Delft University of Technology
Delft Center for Systems and Control

Technical report 09-019

Probabilistic model uncertainty bounding in prediction error identification based on alternative test statistics

P.M.J. Van den Hof, S.G. Douma, A.J. den Dekker and X. Bombois

8 January 2009

Delft Center for Systems and Control
Delft University of Technology
Mekelweg 2, 2628 CD Delft
The Netherlands
Phone: +31-15-27 8xxxx
Fax: +31-15-27 86679
http://www.dcsc.tudelft.nl

Probabilistic model uncertainty bounding in prediction error identification based on alternative test statistics [⋆]

Paul M.J. Van den Hof [2], Sippe G. Douma [1], Arnold J. den Dekker and Xavier Bombois

Delft Center for Systems and Control, Delft University of Technology,
Mekelweg 2, 2628 CD Delft, The Netherlands

Abstract

In prediction error identification probabilistic model uncertainty bounds are generally derived from the statistical properties of the parameter estimator. The probabilistic bounds are then based on an (asymptotic) normal distribution of the parameter estimator, accompanied by a covariance matrix, which generally has to be estimated from data too. When the primal interest of the identification is in quantifying the parameter uncertainty on the basis of one single experiment, alternative methods exist that do not require the specification of the full pdf of the parameter estimator. By using the freedom to choose alternative test statistics as a basis for the uncertainty bounding, alternative bounds can be derived that are computationally more attractive and that are less dependent on (asymptotic) assumptions. The alternative method applies to both linearly (ARX) and nonlinearly parametrized model structures. It is powerful in handling the nonasymptotic finite data case, showing that for Output Error (OE) models and for Instrumental Variable (IV) estimates there exist exact uncertainty bound for finite data.

Key words: system identification, model uncertainties, uncertainty bounding, statistical tools, prediction error, hypothesis testing.

1 Introduction

Dynamical models that are identified on the basis of measurement data are usually accompanied by an indication of their reliability. The variance of estimated parameters or the variance of estimated frequency responses is generally used as an indication of this reliability (or precision); it is commonly constructed on the basis of prior information on the data generating system and on 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.

There are several identification paradigms in which model uncertainty sets can be identified on the basis of measurement data. The areas of set membership identification [19] and $\mathcal{H}_\infty$ identification [4,20] have been particularly devoted to this problem, aiming at the construction of hard-bounded errors on estimated nominal models. While hard-bounded model uncertainty sets have the advantage that they allow hard guarantees on robustness properties of designed controllers, they can suffer from substantial conservatism when noise disturbances that affect the measurement data are of a random-type. This is extensively discussed in e.g., [21].

In the probabilistic (prediction error) approach to system identification, model uncertainty quantification is based on covariance information on estimated parameters, in conjunction with a presumed (or asymptotically achieved if the number of data tends to infinity) Gaussian probability density function, see e.g., [18,24]. This description leads to probabilistic confidence bounds on estimated parameters, from which also probabilistic confidence bounds on related model properties can be derived, such as frequency responses and poles/zeros, with any prechosen level of probability. In this standard context the uncertainty bounds are valid asymptoti-

[⋆] Earlier versions of the results of this paper were presented at the 14th IFAC Symp. System Identification, 27-29 March 2006, Newcastle, Australia, and 17th IFAC World Congress, 6-11 July 2008, Seoul, Korea.

[1] Now with Shell International Exploration and Production, Rijswijk, The Netherlands.

[2] Corresponding author Tel. +31 15 2784509; Fax +31 15 2786679; Email: p.m.j.vandenhof@tudelft.nl

cally (in the number of data) while bias effects are neglected.

In classical prediction error identification, explicit and exact expressions for the parameter covariance matrix are available for model structures that are linear-in-the-parameters in the situation that the model structures are correct, i.e. the data generating system is part of the model set, $\mathcal{S} \in \mathcal{M}$. For linear regression models with deterministic regressors (such as FIR and generalized FIR [25,14]) this holds for finite data length, for ARX models this holds asymptotically. For general model structures, and under the assumption $\mathcal{S} \in \mathcal{M}$, approximate expressions for the asymptotic parameter covariance matrix can be obtained by using first order Taylor expansions. However, in this situation exact system knowledge is also required to compute these approximate expressions for the covariance matrix.
Only in case of linear parametrizations, results are available for model uncertainty bounding when the model structures are not correct ($\mathcal{S} \notin \mathcal{M}$), see e.g. [13,17] and [14] Chapter 7.

