Effect of model structure and signal-to-noise ratio on finite-time uncertainty bounding in prediction error identification

X. Bombois, A.J. den Dekker, M. Barenthin, P.M.J. Van den Hof

***Abstract*— In prediction error identification, confidence regions are most commonly derived from the asymptotic statistical properties of the parameter estimator. Therefore, these confidence regions are only asymptotically valid and, for finite samples, their actual coverage rate can be smaller than the desired coverage rate. In this paper, we analyze the influence of the SNR and of the type of model structure on the difference between the actual and desired coverage rates. In addition, we propose alternatives to the classical approach to constructing probabilistic confidence regions for Box-Jenkins systems.**

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 uncertainty. Probabilistic confidence regions for the system parameters are generally used as an indication of this uncertainty. An (exact) $100\ \alpha\%$ confidence region is a region in the parameter space that contains the true parameter θ_0 with probability α [11]. 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. Confidence regions are indeed extensively used in recent new approaches for robustness analysis [3] and for optimal experiment design for control (see e.g. [9], [4], [2]).

In prediction error identification, confidence regions are most commonly derived from the (asymptotic) statistical properties of the parameter estimator, see e.g., [10]. This leads to ellipsoidal confidence regions U centered at the identified parameter vector and shaped by the covariance matrix of this identified parameter vector. It is important to note that the property used to derive the confidence ellipsoid is an asymptotic property i.e. a property which is only valid when the number N of data tends to infinity ($N \rightarrow \infty$). Consequently, if N is small, the actual coverage rate of U i.e. $Pr(\theta_0 \in U)$ can be smaller than the desired coverage rate α. Based on this observation, Campi and co-authors developed a new confidence region bounding scheme which, at the cost of an heavy computational burden, guarantees the coverage rate also for small values of N (see e.g. [5]). However, perhaps surprisingly, few attention has been paid to the following questions: how significant can the difference between the actual and desired coverage rate be and what are the factors influencing this difference when N is fixed? In our earlier contributions [6], [7], we started this analysis by comparing, for different values of the number N of data, the actual and desired coverage rates when the true system has an ARX or an OE structure. The actual coverage rate is estimated using Monte-Carlo simulations.

When we compare the results in the ARX and OE cases, we observe that, for similar situations, the actual coverage rate is (significantly) closer to the desired one when the true system has an ARX structure than when the true system has an OE structure. A contribution of this paper is to explain this phenomenon by observing that the statistical property used to build the classical confidence region in the OE case requires a first-order Taylor approximation in order to be established while the ARX case does not require this approximation. This first-order approximation is in fact required for all model structures for which the prediction error is not linear (affine) in the parameter vector i.e. OE, ARMAX, BJ. That a first-order approximation is required in these cases is a well-known fact [10], however, the consequences of this first-order approximation on the actual coverage rate $Pr(\theta_0 \in U)$ of the confidence region U are, to the best of our knowledge, unexplored. Another contribution of this paper is to show that the actual coverage rate for BJ, ARMAX or OE structures, which, for small value of N, differs from the desired rate α more severely than in the ARX case, can nevertheless (significantly) be improved by increasing the power of the excitation signal i.e. by increasing the signal-to-noise ratio.

If, for some reasons, both the number of data and the signal-to-noise ratio of an identification experiment is small, the reliability of the obtained confidence region U will be poor i.e. $Pr(\theta_0 \in U) << \alpha$. In these cases, alternative uncertainty bounding schemes should be considered. In our previous contributions [6], [7], we presented such alternative schemes for ARX and OE true systems. In this paper, we extend these results to the more realistic case of a Box-Jenkins true system. In these new schemes, instead of using the (asymptotic) statistics of the identified parameter vector $\hat{\theta}_N$, we use the (asymptotic) statistics of the so-called Fisher score and likelihood ratio as a basis for constructing confidence regions, see e.g., [1]. These new schemes can be considered as alternatives for the uncertainty bounding scheme of [5].

