Likelihood Based Uncertainty Bounding in Prediction Error Identification using ARX models: A Simulation Study

Arnold J. den Dekker, Xavier Bombois and Paul M.J. Van den Hof

Abstract— The purpose of this paper is to evaluate the reliability and finite sample properties of different likelihood based methods for constructing probabilistic parameter confidence regions in prediction error identification using ARX (Auto Regression with eXogenous inputs) models. The paper presents alternatives for the "classical" approach to constructing probabilistic confidence regions in prediction error identification.

I. INTRODUCTION

Prediction error methods have become a wide-spread technique for system identification. Parametric dynamical models that are identified on the basis of measurement data are usually accompanied by an indication of their reliability. Probabilistic confidence regions for the estimated parameters are generally used as an indication of this reliability (or precision). A $100(1 - \alpha)\%$ confidence region is a region in the parameter space that attempts to "cover" the true parameter with probability $(1 - \alpha)$ [8]. These regions are commonly constructed on the basis of prior information on the data generating system and the noise disturbances acting on the measurement data. The presence of the noise disturbances together with a finite length of measurement data is generally the underlying reason for the finite precision of estimated parameters/models. Apart from its intrinsic importance in classical statistical parameter estimation, the need for quantifying model uncertainties has lately become apparent also in many other fields of model applications. When identified models are used as a basis for model-based control, monitoring, simulation or any other model-based decision-making, then robustness requirements impose additional constraints on model uncertainties, which can be taken into account to guarantee robustness properties of the designed algorithms. Different methods to construct confidence regions for the parameters exist. In prediction error identification, confidence regions are most commonly derived from the (asymptotic) statistical properties of the parameter estimator, see e.g., [7]. Alternatively, the (asymptotic) statistics of the so-called Fisher score and likelihood ratio may be used as a basis for constructing confidence regions, see e.g., [1].

The goal of this paper is to evaluate, validate and compare the reliability of different methods for constructing confidence regions (for finite data lengths) for ARX models. This is done by means of well-chosen Monte-Carlo simulation

The authors are with the Delft Center for Systems and Control, Delft University of Technology, Mekelweg 2, 2628 CD Delft, The Netherlands `a.j.dendekker@tudelft.nl`

experiments, recording whether the true values of the parameters are contained within the confidence regions for each realization of the data.

II. 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 $y(t)$ the stochastic (measurable) output signal, $u(t)$ the deterministic (measurable) input signal and $e(t)$ (non-measurable) 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 [7]. 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 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

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

The model of the observations is given by

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

with $\epsilon(t,\theta)$ the prediction errors. Since $(\mathcal{S} \in \mathcal{M})$, the prediction errors evaluated at θ_0

$$\epsilon(t,\theta_0) = y(t) - \hat{y}(t|t-1;\theta_0) \tag{8}$$

are equal to $e(t)$ and thus zero mean, independent, normally distributed, with probability density function (PDF)

$$f_\epsilon(\epsilon(t,\theta_0);\theta_0) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}\epsilon^2(t,\theta_0)\right] \tag{9}$$

with σ^2 the variance of $e(t)$ and θ_0 the true value of the parameter vector θ. The joint probability distribution of the observations $y^N = \{y(t)\}_{t=1,\cdots,N}$ (conditioned on the given deterministic input sequence u^N) is given by:

$$\begin{aligned}
f_y(y^N;\theta_0) &= \prod_{t=1}^{N} f_\epsilon(y(t) - \hat{y}(t|t-1;\theta_0)) \\
&= \prod_{t=1}^{N} f_\epsilon(\epsilon(t,\theta_0);\theta_0).
\end{aligned} \tag{10}$$

Taking the logarithm yields:

$$\log f_y(y^N;\theta_0) = \sum_{t=1}^{N} \log f_\epsilon(\epsilon(t,\theta_0);\theta_0), \tag{11}$$

which can be written as

$$\log f_y(y^N;\theta_0) = -\frac{N}{2}\log(2\pi) - N\log\sigma - \frac{N}{2\sigma^2}V_N(\theta_0) \tag{12}$$

with

$$V_N(\theta_0) = \frac{1}{N}\sum_{t=1}^{N}\epsilon(t,\theta_0)^2. \tag{13}$$

A. The Fisher score

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

$$S(\theta) = \frac{\partial \log f_y(y^N;\theta)}{\partial \theta} = \frac{-N}{2\sigma^2}\frac{\partial V_N(\theta)}{\partial \theta}. \tag{14}$$

It can be shown that the Fisher score (14) evaluated at the true value θ_0 of θ has mean zero [8]:

$$\mathbb{E}[S(\theta_0)] = 0. \tag{15}$$

B. The Fisher information matrix

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

$$\begin{aligned}
F(\theta_0) &= \mathbb{E}\left[S(\theta_0)S^T(\theta_0)\right] \\
&= \frac{N^2}{4\sigma^4}\mathbb{E}\left[\left.\left(\left(\frac{\partial V_N(\theta)}{\partial \theta}\right)\left(\frac{\partial V_N(\theta)}{\partial \theta}\right)^T\right)\right|_{\theta=\theta_0}\right]
\end{aligned} \tag{16}$$

which is known as the Fisher information matrix [4]. It can be shown that $F(\theta_0)$ may alternatively be written as

$$\begin{aligned}
F(\theta_0) &= -\mathbb{E}\left[\left.\frac{\partial^2 \log f_y(y^N;\theta)}{\partial \theta^2}\right|_{\theta=\theta_0}\right] \\
&= \frac{N}{2\sigma^2}\mathbb{E}\left[\left.\frac{\partial^2 V_N(\theta)}{\partial \theta^2}\right|_{\theta=\theta_0}\right]
\end{aligned} \tag{17}$$

Furthermore, by the multivariate central limit theorem, it is generally derived that, for $N \to \infty$,

$$S(\theta_0) \to \mathcal{N}(0, F(\theta_0)), \tag{18}$$

that is, the Fisher score is asymptotically normally distributed with expectation value zero and covariance matrix $F(\theta_0)$ [10].