If we restrict to variance-induced parameter uncertainty (and neglect bias terms), then for Gaussian distributions parameter confidence bounds that are constructed on the basis of the (exact) covariance matrix of the parameter estimator, lead to the smallest possible parameter uncertainty regions for a given probability level. However, usually the exact covariance matrix is not available, and a replacement has to be made with an estimated covariance.

In this paper it will be shown that utilizing the statistical properties of the *estimator* is not necessarily the only way to arrive at uncertainty bounds for estimated parameters. Some alternatives are studied, where the aim is to specify parameter uncertainty regions that do not (or at least not as much) rely on asymptotic assumptions but for which exact probabilistic expressions can be made. It will be shown that the quantification of parameter uncertainty on the basis of only *one experiment* can be done without the full analysis of the parameter estimator, by using alternative test statistics that underly the uncertainty bounding procedure. This will be shown to facilitate uncertainty bounding in several ways, as well as give rise to results that show potentials for application in finite-time analysis. Finite-time analysis of estimated parameters is an important problem, however with few results so far. For some results see e.g. [2,27] and the plenary paper [3].

Throughout this paper it will be assumed that in an open-loop experimental situation input signals are fully under control of the experimenter and therefore are considered to be deterministic. This implies that in those circumstances any uncertainty in the estimated parameters originates from random disturbances on the output signals. This situation is particularly suited for a posteriori quantifying uncertainty in estimated parameters, rather than a priori quantifying uncertainty bounds for estimators, prior to doing the experiment. In a closed-loop experimental set-up, the input signals become affected by the output disturbances and therefore become random also.

After presenting the principle concepts of constructing parameter uncertainty regions on the basis of estimator statistics, alternative approaches are illustrated in a simple example, and next generalized to ARX models. Subsequently it is shown how the approach can be extended to model structures that are not linear-in-the-parameters (Output Error). Attention will be restricted to the situation that there are variance errors only ($\mathcal{S} \in \mathcal{M}$). Extension of the results to the situation of including bias errors also ($\mathcal{S} \notin \mathcal{M}$) is available in [7]. Preliminary material underlying the analysis was presented in the conference contributions [8–10,5,6].

2 Prediction error identification setting

We will consider prediction error (PE) models parametrized by a parameter vector θ, corresponding to a plant model $G(q, \theta)$ and a noise model $H(q, \theta)$, with q the standard shift operator. In a standard prediction error framework [18,24] a model is identified from measurement data $Z^N := \{y, u\}_N$ of data length N according to

$$\hat{\boldsymbol{\theta}}_N = \arg\min_{\theta} V_N(\theta, Z^N) \tag{1}$$

with $V_N(\theta, Z^N) = \frac{1}{N}\sum_{t=1}^{N} \varepsilon^2(t, \theta)$, where the residuals $\varepsilon(t, \theta)$ are constructed as

$$\varepsilon\left(t, \theta\right) = H^{-1}(q, \theta)\left[y(t) - G\left(q, \theta\right) u(t)\right], \tag{2}$$

and with y, u respectively the output and input signal of the plant. The measurement data is assumed to be generated according to the system $\mathcal{S}$:

$$y(t) = G_0(q)u(t) + v(t) \tag{3}$$

where $G_0(q)$ a linear time-invariant dynamical system, $u(t)$ is a measured input sequence, and $v(t)$ denotes an additional unknown contribution to $y(t)$. It is assumed that $v(t) = H_0(q)e(t)$ with H_0 a linear time-invariant monic stable and stably invertible filter, and e a stationary stochastic zero-mean white noise process with variance σ_e^2. Furthermore, we denote by $\mathcal{M}$ the set of models $\{G(q, \theta), H(q, \theta)\}_{\theta\in\Theta}$ with Θ representing the particular range of parameters determining the model sets, typically $\Theta \subset \mathbb{R}^d$. Throughout this paper it is assumed that $\mathcal{S} \in \mathcal{M}$, implying that there exists a parameter θ_0 such that $G(q, \theta_0) = G_0(q)$ and $H(q, \theta_0) = H_0(q)$. Boldface variables ($\hat{\boldsymbol{\theta}}_N$) are used to indicate random variables, in order to distinguish them from single realizations thereof ($\hat{\theta}_N$).