X. Bombois, A.J. den Dekker and P.M.J. Van den Hof are with the Delft Center for Systems and Control, Delft University of Technology, Delft, The Netherlands, x.j.a.bombois@tudelft.nl

M. Barenthin is with the Automatic Control, School of Electrical Engineering, KTH, Stockholm, Sweden

II. Classical uncertainty bounding

We suppose that we have collected N input-output data $Z^N = \{y(t),\ u(t) \mid t = 1...N\}$ from the following data-generating system:

$$\mathcal{S}:\ y(t) = G_0(z)u(t) + H_0(z)e(t) \tag{1}$$

where $e(t)$ is a white noise sequence with variance σ_e^2 independent of $u(t)$, and $G_0(z)$ and $H_0(z)$ are two unknown stable LTI transfer functions. The transfer function $H_0(z)$ is furthermore supposed to be monic and inversely stable. We also suppose that we have chosen a model structure $\mathcal{M} = \{G(z,\theta),\ H(z,\theta) \mid \theta \in \mathbf{R}^k\}$ which contains the true system i.e. there exists a parameter vector θ_0 such that $G(z,\theta_0) = G_0(z)$ and $H(z,\theta_0) = H_0(z)$. Depending on the true system, different types of model structures are available: OE, ARX, ARMAX and BJ model structures.

Using the model structure $\mathcal{M}$ and the data Z^N, it is possible to deduce an estimate $\hat{\theta}_N$ of θ_0 using classical prediction error theory [10]: $\hat{\theta}_N = arg\min_\theta V_N(\theta)$ with

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

and with $\epsilon(t,\theta)$ the so-called prediction error defined as: $\epsilon(t,\theta) = H^{-1}(z,\theta)\ \ (y(t) - G(z,\theta)u(t))$.

Besides an estimate $\hat{\theta}_N$ of θ_0, the classical prediction error theory allows us to build around $\hat{\theta}_N$ an ellipsoidal confidence region for θ_0, and this by making use of the asymptotic statistics of $\hat{\theta}_N$. Indeed, it can be proven that, asymptotically (i.e. for $N \to \infty$), the identified parameter vector $\hat{\theta}_N$ is normally distributed around θ_0 with a covariance matrix that can be estimated by

$$\begin{gathered} P = \sigma_e^2\ \left(\Psi^T\Psi\right)^{-1} \\ \text{with } \Psi^T = \left(\psi(1,\hat{\theta}_N),\ \psi(2,\hat{\theta}_N),...,\psi(N,\hat{\theta}_N)\right) \end{gathered} \tag{3}$$

and $\psi(t,\theta) \triangleq -\frac{\partial\epsilon(t,\theta)}{\partial\theta}$. This asymptotic property can also be equivalently rewritten as:

$$\frac{1}{\sigma_e^2}\left(\hat{\theta}_N - \theta_0\right)^T \Psi^T\Psi\left(\hat{\theta}_N - \theta_0\right) \sim As\chi_k^2 \tag{4}$$

where χ_k^2 is the χ^2-distribution with k degrees of freedom (k is the dimension of θ). This property leads to the the following (asymptotically valid) $100\alpha\%$ confidence region for the true parameter vector θ_0: $\{\theta \mid \frac{1}{\sigma_e^2}(\theta - \hat{\theta}_N)^T\Psi^T\Psi(\theta - \hat{\theta}_N) < \chi\}$ with χ defined as $Pr(\chi_k^2 < \chi) = \alpha$. In this expression, the variance σ_e^2 of the noise is generally unknown and is therefore often replaced by its estimate $\hat{\sigma}_e^2 = \frac{1}{N-k}\sum_{t=1}^{N}\epsilon^2(t,\hat{\theta}_N)$. In this case, the χ_k^2-distribution in (4) becomes a F-distribution with degrees of freedom k and $N-k$ (see [10][page 558]) i.e.

$$\frac{1}{\hat{\sigma}_e^2}\frac{1}{k}\left(\hat{\theta}_N - \theta_0\right)^T \Psi^T\Psi\left(\hat{\theta}_N - \theta_0\right) \sim AsF_{k,N-k} \tag{5}$$

Consequently, if we choose β such that $Pr(F_{k,N-k} < \frac{\beta}{k}) = \alpha$, we obtain the following (asymptotically valid) $100\alpha\%$ confidence region for the true parameter vector θ_0:

$$U = \{\theta \mid \frac{1}{\hat{\sigma}_e^2}(\theta - \hat{\theta}_N)^T\Psi^T\Psi(\theta - \hat{\theta}_N) < \beta\} \tag{6}$$

It is important to note that the property (4) used to derive the confidence ellipsoid U (via (5)) is an asymptotic property. Consequently, if N is small, the actual coverage rate of U i.e. $Pr(\theta_0 \in U)$ can be smaller than the desired coverage rate α. This leads us to the question: Given a number N of data, what are the factors influencing the difference between the actual and desired coverage rates? This analysis should in particular explain the significant difference between the result for ARX and OE true systems that we observe when comparing the results in [6] and in [7]. In these papers, we indeed observe that, for equal N, the actual coverage rate is (significantly) smaller when the true system has an OE structure ($H_0 = 1$) than when the true system has an ARX structure (H_0 is one over the denominator of G_0). More precisely, we considered in [6] and in [7] the same plant system[1] G_0 and, e.g. for $N = 50$, the observed coverage rate was 0.93 in the ARX case while this coverage rate is only 0.55 in the OE case.