C. The likelihood function and the maximum likelihood estimator

Next, suppose that we substitute the available observations y^N for the corresponding indeterminate variables in (10) and regard the resulting expression as a function of the parameter vector θ for fixed observations y^N. To express this, we write $f_y(\theta;y^N)$. This function is called the likelihood function. The maximum likelihood estimator (MLE) of θ_0 is given by

$$\hat{\theta}_N = \arg\max_\theta f_y(\theta;y^N) = \arg\min_\theta V_N(\theta). \tag{19}$$

Note that the MLE of σ^2 is given by

$$\hat{\sigma}^2 = V_N(\hat{\theta}_N) \tag{20}$$

It can be shown [4] that, for $N \to \infty$,

$$\hat{\theta}_N \to \mathcal{N}(\theta_0, F^{-1}(\theta_0)). \tag{21}$$

Furthermore, Wald [9] has shown that, under very general conditions, the MLE $\hat{\theta}_N$ is known to be a consistent estimator.

D. Test statistics

The (asymptotic) statistical properties of the Fisher score and the MLE described above provide a basis for constructing confidence regions. As there is a close connection between confidence regions and hypothesis tests, let us first consider the latter. It follows from (18) and (21) that the quadratic forms

$$S(\theta_0)^T F^{-1}(\theta_0)S(\theta_0) \tag{22}$$

and

$$(\hat{\theta}_N - \theta_0)^T F(\theta_0)(\hat{\theta}_N - \theta_0) \tag{23}$$

both have asymptotically (i.e., for $N \to \infty$) a χ_n^2 distribution, i.e., a chi-square distribution with n degrees of freedom, with n the dimension of the parameter vector θ [5], [6]. Then, if we want to test the null hypothesis

$$H_0 : \theta_0 = \theta \tag{24}$$

against the alternative hypothesis

$$H_1 : \theta_0 \neq \theta, \tag{25}$$

the test statistics

$$T_R = S(\theta)^T F^{-1}(\theta) S(\theta), \qquad (26)$$

which is known as the Rao (or score) test statistic [6], and

$$T_T = (\hat{\theta}_N - \theta)^T F(\theta)(\hat{\theta}_N - \theta) \qquad (27)$$

may be used. Asymptotically, both test statistics have a χ_n^2 distribution under H_0. A third test statistic, which is known as the Wald test statistic [6], is given by

$$T_W = (\hat{\theta}_N - \theta)^T F(\hat{\theta}_N)(\hat{\theta}_N - \theta). \qquad (28)$$

Its use is justified by the fact that the MLE is a consistent estimator (i.e., $\hat{\theta}_N \to \theta_0$ if $N \to \infty$). Asymptotically, the test statistic (28) will also have a χ_n^2 distribution under H_0 [6]. A fourth test statistic is based on a comparison of maximized likelihood functions under both hypotheses. Since the models underlying these hypotheses are nested, the generalized likelihood ratio

$$L_G = \frac{f_y(\theta; y^N)}{\sup_\theta f_y(\theta; y^N)} = \frac{f_y(\theta; y^N)}{f_y(\hat{\theta}_N; y^N)} \qquad (29)$$

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

$$T_{LR} = -2 \log L_G \qquad (30)$$

has asymptotically a χ_n^2 distribution under H_0 [6]. 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, where the significance level is defined as the probability of rejecting H_0 when H_0 is true. 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 [1] .

E. Parameter inference

Based on the test statistics described in subsection II-D, the following asymptotically valid $100(1-\alpha)\%$ confidence regions for the true parameter θ_0 can be specified:

- Confidence region based on the likelihood ratio test statistic T_{LR}:

$$\theta_0 \quad \in \quad \mathcal{D}_\theta(1-\alpha, \hat{\theta}_N) \text{ w.p. } 1-\alpha, \text{ with}$$

$$\mathcal{D}_\theta(1-\alpha, \hat{\theta}_N) \quad := \quad \left\{ \theta \Big| 2 \log \frac{f_y(\hat{\theta}_N; y^N)}{f_y(\theta; y^N)} \leq \chi_{n,1-\alpha}^2 \right\}$$

$$(31)$$

where $\chi_{n,1-\alpha}^2$ is the $1-\alpha$ quantile of the chi-square distribution with n degrees of freedom (cfr. [8]). Using