Under the given conditions and the standard regularity conditions in PE identification, as well as under persistency of excitation conditions on the input signal, the parameter estimator $\hat{\boldsymbol{\theta}}_N$ satisfies an asymptotic Gaussian distribution:

$$\sqrt{N}(\hat{\boldsymbol{\theta}}_N - \theta_0) \underset{N\to\infty}{\to} \mathcal{N}(0, P_0) \tag{4}$$

with P_0 the covariance matrix of the asymptotic distribution. This asymptotic result can be rewritten in a quadratic form, leading to a χ^2 distribution:

$$N(\hat{\boldsymbol{\theta}}_N - \theta_0)^T P_0^{-1}(\hat{\boldsymbol{\theta}}_N - \theta_0) \underset{N\to\infty}{\to} \chi^2_d \tag{5}$$

with d the number of degrees of freedom in the χ^2 distribution, being equal to the dimension of the parameter vector.

Remark 1 *In the sequel of the paper we will make the assumption that the noise process e has a Gaussian distribution with zero mean and variance σ_e^2. Most of the asymptotic results however will not be dependent on this assumption. Whenever this is the case it will be mentioned explicitly.*

3 Probabilistic uncertainty bounds and hypothesis testing

3.1 Uncertainty bounds in prediction error identification

The intention of a probabilistic uncertainty bound is to formulate a bounded set of parameter values to which the real underlying parameter θ_0 belongs with a predefined level of probability. In prediction error identification, it is standard to derive such an uncertainty bound on the basis of (5), with a reasoning as follows:
Equation (5) implies that the random variable $\hat{\boldsymbol{\theta}}_N$ satisfies:

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

with $\mathcal{D}(\alpha, \theta_0) := \left\{\theta \mid N(\theta - \theta_0)^T P_0^{-1}(\theta - \theta_0) \le \chi^2_{d,\alpha}\right\}$ and $\chi^2_{d,\alpha}$ corresponds to a probability level α in the χ^2_d-distribution, i.e. it is the α quantile of the χ^2 distribution with d degrees of freedom. However in order to quantify the uncertainty we are interested in making a probabilistic expression on θ_0 rather than on $\hat{\boldsymbol{\theta}}_N$. This is being achieved by basically reverting the expression, realizing that for every realization $\hat{\theta}_N$ of $\hat{\boldsymbol{\theta}}_N$ it holds that

$$\hat{\theta}_N \in \mathcal{D}(\alpha, \theta_0) \Leftrightarrow \theta_0 \in \mathcal{D}(\alpha, \hat{\theta}_N).$$

As a result

$$\theta_0 \in \mathcal{D}(\alpha, \hat{\theta}_N) \ \text{ with probability } \alpha, \tag{6}$$

with $\mathcal{D}(\alpha, \hat{\theta}_N) := \left\{\theta \mid N(\hat{\theta}_N - \theta)^T P_0^{-1}(\hat{\theta}_N - \theta) \le \chi^2_{d,\alpha}\right\}$.
If a parameter estimate $\hat{\theta}_N$ is obtained from one single experiment, a $100 \times \alpha\%$ confidence interval for θ_0 is specified by $\mathcal{D}(\alpha, \hat{\theta}_N)$, implying that asymptotic in N, in $100 \times \alpha\%$ of the realizations $\hat{\theta}_N$ of the random variable $\hat{\boldsymbol{\theta}}_N$ the expression $\theta_0 \in \mathcal{D}(\alpha, \hat{\theta}_N)$ will hold true.