III. Factors influencing the actual coverage rate

In order to determine the factors influencing the difference between actual and desired coverage rates, it is important to understand why property (4) is only valid for $N \to \infty$. For this purpose, we could use the reasoning in [10][Chapter 9]. However, here, we prefer to use the following (simpler) alternative reasoning which is a generalization of the reasoning in [8].

To prove (4), a first step is to deduce the following first order Taylor approximation of $\epsilon(t,\theta_0)$ around the identified $\hat{\theta}_N$:

$$\epsilon(t,\theta_0) \approx \epsilon(t,\hat{\theta}_N) + \frac{\partial\epsilon(t,\theta)^T}{\partial\theta}|_{\hat{\theta}_N}\ (\theta_0 - \hat{\theta}_N) \tag{7}$$

By noticing that $\epsilon(t,\theta_0) = e(t)$ (the true system lies in the model structure) and by using the definition of $\psi(t,\theta)$, we obtain:

$$\epsilon(t,\hat{\theta}_N) \approx e(t) - \psi^T(t,\hat{\theta}_N)\ (\hat{\theta}_N - \theta_0) \tag{8}$$

Now, since $\hat{\theta}_N$ minimizes the cost function $V_N(\theta)$, we have that the gradient of this cost function is zero when evaluated at $\hat{\theta}_N$. This delivers thus the following equality:

$$\frac{1}{N}\sum_{t=1}^{N}\psi(t,\hat{\theta}_N)\epsilon(t,\hat{\theta}_N) = 0 \tag{9}$$

[1] With respect to the OE case, the signal-to-noise ration (SNR) (i.e. $\frac{\|H_0(z)e(t)\|_2^2}{\|u(t)\|_2^2}$) was 30 times smaller in the ARX case. Consequently, the bad results in the OE case cannot be due to a bad signal-to-noise ratio.

Replacing $\epsilon(t, \hat{\theta}_N)$ in (9) by its approximation (8) yields:

$$\left(\Psi^T \Psi\right) \ (\hat{\theta}_N - \theta_0) \approx \Psi^T \mathbf{e} \tag{10}$$

with $\mathbf{e}^T = (e(1), e(2), ..., e(N))$ and Ψ defined in (3). Let us introduce the *economy size* singular value decomposition of Ψ^T: $\Psi^T = U \Sigma V^T$ where Σ is a diagonal matrix of dimension k and $U \in \mathbf{R}^{k \times k}$, $V^T \in \mathbf{R}^{k \times N}$ are such that $U^T U = V^T V = I_k$. The expression (10) can then be rewritten as:

$$V^T \Psi (\hat{\theta}_N - \theta_0) \approx V^T \mathbf{e} \tag{11}$$

Using the central limit theorem, it can be shown [13] that the right-hand side of (11) is asymptotically a normally distributed random vector with zero mean and covariance matrix $\sigma_e^2 I_k$. Consequently, the left-hand side is also (asymptoticaly) a normally distributed random vector with zero mean and a covariance matrix $\sigma_e^2 I_k$. We can thus write that $\rho = \frac{1}{\sigma_e} V^T \Psi (\hat{\theta}_N - \theta_0) \sim As\mathcal{N}(0, I_k)$ and thus that $\rho^T \rho \sim As\chi_k^2$. This observation and the fact that $\Psi^T = U \Sigma V^T$ with $V^T V = I_k$ leads to the distribution (4) (which can be transformed into (5) by replacing σ_e^2 by $\hat{\sigma}_e^2$).

It is important to realize that the property (4) (and thus also the property (5)) is obtained by making two approximations i.e.

1) the first-order approximation (8). Note that the impact of this approximation reduces for large value of N and disappears when $N \to \infty$ since $\hat{\theta}_N$ is a consistent estimate of θ_0
2) the fact that $V^T \mathbf{e}$ is $\mathcal{N}(0, \sigma_e^2 I_k)$ only for $N \to \infty$.

In [8], [13], simulations show that $V^T \mathbf{e}$ converges relatively fast to a normal distribution. In the following two subsections, we will analyze the factors influencing the Taylor approximation (8).