(12), $\mathcal{D}_\theta(1-\alpha, \hat{\theta}_N)$ may also be written as:

$$\mathcal{D}_\theta(1-\alpha, \hat{\theta}_N) := \left\{ \theta \Big| V_N(\theta) - V_N(\hat{\theta}_N) \leq \frac{\sigma^2}{N} \chi_{n,1-\alpha}^2 \right\}$$

$$(32)$$

- Confidence region based on the Wald test statistic T_W:

$$\theta_0 \in \mathcal{D}_\theta(1-\alpha, \hat{\theta}_N) \text{ w.p. } 1-\alpha, \text{ with}$$

$$\mathcal{D}_\theta(1-\alpha, \hat{\theta}_N) :=$$

$$\left\{ \theta \Big| (\hat{\theta}_N - \theta)^T F(\hat{\theta}_N)(\hat{\theta}_N - \theta) \leq \chi_{n,1-\alpha}^2 \right\} \qquad (33)$$

- Confidence region based on the Rao test statistic T_R:

$$\theta_0 \quad \in \quad \mathcal{D}_\theta(1-\alpha) \text{ w.p. } 1-\alpha, \text{ with}$$

$$\mathcal{D}_\theta(1-\alpha) \quad := \quad \left\{ \theta | S(\theta)^T F^{-1}(\theta) S(\theta) \leq \chi_{n,1-\alpha}^2 \right\}$$

$$(34)$$

Note that this confidence region can be constructed without calculation of the MLE $\hat{\theta}_N$. However, its construction is computationally expensive, requiring the evaluation of $S(\theta)$ and $F(\theta)$ at a sufficient number of points to produce a contour.

- Confidence region based on the test statistic T_T:

$$\theta_0 \in \mathcal{D}_\theta(1-\alpha, \hat{\theta}_N) \text{ w.p. } 1-\alpha, \text{ with}$$

$$\mathcal{D}_\theta(1-\alpha, \hat{\theta}_N) :=$$

$$\left\{ \theta \Big| (\hat{\theta}_N - \theta)^T F(\theta)(\hat{\theta}_N - \theta) \leq \chi_{n,1-\alpha}^2 \right\} \qquad (35)$$

Note that the construction of (35), requiring repeated evaluation of $F(\theta)$, is computationally expensive as well.

Sometimes calculating $F(\theta)$, which is sometimes denoted as the *expected* information matrix, is difficult and the so-called *observed* Fisher information matrix, given by

$$J(\theta) \quad = \quad -\left(\frac{\partial^2 \log f_y(\theta; y^N)}{\partial \theta^2} \right) \qquad (36)$$

is used instead. The observed Fisher information matrix can be calculated by evaluating the likelihood function. If the derivatives can not be found analytically, they can be approximated by finite differences.

In the remainder of this paper, we will assume the noise variance σ^2 to be known, but the analysis that follows can easily be extended to include the case of unknown noise variance.

III. ARX MODELLING

The ARX model set is determined by

$$G(q, \theta) = \frac{q^{-n_k} B(q^{-1}, \theta)}{A(q^{-1}, \theta)}, \quad H(q, \theta) = \frac{1}{A(q^{-1}, \theta)}, \quad (37)$$

with q^{-1} the backward shift operator, n_k the delay, and

$$A(q^{-1}, \theta) \quad = \quad 1 + a_1 q^{-1} + \cdots + a_{n_a} q^{-n_a}$$

$$B(q^{-1}, \theta) \quad = \quad b_0 + b_1 q^{-1} + \cdots + b_{n_b-1} q^{-n_b+1} \quad (38)$$

with $\theta^T = [a_1 \cdots a_{n_a} b_0 \cdots b_{n_b-1}]$. For an ARX model structure the one-step ahead predictor can be written as [7]

$$\hat{y}(t|t-1;\theta) = \varphi^T(t)\theta \qquad (39)$$

with

$$\varphi^T(t) = [-y(t-1)\cdots-y(t-n_a)u(t-n_k)\cdots u(t-n_k-n_b+1)]. \qquad (40)$$

having dimension $n = n_a + n_b$. Let us denote

$$\Phi = \begin{pmatrix} \varphi^T(1) \\ \vdots \\ \varphi^T(N) \end{pmatrix} \text{ and } \mathbf{y} = [y(1)\cdots y(N)]^T \qquad (41)$$

If the data generating system belongs to the model class ($\mathcal{S} \in \mathcal{M}$), then it holds that $\mathbf{y} = \Phi\theta_0 + \mathbf{e}$, with $\mathbf{e}$ an N dimensional vector of samples from a white noise process. If we furthermore assume that this white noise is Gaussian distributed (with variance σ^2), such as we do in this paper, it follows that the MLE $\hat{\theta}_N$ of θ_0 is obtained by minimizing the quadratic prediction error criterion

$$\hat{\theta}_N = \arg\min_\theta V_N(\theta); \quad V_N(\theta) = \frac{1}{N}\sum_{t=1}^N \epsilon(t,\theta)^2, \qquad (42)$$

with $\epsilon(t,\theta) = y(t) - \hat{y}(t|t-1;\theta)$. Then it follows that

$$\hat{\theta}_N = (\Phi^T\Phi)^{-1}\Phi^T\mathbf{y} = \theta_0 + (\Phi^T\Phi)^{-1}\Phi^T\mathbf{e} \qquad (43)$$