3.2 Relation with hypothesis testing

The step made above, moving from a probabilistic expression on $\hat{\boldsymbol{\theta}}_N$ to a confidence interval on θ_0 can also be phrased in terms of a hypothesis test. We basically test the null hypothesis $H_0 : \theta = \theta_0$ against the alternative hypothesis $H_1 : \theta \neq \theta_0$, on the basis of the particularly chosen test statistic

$$N(\hat{\boldsymbol{\theta}}_N - \theta)^T P_0^{-1}(\hat{\boldsymbol{\theta}}_N - \theta). \tag{7}$$

Under the null hypothesis ($\theta = \theta_0$) the test statistic is known to have an asymptotic χ^2_d-distribution, according to (5). This allows one to compose tests (i.e. to 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. When the test statistic has been selected, and a significance level α has been chosen, a $100 \times \alpha\%$ confidence region for θ_0 is then constituted by the set of all values θ for which the null hypothesis $\theta = \theta_0$ would be accepted [1]. This leads to the same expression (6) for θ_0.

Note though that the choice of test statistic (7) is a freedom that is available to the user. In the situation above, the test statistic is chosen directly related to the probability density function (pdf) of the estimator (4). [3] Alternative choices may lead to different confidence regions. This freedom will be explored in the sequel of this paper. Additionally it has to be noted that in many situations P_0 might be unknown, being dependent on the system $\mathcal{S}$. In those situations P_0 needs to be estimated in order to calculate (approximate) confidence bounds (6).

4 Illustrative Example

In order to illustrate the effect that can be obtained by choosing alternative test statistics, we consider the following - very simple - example. Consider the data generating system $\mathbf{y} = \theta_0 \boldsymbol{x}_1 + \boldsymbol{x}_2$, and one available measurement $\{y, x_1\}$ of $\mathbf{y}$ and $\boldsymbol{x}_1$. It is given that $\boldsymbol{x}_1, \boldsymbol{x}_2$ are random numbers that are Gaussian distributed, with an unknown correlation, and with $\boldsymbol{x}_2 \in \mathcal{N}(0, 2)$. We consider the following estimator of θ_0:

$$\hat{\boldsymbol{\theta}} = \frac{\mathbf{y}}{\boldsymbol{x}_1}. \tag{8}$$

Under the above conditions the estimator (8) satisfies

$$\hat{\boldsymbol{\theta}} = \frac{\mathbf{y}}{\boldsymbol{x}_1} = \theta_0 + \frac{\boldsymbol{x}_2}{\boldsymbol{x}_1}. \tag{9}$$

[3] Note that for Gaussian distributions a symmetric (ellipsoidal) norm-bounded α-probability region corresponds to the smallest possible region satisfying a probability of α. For other distributions the smallest region corresponds to the contours of level sets of the probability density function.

Since $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are correlated, the probability density function of this estimator will generally not be Gaussian [4]. Therefore, evaluation of parameter uncertainty regions on the basis of a test statistic that is directly derived from the pdf $f_{\hat{\boldsymbol{\theta}}}$ will generally be cumbersome.

However since $\boldsymbol{x}_1(\hat{\boldsymbol{\theta}} - \theta_0) = \boldsymbol{x}_2$, it follows that this random variable has a Gaussian distribution

$$\boldsymbol{x}_1(\hat{\boldsymbol{\theta}} - \theta_0) \in \mathcal{N}(0, 2)$$

and consequently $(\hat{\boldsymbol{\theta}} - \theta_0)\frac{\boldsymbol{x}_1^2}{2}(\hat{\boldsymbol{\theta}} - \theta_0) \in \chi_1^2$.

We can now construct an uncertainty region for θ_0 on the basis of a hypothesis test as described in the previous section, and the alternative test statistic

$$(\hat{\boldsymbol{\theta}} - \theta)\frac{\boldsymbol{x}_1^2}{2}(\hat{\boldsymbol{\theta}} - \theta)$$

which under hypothesis $\theta = \theta_0$ is known to satisfy a χ_1^2-distribution. We therefore select all values of θ that, together with the observed values $\hat{\theta}$ and x_1 are within the α quantile of the χ_1^2 distribution of $\boldsymbol{x}_2^2$. This set is exactly given by

$$\mathcal{D}(\alpha, \hat{\theta}) = \left\{ \theta \mid \frac{x_1^2}{2}(\hat{\theta} - \theta)^2 \leq \chi_{1,\alpha}^2 \right\}. \qquad (10)$$

As a result it holds that $\theta_0 \in \mathcal{D}(\alpha, \hat{\theta})$ w.p. α.