A. the ARX case

Suppose that both the true system and the model structure are ARX (with a given delay n_k) i.e.

$$G(z, \theta) = \frac{B(z,\theta)}{A(z,\theta)} \quad H(z, \theta) = \frac{1}{A(z,\theta)}$$
$$B(z, \theta) = z^{-n_k} \ \left(b_0 + b_1 z^{-1} + ... + b_{n_b} z^{-n_b}\right)$$
$$A(z, \theta) = 1 + a_1 z^{-1} + ... + a_{n_a} z^{-n_a}$$
$$\theta^T = (a_1 \ a_2 \ ... \ a_{n_a} \ b_0 \ ... \ b_{n_b})$$

In this case, the true system can be rewritten as:

$$y(t) = \phi(t)^T \theta_0 + e(t) \tag{12}$$

with $\phi(t) = (-y(t-1) \ ... \ -y(t-n_a) \ u(t-n_k) \ ... \ u(t-n_b-n_k))^T$ the so-called regression vector. Moreover, we have that:

$$\epsilon(t, \theta) = y(t) - \phi(t)^T \theta \tag{13}$$

Consequently, using (12) and (13), we can write: $\epsilon(t, \hat{\theta}_N) = e(t) - \phi^T(t) \ (\hat{\theta}_N - \theta_0)$, and we see thus that (8) is not an approximation in the ARX case.

This observation constitutes an explanation for the phenomenon observed in [6], [7] i.e. the fact that the actual coverage rate of the ellipsoidal confidence region U is closer to the desired one in the ARX case than in the OE case (where (8) is an approximation).

B. the role of the signal-to-noise ratio for non-ARX model structures

Let us now consider the case where (8) is an approximation (BJ, OE, ARMAX systems). In order to reduce the impact of the approximation (8) on the actual coverage rate of the confidence region U, we need to ensure that the identified parameter vector $\hat{\theta}_N$ is close to the true parameter vector θ_0. As mentioned above, this can of course be achieved by taking a large number of data. However, this is of course not the only possibility. Indeed, for a given (small) value of N, $\hat{\theta}_N$ can also be made closer[2] to θ_0 by increasing the power of the input signal $u(t)$, or in other words by increasing the signal-to-noise ratio (SNR).

Consequently, we would expect that the actual coverage rate becomes closer to the desired one if, for a given value of N, we increase the signal-to-noise ratio. This claim will be verified in the next subsection via simulations. Equivalently, in the light of Section III-A and if we can neglect the eventual influence of the SNR on the convergence rate of $V^T \mathbf{e}$, we expect that the chosen SNR will not affect the actual coverage rate in the ARX case since (8) is then not an approximation.

C. Simulation result

In order to verify our results of the previous section, we will analyze the influence of the SNR on the actual coverage rate of the confidence region U when U is designed to contain the true parameter vector with a probability of 0.95 (i.e. the desired coverage rate α is 0.95). For this purpose, we consider first the same OE true system as in [7]:

$$\mathcal{S}_{OE} : y(t) = \overbrace{\frac{z^{-1}(b_0 + b_1 z^{-1})}{1 + f_1 z^{-1} + f_2 z^{-2}}}^{G_0(z)} u(t) + e(t),$$

with $e(t)$ a white noise of variance $\sigma_e^2 = 1$ and with $b_1 = 0.1047, b_2 = 0.0872, f_1 = -1.5578, \ f_2 = 0.5769$. The true parameter vector $\theta_0 \in \mathbf{R}^4$ is here given by: $(b_0, b_1, f_1, f_2)^T$.

For different data lengths N, we perform K Monte-Carlo simulations on the system $\mathcal{S}_{OE}$. More precisely, K data sets $Z^N = \{y(t), \ u(t) \mid t = 1...N\}$ are generated using $\mathcal{S}_{OE}$ with $u(t)$ a white noise with given variance σ_u^2. From each data set, $\hat{\theta}_N \in \mathbf{R}^4$ is identified using a full-order

[2]We neglect the bias of the finite-time estimate $\hat{\theta}_N$ i.e. the fact that, for finite N, $E\hat{\theta}_N \neq \theta_0$.