Furthermore, the expressions

$$V_N(\theta) = \frac{1}{N}(\mathbf{y} - \Phi\theta)^T(\mathbf{y} - \Phi\theta), \qquad (44)$$

$$S(\theta) = \frac{-N}{2\sigma^2}\frac{\partial V_N(\theta)}{\partial\theta} = \frac{1}{\sigma^2}\Phi^T(\mathbf{y} - \Phi\theta), \qquad (45)$$

$$\begin{aligned} F(\theta_0) &= \frac{N^2}{4\sigma^4}\mathbb{E}\left[\left(\frac{\partial V_N(\theta)}{\partial\theta}\right)\left(\frac{\partial V_N(\theta)}{\partial\theta}\right)^T\bigg|_{\theta=\theta_0}\right] \\ &= \frac{1}{\sigma^4}\mathbb{E}\left[\Phi^T\mathbf{e}\mathbf{e}^T\Phi\right], \end{aligned} \qquad (46)$$

and

$$J = \frac{1}{\sigma^2}\Phi^T\Phi \qquad (47)$$

can be derived for the quadratic prediction error criterion, the Fisher score, the expected Fisher information matrix and the observed Fisher information matrix, respectively. Note that the observed Fisher information matrix (47) is independent of θ. Since $e(t)$ is a white noise sequence and $\varphi^T(t)$ is uncorrelated with $e(s)$ for $s > t$ (cfr. [7]), (46) simplifies to

$$F(\theta_0) = \frac{1}{\sigma^2}\mathbb{E}\left[\Phi^T\Phi\right]. \qquad (48)$$

If we know the true system, the term $\mathbb{E}\left[\Phi^T\Phi\right]$ in (48) can be calculated analytically. This can be seen as follows. It can be shown that

$$\varphi(t) = s_u(t) + s_e(t), \qquad (49)$$

where the $n \times 1$ vector $s_u(t)$ is given by

$$s_u(t) = \frac{\Lambda_G(q^{-1},\theta_0)}{H(q^{-1},\theta_0)}u(t) \qquad (50)$$

with $\Lambda_G(q^{-1},\theta)$ the $n \times 1$ gradient vector of the transfer function $G(q^{-1},\theta)$ with respect to θ. Equivalently, the $n \times 1$ vector $s_e(t)$ in (49) is given by

$$s_e(t) = \frac{\Lambda_H(q^{-1},\theta_0)}{H(q^{-1},\theta_0)}e(t), \qquad (51)$$

with $\Lambda_H(q^{-1},\theta)$ the $n \times 1$ gradient vector of the transfer function $H(q^{-1},\theta)$ with respect to θ. Now, let us define $R(\theta_0) = \mathbb{E}\left[\Phi^T\Phi\right]$. Then it can be shown that

$$R(\theta_0) = \sum_{t=1}^N s_u(t)s_u^T(t) + \sum_{t=1}^N \mathbb{E}\left[s_e(t)s_e^T(t)\right] \qquad (52)$$

Using a state space representation of (51), i.e.,

$$\begin{aligned} x(t+1) &= A(\theta_0)x(t) + K(\theta_0)e(t) \\ s_e(t) &= C(\theta_0)x(t) + D(\theta_0)e(t), \end{aligned} \qquad (53)$$

it follows that, for all t,

$$\mathbb{E}\left[s_e(t)s_e^T(t)\right] = C(\theta_0)P(\theta_0)C^T(\theta_0) + D(\theta_0)D^T(\theta_0)\sigma^2, \qquad (54)$$

with $P(\theta_0) = \mathbb{E}[x(t)x^T(t)]$. $P(\theta_0)$ is obtained as the positive definite solution of the stationary Liapunov equation:

$$P(\theta_0) = A(\theta_0)P(\theta_0)A^T(\theta_0) + K(\theta_0)K^T(\theta_0)\sigma^2. \qquad (55)$$

Using the expressions derived in subsection II-E, the following asymptotically valid $100(1-\alpha)\%$ confidence regions for θ_0 can be derived.

A. Confidence region based on the likelihood ratio test statistic

$$\theta_0 \in \mathcal{D}_\theta(1-\alpha,\hat{\theta}_N) \text{ w.p. } 1-\alpha, \text{ with}$$

$$\mathcal{D}_\theta(1-\alpha,\hat{\theta}_N) :=$$

$$\left\{\theta|V_N(\theta) - V_N(\hat{\theta}_N) \leq \frac{\sigma^2}{N}\chi^2_{n,1-\alpha}\right\} \qquad (56)$$

It can easily been shown (see Appendix) that

$$N(V_N(\theta) - V_N(\hat{\theta}_N)) = (\theta - \hat{\theta}_N)^T\Phi^T\Phi(\theta - \hat{\theta}_N). \qquad (57)$$

Therefore, $\mathcal{D}_\theta(1-\alpha,\hat{\theta}_N)$ in (56) may also be written as:

$$\mathcal{D}_\theta(1-\alpha,\hat{\theta}_N) :=$$

$$\left\{\theta|(\theta - \hat{\theta}_N)^T\Phi^T\Phi(\theta - \hat{\theta}_N) \leq \sigma^2\chi^2_{n,1-\alpha}\right\}. \qquad (58)$$

This confidence region corresponds with the one implemented in the Matlab System Identification Toolbox and the one derived by Douma and Van den Hof using an alternative paradigm for probabilistic uncertainty bounding based on the analysis of data-dependent mappings of the parameter

estimator [2], [3]. Alternatively, using (47), (58) may be written as

$$\mathcal{D}_\theta(1 - \alpha, \hat{\theta}_N) :=$$

$$\left\{ \theta | (\theta - \hat{\theta}_N)^T J(\theta - \hat{\theta}_N) \leq \chi^2_{n,1-\alpha}, \right\} \qquad (59)$$

with J the observed Fisher information matrix.