The interpretation of this probabilistic expression is that when we construct the uncertainty region $\mathcal{D}(\alpha, \hat{\theta})$ for n experiments, i.e. n realizations of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, the constructed region (10) will contain the true parameter only a number of αn times if $n \rightarrow \infty$. Note that $\mathcal{D}(\alpha, \hat{\theta})$ is dependent on the particular experiment that is done, i.e. the uncertainty set is dependent on x_1.

The result of the example is illustrated in Figure 1, where the (unknown) pdf $f_{\hat{\boldsymbol{\theta}}}$ of $\hat{\boldsymbol{\theta}}$ is sketched together with the smallest 90% uncertainty regions for $\hat{\boldsymbol{\theta}}$, as well as the 90% uncertainty region that is symmetric around $\mathbb{E}\hat{\boldsymbol{\theta}}$. The Figure also shows the parameter interval that relates to all realizations $\hat{\theta}$ of $\hat{\boldsymbol{\theta}}$ for which correctly holds that $\theta_0 \in \mathcal{D}(\alpha, \hat{\theta})$.

In the classical approach the parameter uncertainty region is determined on the basis of $\mathbb{E}\hat{\boldsymbol{\theta}}$ and $cov(\hat{\boldsymbol{\theta}})$. For a Gaussian distribution one then arrives at the smallest possible parameter uncertainty regions corresponding to a fixed probability level. However the above quantities need to be known. The alternative analysis does not require full analysis of the pdf of the parameter estimator, at the possible cost of delivering larger parameter uncertainty sets, but with exact probabilistic expressions connected to it. Next it will be shown how the alternative paradigm can be applied to the quantification of uncertainty in identified ARX / linear regression models.

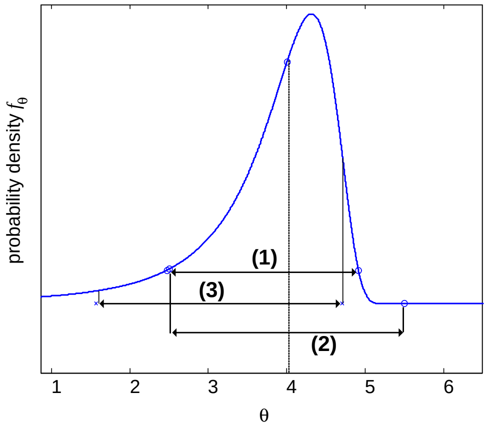

Fig. 1. Probability density function $f_{\hat{\boldsymbol{\theta}}}$ of $\hat{\boldsymbol{\theta}}$ (8) and three uncertainty regions each corresponding to a confidence level $\alpha = 0.9$. The smallest (1) and symmetric (2) 90% regions are tied to the pdf of $\hat{\boldsymbol{\theta}}$. The computed 90% region (3) is determined by all θ for which $\theta_0 \in \mathcal{D}(\alpha, \theta)$; it is the collection of all realizations $\hat{\theta}_N$ on the basis of which a correct uncertainty bound for θ_0 will be specified; it is based on a symmetric 90% probability region of random variable $\boldsymbol{x}_2$.