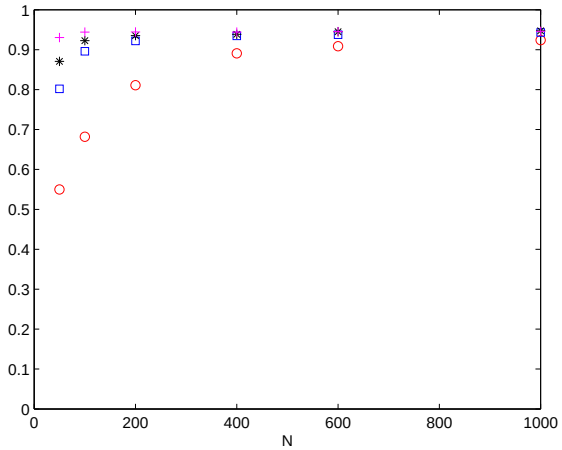

Fig. 1. OE system: observed coverage rate ($K = 10000$) versus the number of data points when $\sigma_u^2 = 1$ (red 'o'), when $\sigma_u^2 = 5$ (blue '□'), when $\sigma_u^2 = 10$ ('black $*$') and when $\sigma_u^2 = 100$ ('magenta $+$'). The desired coverage rate is 0.95.

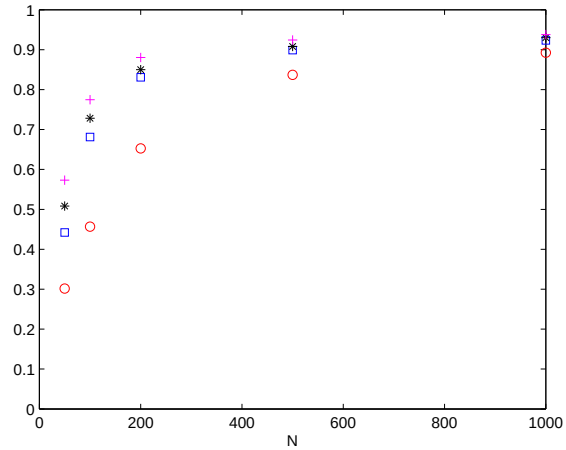

Fig. 2. BJ system: observed coverage rate ($K = 10000$) versus the number of data points when $\sigma_u^2 = 1$ (red 'o'), when $\sigma_u^2 = 5$ (blue '□'), when $\sigma_u^2 = 10$ ('black $*$') and when $\sigma_u^2 = 100$ ('magenta $+$'). The desired coverage rate is 0.95.

OE model structure and it is recorded whether or not the corresponding confidence region U contains the true value θ_0. The observed coverage $\gamma_{0.95}$ is then defined as the percentage of the total number of data sets K, for which the true parameter values θ_0 lay within the confidence region. In this study, we used $K = 10000$.

The circles 'o' in Figure 1 represent the observed coverage rates $\gamma_{0.95}$ for different values of N when the power σ_u^2 of the input signal is equal to 1. The whole Monte-Carlo procedure is then repeated for other values of σ_u^2 (i.e. for other values of the SNR) and the corresponding values of $\gamma_{0.95}$ are also represented in Figure 1. As expected, we observe that, for a given value of N, the actual coverage rate (estimated here by $\gamma_{0.95}$) becomes closer to the desired coverage rate of 0.95 when σ_u^2 increases. It is to be noted that the coverage rates we obtain for high values of σ_u^2 are very close to the coverage rates we obtain when the same system has the ARX form [6]. Note that, in the ARX case, the simulations show that the coverage rate is independent of the SNR.

We have also repeated the same operation for a BJ true system (1) for which $G_0(z)$ is identical as in $\mathcal{S}_{OE}$ and $H_0(z) = (1 + c_1 z^{-1} + c_2 z^{-2})(1 + d_1 z^{-1} + d_2 z^{-2})$ with $c_1 = 0.02$, $c_2 = 0.07$, $d_1 = -1.2$ and $d_2 = 0.63$. The variance σ_e^2 of $e(t)$ is here also chosen equal to one. The corresponding results are given in Figure 2. In this figure, we observe the same increase of the reliability of U when the SNR increases. We observe also that, for given N and σ_u^2, $Pr(\theta_0 \in U)$ is smaller in the BJ case than in the OE case. However, for a fair comparison, we have to note that $\|H_0(z)e(t)\|_2^2$ is four times larger in the BJ case. Let us thus compare the results for similar SNR and observe that they are then very close. For example, for $N = 200$, $Pr(\theta_0 \in U) = 0.83$ when $\sigma_u^2 = 5$ in the BJ case and $Pr(\theta_0 \in U) = 0.81$ for $\sigma_u^2 = 1$ in the OE case.

IV. Alternative uncertainty bounding schemes for BJ systems

If, for some reasons, both the number of data and the signal-to-noise ratio of an identification experiment are small, the reliability of the obtained confidence region U will be poor i.e. $Pr(\theta_0 \in U) << \alpha$ (see Figures 1 and 2). In those cases, alternative uncertainty bounding schemes are required. In [6], [7], we presented several of these alternative uncertainty bounding schemes for the special cases where the true system has an ARX or an OE model structure. In this contribution we extend those results to the more realistic case of a Box-Jenkins (BJ) true system.