B. Confidence region based on the Wald test statistic

$$\theta_0 \in \mathcal{D}_\theta(1 - \alpha, \hat{\theta}_N) \text{ w.p. } 1 - \alpha, \text{ with}$$

$$\mathcal{D}_\theta(1 - \alpha, \hat{\theta}_N) :=$$

$$\left\{ \theta | (\hat{\theta}_N - \theta)^T R(\hat{\theta}_N)(\hat{\theta}_N - \theta) \leq \sigma^2 \chi^2_{n,1-\alpha} \right\}, \qquad (60)$$

with $R(\hat{\theta}_N)/\sigma^2$ the expected Fisher information matrix evaluated at $\theta = \hat{\theta}_N$. Note that substitution of the observed Fisher information matrix (47) for the expected Fisher information matrix in (60) yields (58).

C. Confidence region based on the Rao test statistic

$$\theta_0 \in \mathcal{D}_\theta(1 - \alpha, \hat{\theta}_N) \text{ w.p. } 1 - \alpha, \text{ with}$$

$$\mathcal{D}_\theta(1 - \alpha, \hat{\theta}_N) :=$$

$$\left\{ \theta | (\mathbf{y} - \Phi\theta)^T \Phi R^{-1}(\theta) \Phi^T (\mathbf{y} - \Phi\theta) \leq \sigma^2 \chi^2_{n,1-\alpha} \right\} \quad (61)$$

Since it follows from (43) that $\Phi^T(\mathbf{y} - \Phi\theta) = \Phi^T \Phi(\hat{\theta}_N - \theta)$, (61) may also be written as

$$\theta_0 \in \mathcal{D}_\theta(1 - \alpha, \hat{\theta}_N) \text{ w.p. } 1 - \alpha, \text{ with}$$

$$\mathcal{D}_\theta(1 - \alpha, \hat{\theta}_N) :=$$

$$\left\{ \theta | (\hat{\theta}_N - \theta)^T \Phi^T \Phi R^{-1}(\theta) \Phi^T \Phi(\hat{\theta}_N - \theta) \leq \sigma^2 \chi^2_{n,1-\alpha} \right\}. \qquad (62)$$

D. Confidence region based on the test statistic T_T

$$\theta_0 \in \mathcal{D}_\theta(1 - \alpha, \hat{\theta}_N) \text{ w.p. } 1 - \alpha, \text{ with}$$

$$\mathcal{D}_\theta(1 - \alpha, \hat{\theta}_N) :=$$

$$\left\{ \theta | (\hat{\theta}_N - \theta)^T R(\theta)(\hat{\theta}_N - \theta) \leq \sigma^2 \chi^2_{n,1-\alpha} \right\}. \qquad (63)$$

IV. SOME RESULTS FROM ASYMPTOTIC THEORY

Following [7], an asymptotically valid expression for the covariance matrix of the estimator $\hat{\theta}_N$ is given by

$$P_\theta = \frac{\sigma^2}{2\pi N} \left(\int_{-\pi}^{\pi} \left(\Gamma_G(e^{i\omega}, \theta_0) \Phi_u(\omega) + \Gamma_H(e^{i\omega}, \theta_0) \sigma^2 \right) d\omega \right)^{-1} \qquad (64)$$

with

$$\Gamma_G(e^{i\omega}, \theta_0) = \frac{\Lambda_G(e^{i\omega}, \theta_0) \Lambda_G^*(e^{i\omega}, \theta_0)}{H(e^{i\omega}, \theta_0) H^*(e^{i\omega}, \theta_0)}, \qquad (65)$$

and

$$\Gamma_G(e^{i\omega}, \theta_0) = \frac{\Lambda_H(e^{i\omega}, \theta_0) \Lambda_H^*(e^{i\omega}, \theta_0)}{H(e^{i\omega}, \theta_0) H^*(e^{i\omega}, \theta_0)} \qquad (66)$$

and $\Phi_u(\omega)$ the power spectrum of $u(t)$ (where it has been assumed that $u(t)$ a quasi-stationary signal). An asymptotically valid $100(1 - \alpha)\%$ confidence region for θ_0 is then given by

$$\theta_0 \in \mathcal{D}_\theta(1 - \alpha, \hat{\theta}_N) \text{ w.p. } 1 - \alpha, \text{ with}$$

$$\mathcal{D}_\theta(1 - \alpha, \hat{\theta}_N) :=$$

$$\left\{ \theta | (\hat{\theta}_N - \theta)^T P_\theta^{-1} (\hat{\theta}_N - \theta) \leq \chi^2_{n,1-\alpha} \right\}. \qquad (67)$$

Although P_θ can only be calculated exactly if the true system is known, which is, of course, not the case in practice, expression (64) is often used for experimental design purposes. P_θ is then usually estimated by substitution of some initial value for θ_0 in (64). In our simulation study described in section V, the confidence region (67) will also be included so as to test its validity for finite data lengths.