5 Uncertainty bounding in ARX models

5.1 *Identification setting and standard approach*

The ARX model set is determined by

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

with n_k the delay and

$$A(q^{-1}, \theta) = 1 + a_1 q^{-1} + \cdots + a_{n_a} q^{-n_a}$$
$$B(q^{-1}, \theta) = b_0 + b_1 q^{-1} + \cdots + b_{n_b - 1} q^{-n_b + 1}$$

with $\theta^T = [a_1 \cdots a_{n_a} \; b_0 \cdots b_{n_b - 1}]$, having a dimension $d = n_a + n_b$. In prediction error identification, the one step ahead predictor is considered, written as $\hat{y}(t|t-1;\theta) = \varphi^T(t)\theta$ with $\varphi^T(t) = [-y(t-1) \cdots - y(t-n_a) \; u(t) \cdots u(t-n_b+1)]$, having dimension d, and y, u the scalar-valued output and input signal. The parameter estimate is obtained by minimizing the quadratic prediction error criterion (1) with $\varepsilon(t, \theta) = y(t) - \hat{y}(t|t-1;\theta)$. By denoting

$$\boldsymbol{\Phi} = \begin{pmatrix} \varphi^T(1) \\ \vdots \\ \varphi^T(N) \end{pmatrix} \text{ and } \mathbf{y} = [y(1) \cdots y(N)]^T \qquad (11)$$

it follows that $\hat{\boldsymbol{\theta}}_N = (\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T\mathbf{y}$. If the data generating system belongs to the model class ($\mathcal{S} \in \mathcal{M}$) then there exists a θ_0 such that $\mathbf{y} = \boldsymbol{\Phi}\theta_0 + \mathbf{e}$ with $\mathbf{e}$ an N-vector of samples

[4] It is plotted in Figure 1 for $\boldsymbol{x}_2 \in \mathcal{N}(0, 2)$ and $\boldsymbol{x}_1 = 3 + \frac{0.5}{\boldsymbol{x}_2}$

from a white noise process, and so

$$\hat{\boldsymbol{\theta}}_N = \theta_0 + (\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T\mathbf{e}. \tag{12}$$

In accordance with the theory as presented in Section 3.1, the parameter estimator has an asymptotic Gaussian distribution

$$\sqrt{N}(\hat{\boldsymbol{\theta}}_N - \theta_0) \underset{N\to\infty}{\to} \mathcal{N}(0, P_0) \tag{13}$$

where for the ARX situation the covariance matrix of the asymptotic distribution is given by [18]:

$$P_0 = (\lim_{N\to\infty} \mathbb{E}[\frac{1}{N}\boldsymbol{\Phi}^T\boldsymbol{\Phi}])^{-1} \cdot \sigma_e^2. \tag{14}$$

When building a parameter uncertainty bound for θ_0 on the basis of a single experiment, by applying the approach as outlined in Section 3, this can be done on the basis of the test statistic

$$N(\hat{\boldsymbol{\theta}}_N - \theta)^T P_0^{-1}(\hat{\boldsymbol{\theta}}_N - \theta) \tag{15}$$

which under hypothesis $\theta = \theta_0$ is known to have an asymptotic χ_d^2 distribution. This leads to the following result:

Result 1 *On the basis of the test statistic (15), it follows that asymptotically in N, $\theta_0 \in \mathcal{D}(\alpha, \hat{\theta}_N)$ w.p. α, with*

$$\mathcal{D}(\alpha, \hat{\theta}_N) := \{\theta \mid N(\hat{\theta}_N - \theta)^T P_0^{-1}(\hat{\theta}_N - \theta) \leq \chi^2_{d,\alpha}\}. \tag{16}$$

This result is built on the asymptotic normality of the term $(\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T\mathbf{e}$.

5.2 Alternative 1 to ARX uncertainty bounding

As an alternative to analyzing the full pdf of $\hat{\boldsymbol{\theta}}_N$, the expression (12) for the parameter estimator can be rewritten in the form

$$\frac{1}{\sqrt{N}}\boldsymbol{\Phi}^T\boldsymbol{\Phi}(\hat{\boldsymbol{\theta}}_N - \theta_0) = \frac{1}{\sqrt{N}}\boldsymbol{\Phi}^T\mathbf{e}, \tag{17}$$

where the factor $1/\sqrt{N}$ is chosen because of the fact that due to the Central Limit Theorem ([18])

$$\frac{1}{\sqrt{N}}\boldsymbol{\Phi}^T\mathbf{e} \underset{N\to\infty}{\to} \mathcal{N}(0, Q_0) \quad \text{with } Q_0 = \lim_{N\to\infty} \mathbb{E}\frac{1}{N}\boldsymbol{\Phi}^T\boldsymbol{\Phi}\cdot\sigma_e^2. \tag{18}$$