Unlike the classical approach, the alternative results presented in [6], [7] use the (asymptotic) properties of other quantities than $\hat{\theta}_N$ as a basis to construct confidence regions for θ_0. These other quantities (also called test statistics) are here the so-called likelihood ratio and the so-called Fisher score. Note that, like (4), the statistical properties of these quantities are generally also asymptotic and that, consequently, the actual coverage rate for finite N can be different from the desired coverage rate α. However, our former results for the ARX and OE cases show that the reliability of these alternative confidence region can outperform the reliability of the classical confidence region U.

Let us first introduce some concepts. We suppose the noise $e(t)$ in (1) to be Gaussian and the input sequence applied to the true system to generate the data Z^N is considered as deterministic. In this case, it can be shown [10][page 216] that the joint probability distribution $f(y^N; \theta_0)$ of the output

sequence $y^N = \{y(t) \mid t = 1...N\}$ generated by (1) is given by:

$$f(y^N;\theta_0) = \prod_{t=1}^{N} \frac{1}{\sqrt{2\pi\sigma_e^2}} \exp\left[-\frac{1}{2\sigma_e^2}\epsilon^2(t,\theta_0)\right] \tag{14}$$

For a given output sequence y^N, we can consider (14) as a function of θ and this function of θ is called the likelihood function and is denoted by $f(\theta; y^N)$. Note that the prediction error estimate $\hat{\theta}_N$ is here also the maximum likelihood estimate i.e. $\hat{\theta}_N = arg\ \min_\theta\ f(\theta; y^N)$. Note also that:

$$\log f(\theta; y^N) = -\frac{N}{2}\log(2\pi) - N\log\sqrt{\sigma_e^2} - \frac{N}{2\sigma_e^2}V_N(\theta) \tag{15}$$

A. Likelihood ratio-based uncertainty bounding for BJ systems

Since, for Gaussian $e(t)$, the prediction error estimate $\hat{\theta}_N$ is equivalent to the maximum likelihood estimate, the so-called generalized likelihood ratio is defined as:

$$T_{LR} = -2\ log\left(\frac{f(\theta_0; y^N)}{f(\hat{\theta}_N; y^N)}\right)$$

Using (15), the expression for T_{LR} can be simplified:

$$T_{LR} = \frac{N}{\sigma_e^2}\ \left(V_N(\theta_0) - V_N(\hat{\theta}_N)\right). \tag{16}$$

If we regard the observations y^N as random variables, it can be proven that the generalized likelihood ratio T_{LR} is χ_k^2-distributed when $N \to \infty$ [11]. Consequently, the following (asymptotically valid) $100\alpha\%$ confidence region for θ_0 can be derived: $\left\{\theta \mid \frac{N}{\sigma_e^2}\ \left(V_N(\theta) - V_N(\hat{\theta}_N)\right) \leq \chi\right\}$, where χ is determined such that $Pr(\chi_k^2 < \chi) = \alpha$. If we replace the unknown σ_e^2 by $\hat{\sigma}_e^2$, the χ^2-distribution becomes a F-distribution (see also Section II). This leads to the following (asymptotically valid) $100\alpha\%$ confidence region [7]

$$U_{LR} = \left\{\theta \mid \frac{N}{\hat{\sigma}_e^2}\ \left(V_N(\theta) - V_N(\hat{\theta}_N)\right) \leq \beta\right\} \tag{17}$$

where β is defined as in (6). It is to be noted that U_{LR} is no longer an ellipsoid and that its construction is generally computationally expensive, requiring the evaluation of $V(\theta)$ at a sufficient number of points to produce contours. Observe also that the expression for U_{LR} is the same for each model structure. This will not be the case for the uncertainty region based on the Fisher score that will be presented in the next section and where the expression will have to be particularized to the BJ-case.