V. 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 ARX model class:

$$y(t) + a_1 y(t-1) + a_2 y(t-2) = b_0 u(t-1) + b_1 u(t-2) + e(t), \qquad (68)$$

with $a_1 = -1.5578, a_2 = 0.5769, b_0 = 0.1047$ and $b_1 = 0.0872$. 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 $\{e(t)\}$ having a variance $\sigma^2 = 0.5$. 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 ARX structure as the data generating system $(\mathcal{S} \in \mathcal{M})$:

$$G(q, \theta) = \frac{b_0 q^{-1} + b_1 q^{-2}}{1 + a_1 q^{-1} + a_2 q^{-2}}, \qquad (69)$$

$$H(q, \theta) = \frac{1}{1 + a_1 q^{-1} + a_2 q^{-2}} \qquad (70)$$
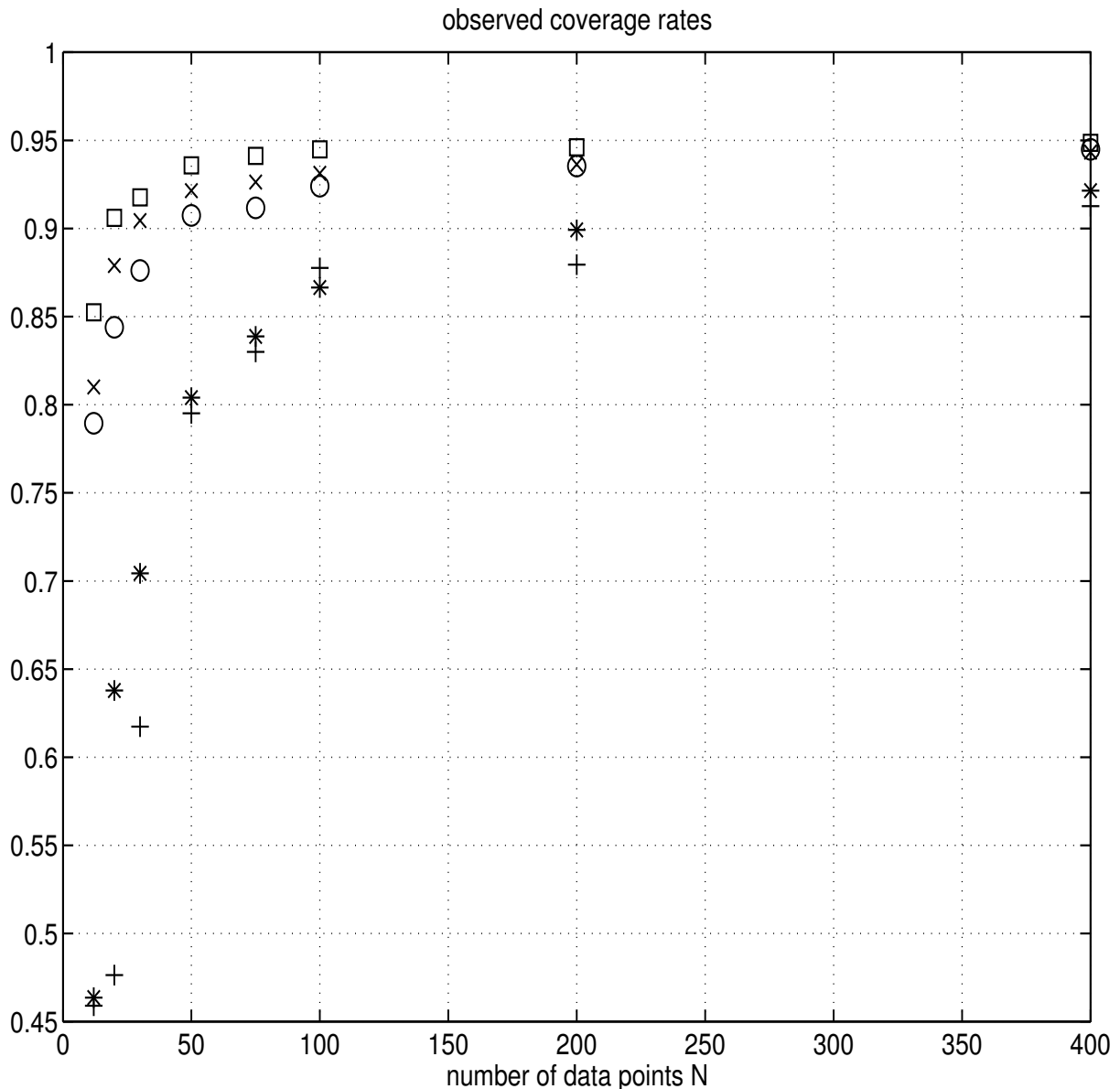
observed coverage rates

Fig. 1. The results of the simulation experiment described in section V, where the data sets are generated using a fixed realization of the input sequence u^N. Observed coverage rates of the confidence regions based on the likelihood ratio test (58)($\square$), the Wald test (60)($\circ$) and the Rao test (61)($\times$), and the confidence regions described by (63)(*) and (67)(+). The data generating system is given by (68) and the ARX model set described by (69) and (70). The nominal confidence level is 0.05. All results have been obtained from 50000 realizations.