This result leads to the suggestion of an alternative test statistic, given by

$$\frac{1}{N}(\hat{\boldsymbol{\theta}}_N - \theta)^T\boldsymbol{\Phi}^T\boldsymbol{\Phi}Q_0^{-1}\boldsymbol{\Phi}^T\boldsymbol{\Phi}(\hat{\boldsymbol{\theta}}_N - \theta) \tag{19}$$

which under hypothesis $\theta = \theta_0$ is known to have an asymptotic χ_d^2 distribution. As a result we can phrase an alternative set for quantifying the parameter uncertainty bound.

Result 2 *On the basis of the test statistic (19), it follows that asymptotically in N, $\theta_0 \in \mathcal{D}(\alpha, \hat{\theta}_N)$ w.p. α, with*

$$\mathcal{D}(\alpha, \hat{\theta}_N) := \{\theta \mid \frac{1}{N}(\hat{\theta}_N - \theta)^T\Phi^T\Phi Q_0^{-1}\Phi^T\Phi(\hat{\theta}_N - \theta) \leq \chi^2_{d,\alpha}\}. \tag{20}$$

This result is built on the asymptotic normality of the term $\frac{1}{\sqrt{N}}\boldsymbol{\Phi}^T\mathbf{e}$.

5.3 Alternative 2 to ARX uncertainty bounding

The approach leading to alternative 1 in the previous subsection can be taken one step further. To this end we introduce the singular value decomposition (svd) of $\boldsymbol{\Phi}^T$, as

$$\boldsymbol{\Phi}^T = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$$

with $\mathbf{U}$ and $\mathbf{V}$ unitary matrices, i.e. $\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = I$, and rewrite the expression (17) for the parameter estimation error into the form:

$$\mathbf{V}^T\boldsymbol{\Phi}(\hat{\boldsymbol{\theta}}_N - \theta_0) = \mathbf{V}^T\mathbf{e}. \tag{21}$$

With the Central Limit Theorem it can be shown that

$$\mathbf{V}^T\mathbf{e} \underset{N\to\infty}{\to} \mathcal{N}(0, \sigma_e^2 I). \tag{22}$$

A proof of (22) is added in the Appendix. This result gives rise to considering the quadratic form of the left hand side of (21):

$$(\hat{\boldsymbol{\theta}}_N - \theta_0)^T\boldsymbol{\Phi}^T\boldsymbol{\Phi}(\hat{\boldsymbol{\theta}}_N - \theta_0)$$

as a basis for the test statistic. Indeed the resulting test statistic

$$\frac{N}{\sigma_e^2}(\hat{\boldsymbol{\theta}}_N - \theta)^T\frac{1}{N}\boldsymbol{\Phi}^T\boldsymbol{\Phi}(\hat{\boldsymbol{\theta}}_N - \theta) \tag{23}$$

consequently has an asymptotic χ_d^2 distribution under hypothesis $\theta = \theta_0$. This leads to the result formulated next.

Result 3 *On the basis of the test statistic (23), it follows that asymptotically in N, $\theta_0 \in \mathcal{D}(\alpha, \hat{\theta}_N)$ w.p. α, with*

$$\mathcal{D}(\alpha, \hat{\theta}_N) := \{\theta \mid \frac{N}{\sigma_e^2}(\hat{\theta}_N - \theta)^T\frac{1}{N}\Phi^T\Phi(\hat{\theta}_N - \theta) \leq \chi^2_{d,\alpha}\}. \tag{24}$$

This result is built on the asymptotic normality of the term $\mathbf{V}^T\mathbf{e}$.

