$$G(q, \theta) = \frac{q^{-1}(b_0 + b_1 q^{-1})}{1 + f_1 q^{-1} + f_2 q^{-2}}; \quad H(q, \theta) = 1, \tag{81}$$

with $\theta^T = [b_0\, b_1 f_1 f_2]$ and it was recorded whether or not the confidence regions described by (48), (59), (66), (72), and (79) contained the true value θ_0. Note that determining whether θ_0 lay within the confidence regions did not require the construction of the full confidence regions. The observed coverage γ_α, for a particular nominal confidence level $1-\alpha$, is defined as the percentage of the total number of data sets K for which θ_0 lay within the confidence region. In this study, we used $K = 50000$. Furthermore, a nominal confidence level $\alpha = 0.05$ was chosen. This means that the (asymptotic) theory predicts an observed coverage of 95%. Figure 1 shows the observed coverage rates $\gamma_{0.05}$ as a function of the number of data points N. The 95% confidence intervals for $\gamma_{0.05}$ can be obtained from the binomial distribution. The maximum width of these confidence intervals was approximately 0.01. The results show that for increasing data lengths, all observed coverage rates tend to 0.95, as predicted by asymptotic theory. For finite data lengths, however, the different confidence regions show different reliability. The lack of fit method (based on the Rao test statistic) turns out to yield the most reliable confidence regions in the sense that the coverage probability equals the nominal probability for all data lengths (as predicted by theory). For the other confidence regions considered, coverage and nominal probabilities differ significantly for small N. Particularly the "classical" confidence region (48) and the alternative (59) turn out to be unreliable for small N. The reliability of the LR based confidence region (72) and the alternative (66) turn out to be relatively high, but suboptimal when compared to the lack-of-fit

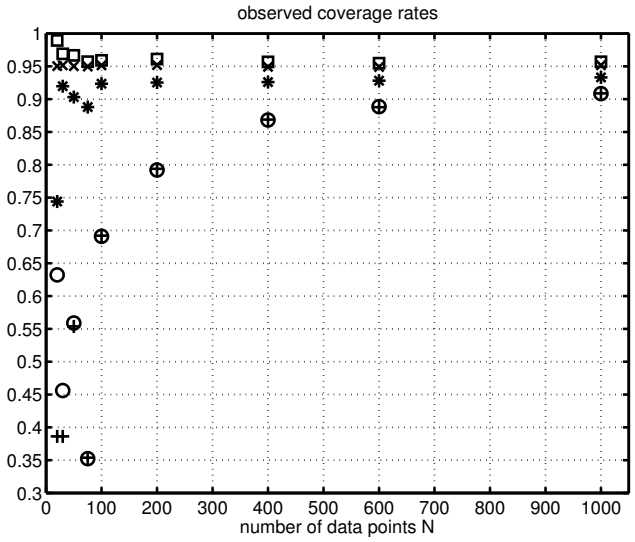

Fig. 1. Results of the simulation experiment described in section 5 . Observed coverage rates of the confidence regions based on the "classical" approach (48)(∘), the alternative approaches (59)(+) and (66)(*), the LR based approach (72)(□), and the lack of fit method (79)(×), as a function of N. The data generating system is given by (80) and the OE model set is described by (81). The nominal confidence level is 0.05. All results were obtained from 50000 realizations.

method. The simulation experiment was repeated for data sequences obtained using different realizations of both the noise contribution e^N and the input sequence u^N in each of the $K = 50000$ data sets (for each value of N). The results were similar to those obtained with a fixed input sequence u^N. More simulation experiments were performed, using alternative data generating systems (all belonging to the OE model class), parameters and nominal confidence rates. All experiments yielded similar results.

6. CONCLUSION

It was shown that the classical method for constructing confidence regions for OE model parameters yields unreliable inference results for small data lengths. In addition, the possibility of constructing exact confidence regions for finite data lengths was demonstrated. It should be noted that whereas the classical regions (ellipsoids) are easy to compute, the construction of the newly proposed (exact) confidence regions is computationally expensive. Therefore, more research should be done towards their practical implementation. The prospect of using randomized algorithms (Tempo et al., 2004) for this purpose is subject of ongoing research, as well as a further generalization of the results to the Box-Jenkins model structure.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge fruitful discussions with Sippe Douma and Asbjörn Klomp.

REFERENCES

M.C. Campi and E. Weyer. Finite sample properties of system identification methods. *IEEE Trans. Autom. Control*, 47(8):1329–1334, 2002.

M.C. Campi and E. Weyer. Guaranteed non-asymptotic confidence regions in system identification. *Automatica*, 41(10):1751–1764, 2005.

A. J. den Dekker, X. Bombois, and P.M.J. Van den Hof. Likelihood based uncertainty bounding in prediction error identification using ARX models: A simulation study. In *Proc. European Control Conference ECC'07*, pages 2879–2886, Kos, Greece, July 2-5, 2007.

J.R. Donaldson and R.B. Schnabel. Computational experience with confidence regions and confidence intervals for nonlinear least squares. *Technometrics*, 27(1):67–82, 1987.

S.G. Douma and P.M.J. Van den Hof. An alternative paradigm for probabilistic uncertainty bounding in prediction error identification. In *Proc. 44th IEEE Conf. Decision and Control and European Control Conference ECC'05, CDC-ECC'05*, pages 4970–4975, Sevilla, Spain, December 12-15, 2005.

S.G. Douma and P.M.J. Van den Hof. Probabilistic model uncertainty bounding: An approach with finite-time perspectives. In *Preprints 14th IFAC Symposium on System Identification*, pages 1021–1026, Newcastle, Australia, March 27-29, 2006.

R. A. Fisher. On the mathematical foundations of theoretical statistics. *Phil. Trans. Roy. Soc. London, Series A*, 222:309–368, 1922.

R.A. Gallant. *Nonlinear statistical models*. John Wiley and Sons, Inc., New York, 1987.

D. Hamilton. Confidence regions for parameter subsets in nonlinear regression. *Biometrika*, 73(1):57–64, 1986.

S.M. Kay. *Fundamentals of Statistical Signal Processing, Volume II Detection Theory*. Prentice Hall PTR, Upper Saddle River, New Jersey, 1998.

L. Ljung. *System Identification - Theory for the User*. Prentice Hall, Upper Saddle River, NJ, 2nd edition, 1999.

L. Ljung. System identification toolbox in matlab. The Mathworks, Inc., 2003.

A. M. Mood, F. A. Graybill, and D. C. Boes. *Introduction to the Theory of Statistics*. McGraw-Hill, Tokyo, 3rd edition, 1974.

S.L. Quinn, T.J. Harris, and D.W. Bacon. Accounting for uncertainty in control-relevant statistics. *Journal of Process Control*, 15(1):675–690, 2005.

G. A. F. Seber and C. J. Wild. *Nonlinear regression*. John Wiley and Sons, New York, 1989.

R. Tempo, G. Calafiore, and F. Dabbene. *Randomized algorithms for analysis and control of uncertain systems*. Springer Verlag, New York, 2004.

A. Wald. Note on the consistency of the maximum likelihood estimate. *Ann. Math. Stat.*, 20:595–601, 1949.

E. Weyer and M.C. Campi. Non-asymptotic confidence ellipsoids for the least-squares estimate. *Automatica*, 38:1539–1547, 2002.

S.S. Wilks. *Mathematical Statistics*. John Wiley and Sons, Inc., New York, 1962.