with $\theta^T = [a_1 \, a_2 \, b_0 \, b_1]$. Thus for each data set the MLE $\hat{\theta}_N$ was calculated and it was recorded whether or not the confidence regions described by (58), (60), (61), (63) and (67) contained the true value θ_0. Note that determining whether the true parameter values 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 the true parameter values 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 asymptotical 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. For $K = 50000$, 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. Of all confidence regions evaluated, the one based on the likelihood ratio test statistic turns out to be the most reliable one. Furthermore, it is clearly seen that the confidence regions (63) and (67) are unreliable for small N. Since (67) was constructed on the basis of the theoretical, asymptotically
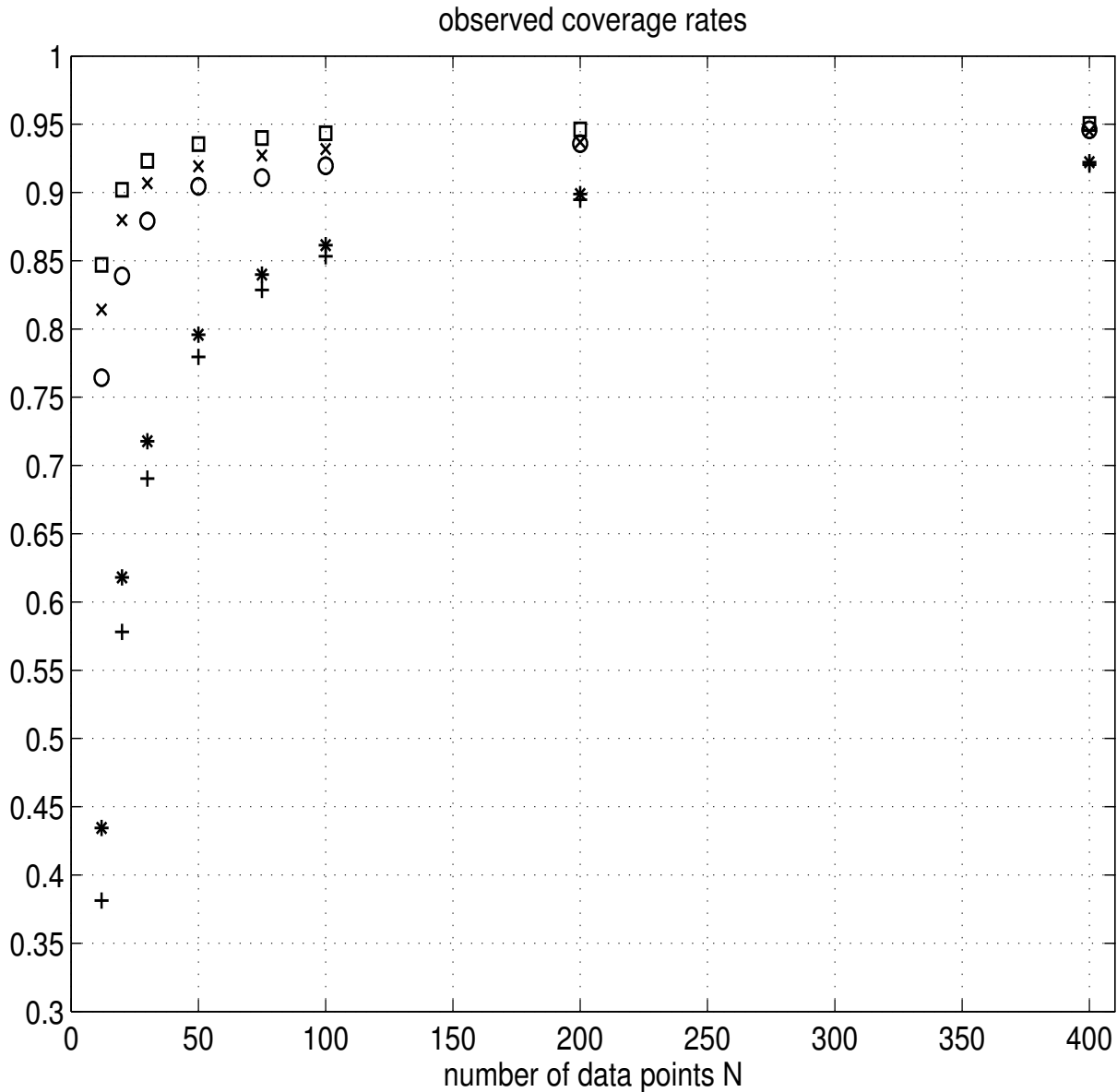
Fig. 2.   The results of the simulation experiment described in section V. Observed coverage rates of the confidence regions based on the likelihood ratio test (58)($\square$), the Wald test (60)($\circ$) and the Rao test (61)($\times$), and the confidence regions described by (63)(*) and (67)(+). The data generating system is given by (68) and the ARX model set described by (69) and (70). Each data set is obtained by generating new realizations of both the input sequence u^N and the noise e^N. The nominal confidence level is 0.05. All results have been obtained from 50000 realizations.

valid, expression for the covariance matrix (64), the results of our simulation experiment indicate that the latter expression is inaccurate for small N.

Next, the whole simulation 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 $K = 50000$ experiments (for each value of N). The results, as shown in Figure 2, are similar to those obtained with a fixed input sequence u^N (see Figure 1).

More simulation experiments have been performed, using alternative model generating systems (all belonging to the ARX model class), parameters and nominal confidence rates. All experiments yielded similar results.