B. Fisher score-based uncertainty bounding for BJ systems

The Fisher score-based uncertainty bounding is based on the statistical properties of the so-called Fisher score. When $e(t)$ is Gaussian, the Fisher score $S(\theta)$ is defined as [12]:

$$S(\theta) = \frac{-N}{2\sigma^2}\frac{\partial V_N(\theta)}{\partial\theta}. \tag{18}$$

It can be shown that the Fisher score (18) evaluated at the true value θ_0 of θ has mean zero [11]:

$$E[S(\theta_0)] = 0. \tag{19}$$

Furthermore, by the multivariate central limit theorem, we can derive [14] that $S(\theta_0)$ is asymptotically normally distributed:

$$S(\theta_0) \sim As\mathcal{N}(0, F(\theta_0)), \tag{20}$$

with $F(\theta_0)$ the so-called Fisher information matrix i.e.:

$$\begin{aligned} F(\theta_0) &= E\left[S(\theta_0)S^T(\theta_0)\right] \\ &= \frac{N^2}{4\sigma^4}E\left[\left(\left(\frac{\partial V_N(\theta)}{\partial\theta}\right)\left(\frac{\partial V_N(\theta)}{\partial\theta}\right)^T\right)\bigg|_{\theta=\theta_0}\right] \end{aligned} \tag{21}$$

The distribution (20) is equivalent to:

$$S^T(\theta_0)F^{-1}(\theta_0)S(\theta_0) \sim As\chi_k^2 \tag{22}$$

which leads to the following (asymptotically valid) $100\alpha\%$ confidence region for θ_0:

$$\left\{\theta \mid S^T(\theta)F^{-1}(\theta)S(\theta) \leq \chi\right\}, \tag{23}$$

with χ such that $Pr(\chi_k^2 < \chi) = \alpha$.

Now, let us particularize this confidence region to the case of a BJ true system and model structure i.e. a model structure where $G(z,\theta)$ and $H(z,\theta)$ are two arbitrary and independently parametrized transfer functions. Applying the definitions, we obtain the following expressions for the Fisher score and the Fisher information matrix:

$$S(\theta_0) = \frac{1}{\sigma_e^2}\sum_{t=1}^{N}\psi(t,\theta_0)\epsilon(t,\theta_0) \tag{24}$$

$$F(\theta_0) = \frac{1}{\sigma_e^2}\sum_{t=1}^{N}E\psi(t,\theta_0)\psi^T(t,\theta_0) \tag{25}$$

where the latter expression is obtained from (21) if we use the fact that $\psi(t,\theta_0)$ is uncorrelated with the white noise $\epsilon(s,\theta_0)$ for $s > t$ (cfr. [10]). It is important to note that $\psi(t,\theta)$ evaluated at θ_0 is equal to:

$$\psi(t,\theta_0) = \underbrace{\frac{\Lambda_G(z,\theta_0)}{H(z,\theta_0)}u(t)}_{s_u(t,\theta_0)} + \underbrace{\frac{\Lambda_H(z,\theta_0)}{H(z,\theta_0)}e(t)}_{s_e(t,\theta_0)}$$

with $\Lambda_G(z,\theta) = \frac{\partial G(z,\theta)}{\partial\theta}$ and $\Lambda_H(z,\theta) = \frac{\partial H(z,\theta)}{\partial\theta}$. Using this expression, $F(\theta_0)$ can be rewritten as follows:

	$\gamma_{LR,0.95}$	$\gamma_{FS,0.95}$
$N = 50$, $\sigma_u^2 = 1$	0.947	0.962
$N = 100$, $\sigma_u^2 = 1$	0.949	0.952
$N = 200$, $\sigma_u^2 = 1$	0.965	0.955
$N = 50$, $\sigma_u^2 = 100$	0.942	0.952
$N = 100$, $\sigma_u^2 = 100$	0.950	0.943
$N = 200$, $\sigma_u^2 = 100$	0.963	0.945

TABLE I

OBSERVED COVERAGE RATE $\gamma_{0.95}$ OF THE CONFIDENCE REGIONS U_{LR} AND U_{FS} FOR DIFFERENT VALUES OF N AND σ_u^2. THE DESIRED COVERAGE RATE α IS 0.95 AND $K = 5000$

$$F(\theta_0) = \left(\frac{1}{\sigma_e^2} \sum_{t=1}^{N} s_u(t,\theta_0) s_u^T(t,\theta_0) \right) + N \left\| \frac{\Lambda_H(z,\theta_0)}{H(z,\theta_0)} \right\|_2^2 \tag{26}$$

where $\|.\|_2$ represents the H_2 norm. Consequently, the confidence region (23) becomes:

$$\left\{ \theta \mid \Gamma(\theta, \sigma_e^2) \leq \chi \right\}$$

$$\Gamma(\theta, \sigma_e^2) = \frac{1}{\sigma_e^2} \left(\sum_{t=1}^{t=N} \psi(t,\theta)\epsilon(t,\theta) \right) R^{-1}(\theta, \sigma_e^2) \left(\sum_{t=1}^{t=N} \psi(t,\theta)\epsilon(t,\theta) \right)$$