VI. CONCLUSIONS AND FUTURE WORKS

A. Conclusions

Different methods for constructing probabilistic parameter confidence regions in prediction error identification using ARX models have been reviewed and their reliability has been evaluated by means of simulation experiments. In the ARX case, all uncertainty regions considered in this paper assume the form:

$$\left\{\theta | (\hat{\theta}_N - \theta)^T X (\hat{\theta}_N - \theta) \leq \chi^2_{n,1-\alpha}\right\} \qquad (71)$$

with X an $n \times n$ matrix. The results of the simulation experiments indicate that those confidence regions for which

X incorporates information on the unknown noise realization e^N are the most reliable. This information is either contained in the regressor vector Φ (likelihood ratio test (58), Rao test (61)), or in the identified parameter vector $\hat{\theta}_N$ (Wald test (60)). For the remaining confidence regions (i.e., (63) and (67)), X does not incorporate any information on the particular noise realization e^N.

Furthermore, the results of the simulation experiments suggest that the confidence region described by (58), which has been obtained using the observed Fisher information matrix, gives the most reliable results. This method, which can be shown to be based on the likelihood ratio test statistic, outperforms alternative methods evaluated in this paper, including the ones based on the Wald and Rao test statistics, especially for finite data lengths. Moreover, since the construction of (58) only requires the observed Fisher information matrix, the method is not only more reliable but also simpler and computationally less expensive than variants such as (60), (61) and (63), which use the expected Fisher information and/or the Fisher score instead.

Finally, it has been shown that the use of the well known (asymptotically valid) theoretical expression for the covariance matrix (64) to construct confidence regions may lead to unreliable results for small numbers of observations. Therefore, one should be careful when applying expression (64) for experimental design purposes.

B. Future Works

More simulation experiments have to be performed to further substantiate the conclusions of the study described in this paper. Furthermore, the evaluation presented should be extended to include other model structures, such as Output Error and Box Jenkins. Both aspects are subject of current research.

VII. ACKNOWLEDGMENTS

The authors gratefully acknowledge fruitful discussions with Sippe Douma and Robert Bos.

REFERENCES

[1] A. Azzalini, *Statistical Inference - Based on the likelihood*, Chapman & Hall, London, 1996.
[2] S.G. Douma and P.M.J. Van den Hof, "An Alternative Paradigm for Probabilistic Uncertainty Bounding in Prediction Error Identification" *Proc. 44th IEEE Conf. Decision and Control and European Control Conference ECC'05, CDC-ECC'05*, December 12-15, 2005, Sevilla, Spain, pp. 4970-4975.
[3] S.G. Douma and P.M.J. Van den Hof, "Probabilistic Model Uncertainty Bounding: An Approach with Finite-Time Perspectives," *Preprints 14th IFAC Symposium on System Identification*, March 27-29, 2006, Newcastle, Australia, pp. 1021-1026.
[4] R. A. Fisher, On the mathematical foundations of theoretical statistics, *Phil. Trans. Roy. Soc. London, Series A*, vol. 222, 1922, pp. 309-368.
[5] F.A. Graybill, *Matrices with Applications in Statistics,*, Wadsworth, California, 1983.
[6] S.M. Kay, *Fundamentals of Statistical Signal Processing, Volume II Detection Theory*, Prentice Hall, Upper Saddle River, NJ;1998.
[7] L. Ljung, *System Identification - Theory for the User*, 2nd edition, Prentice Hall, Upper Saddle River, NJ, 1999.
[8] A.M. Mood, F.A. Graybill and D.C. Boes *Introduction to the Theory of Statistics*, McGraw-Hill, Tokyo, third edition, 1974.
[9] A. Wald, Note on the consistency of the maximum likelihood estimate, *Ann. Math. Stat.*, vol. 20, 1949, pp. 595-601.
[10] S.S. Wilks *Mathematical Statistics*, John Wiley & Sons, Inc., New York, third edition, 1974.

APPENDIX

Proof of (57).

It follows form (44) that the term $N(V_N(\theta) - V_N(\hat{\theta}_N))$ can be rewritten as

$$(\mathbf{y} - \Phi\theta)^T(\mathbf{y} - \Phi\theta) - (\mathbf{y} - \Phi\hat{\theta}_N)^T(\mathbf{y} - \Phi\hat{\theta}_N) =$$
$$\theta^T\Phi^T\Phi\theta - \mathbf{y}^T\Phi(\theta - \hat{\theta}_N) - (\theta - \hat{\theta}_N)^T\Phi^T\mathbf{y} - \hat{\theta}^T\Phi^T\Phi\hat{\theta}_N. \quad (72)$$

Since it follows from (43), that $\Phi^T\mathbf{y} = \Phi^T\Phi\hat{\theta}_N$, (72) may also be written as

$$\begin{aligned}
&\theta^T\Phi^T\Phi\theta - \hat{\theta}_N^T\Phi^T\Phi(\theta - \hat{\theta}_N) \\
-\; &(\theta - \hat{\theta}_N)^T\Phi^T\Phi\hat{\theta}_N - \hat{\theta}^T\Phi^T\Phi\hat{\theta}_N \\
=\; &(\theta - \hat{\theta}_N)^T\Phi^T\Phi(\theta - \hat{\theta}_N) \quad (73)
\end{aligned}$$

This completes the proof. $\square$