$$R(\theta, \sigma_e^2) = \left(\sum_{t=1}^{t=N} s_u(t,\theta) s_u^T(t,\theta) \right) + \sigma_e^2 N \left\| \frac{\Lambda_H(z,\theta)}{H(z,\theta)} \right\|_2^2$$

with χ as defined above. In the likely event that σ_e^2 is unknown, σ_e^2 will be replaced by its estimate $\hat{\sigma}_e^2$ and a F-distribution will be used. The (asymptotically valid) $100\alpha\%$ confidence region for θ_0 becomes:

$$U_{FS} = \left\{ \theta \mid \Gamma(\theta, \hat{\sigma}_e^2) \leq \beta \right\} \tag{27}$$

with β as defined in (6). Note that, like U_{LR}, the confidence region U_{FS} is no longer an ellipsoid and that its construction will be also computationally expensive, requiring the evaluation of $\Gamma(\theta, \hat{\sigma}_e^2)$ at a sufficient number of points to produce contours.

C. Simulation results

Both confidence regions U_{LR} and U_{FS} are constructed based on an asymptotic property. We will nevertheless show in this subsection that the reliability of these confidence regions is very high. To show this, we consider the same Box-Jenkins true system as in Section III-C and we perform $K = 5000$ Monte Carlo simulations with a white noise input with given variance σ_u^2 to estimate the actual coverage rate of the confidence regions U_{LR} and U_{FS} for a desired coverage rate of 95%. Recall that the observed coverage $\gamma_{0.95}$ is defined as the percentage of the total number of data sets K, for which the true parameter values lay within the confidence regions.

In Table I, we represent the observed coverage rate $\gamma_{0.95}$ for the two confidence regions for different values of N and σ_u^2. We see that the observed coverage rates of U_{LR} and U_{FS} is, even for small N and σ_u^2, very close to the desired coverage rate $\alpha = 0.95$; especially when we compare these results with those in Figure 2 that correspond to the classical confidence region U. The SNR does not seem to influence the results for U_{LR} and U_{FS}.

V. CONCLUSIONS

It is well known that the variance of the identified parameter vector $\hat{\theta}_N$ can be reduced either by increasing the number N of data or by increasing the signal-to-noise ratio (SNR). In this paper, we have shown that these two variables (i.e. N and the SNR) are also interchangeable for increasing the reliability of the classical uncertainty bounding scheme (i.e. for decreasing the difference between the actual and desired coverage rates of the classical confidence region U). Another contribution of this paper is to present alternatives for the classical approach to constructing probabilistic confidence regions for Box-Jenkins systems.

REFERENCES

[1] A. Azzalini. *Statistical Inference - Based on the likelihood*. Chapman and Hall, London, 1996.

[2] M. Barenthin and H. Hjalmarsson. Joint input design and H_∞ state feedback with ellipsoidal parametric uncertainties via LMIs. *Automatica*, 44(2):543–551, 2008.

[3] X. Bombois, M. Gevers, G. Scorletti, and B.D.O. Anderson. Robustness analysis tools for an uncertainty set obtained by prediction error identification. *Automatica*, 37(10):1629–1636, 2001.

[4] X. Bombois, G. Scorletti, M. Gevers, P.M.J. Van den Hof, and R. Hildebrand. Least costly identification experiment for control. *Automatica*, 42(10):1651–1662, 2006.

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

[6] 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 *Proceedings of the European Control Conference*, Kos, Greece, July 2007.

[7] A.J. den Dekker, X. Bombois, and P.M.J. Van den Hof. Finite sample confidence regions for parameters in prediction error identification using output error models. In *Proceedings of the 17th IFAC World Congress*, Seoul, South Korea, July 2008.

[8] S. Douma and P.M.J. Van den Hof. Probabilistic uncertainty modelling in output error models with unmodelled dynamics. In *Proceedings of the American Control Conference*, Minneapolis, MA, June 2006.

[9] H. Jansson and H. Hjalmarsson. Input design via LMIs admitting frequency-wise model specifications in confidence regions. *IEEE Transactions on Automatic Control*, 50(10):1534–1549, October 2005.

[10] L. Ljung. *System Identification: Theory for the User*. Prentice-Hall, Englewood Cliffs, NJ, 2nd edition edition, 1999.

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

[12] A. van den Bos. *Parameter Estimation for Scientists and Engineers*. John Wiley and Sons, Inc, Hoboken, New Jersey, 2007.

[13] P.M.J. Van den Hof, S.G. Douma, A.J. den Dekker, and X. Bombois. Probabilistic model uncertainty bounding in prediction error identification based on alternative test statistics. Submitted to Automatica.

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