5.4 Intermediate discussion

The three formulations of uncertainty sets vary in terms of their construction, and they vary in terms of the underlying random variable that is supposed to reach an asymptotic normal distribution. When taking a look at the construction of the three uncertainty sets, it appears that in the several sets unknown quantities occur: P_0, Q_0 and σ_e^2. If we focus for the moment on the effect of P_0 and Q_0, and assume that σ_e^2 is known, it appears that either P_0 (Result 1) or Q_0 (Result 2) need to be known. When considering the sample estimates $\hat{P} = (\frac{1}{N}\Phi^T\Phi)^{-1}\sigma_e^2$ and $\hat{Q} = \frac{1}{N}\Phi^T\Phi\sigma_e^2$ and use them as a replacement of P_0 and Q_0 in the expressions for the uncertainty sets, it appears that all three uncertainty sets become the same, and given by (24). Note however, that this approximation of P_0, Q_0 by a sample estimate is a compromise in the scope of Results 1 and 2, but is not required for Result 3. In this respect the analysis leading to Result 3, appears to be the most powerful analysis of the three.

It has been assumed so far that σ_e^2 is known. The present analysis can easily be extended to the situation where this term also has to be estimated from data. This is further addressed in [6].

It has to be noted that in the analysis of the statistical properties of the test statistics, the matrices P_0 and Q_0 are considered to be matrices that are fixed and not parameter-dependent. Note that when writing $P_0 = P_0(\theta_0)$ and $Q_0 = Q_0(\theta_0)$ this could give rise to alternative test statistics where P_0 and Q_0 in uncertainty bounds (16),(20) would need to be replaced by $P(\theta)$ and $Q(\theta)$, thus eliminating the simple ellipsoidal structure of the uncertainty bounds and requiring computationally more expensive algorithms to compute them. This issue will also be further addressed in a subsequent section.

6 Reflection on finite-time perspectives

The three Results presented in the previous section rely on three different random variables to approach a Gaussian distribution:

$$\begin{aligned} &\text{For Result 1:} && (\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\mathbf{e} \\ &\text{For Result 2:} && \tfrac{1}{\sqrt{N}}\mathbf{\Phi}^T\mathbf{e} \\ &\text{For Result 3:} && \mathbf{V}^T\mathbf{e} \end{aligned}$$

As a result, the question whether the results for the uncertainty bounds will also be feasible for finite values of N will highly depend on the question how fast (with increasing N) the several variables approach the (asymptotic) Gaussian distribution. In order to illustrate this we consider a simulation example, particularly focussing on the behaviour of the random variables related to Results 1 and 3.

Example 1 *A first-order data generating system is modelled with an ARX model of the form*

$$\varepsilon(t,\theta) = (1+\theta_a q^{-1})y(t) + \theta_b u(t),$$

such that $\mathcal{S} \in \mathcal{M}$. Experimental data is simulated driving the data-generating system with an input $u(t)$ and noise disturbance $e(t)$ that are independent Gaussian distributed white noise sequences with variance $\sigma_u^2 = \sigma_e^2 = 1$. The system coefficients are $\theta_b = 0.5$, $\theta_a = 0.9$. The parameters θ_b and θ_a are estimated with a least-squares identification criterion. It is verified here whether the random variables, that underly the several test statistics, indeed satisfy a Gaussian distribution for finite values of N.

The top row of Figure 2 depicts the histogram of the element of $(\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\mathbf{e}$ (Result 1) corresponding with $\hat{\theta}_b$, as a function of data length N and for 5000 Monte Carlo simulations. The bottom row depicts the distribution of the corresponding element of $\mathbf{V}^T\mathbf{e}$ (Result 3). The red solid curves indicate closest Gaussian distributions to the results. Clearly, the bottom row is indistinguishable from a Gaussian distribution, while the top approaches a Gaussian slowly. Similar results are presented in Figure 3 for the element of the parameter vector, corresponding with $\hat{\theta}_a$. Clearly, the bottom row is indistinguishable from a Gaussian distribution, while the top approaches a Gaussian distribution very slowly. The corresponding evaluation of the χ_d^2 tests is reflected in Figure 4. It shows that the theoretical asymptotic PE method (Result 1) differs considerably from the theoretical χ_d^2 distribution for small values of N, while the result with the data-based covariance matrix (Result 3) shows a close fit to the χ_d^2 distribution even for very small values of N. The experimental coverage rates of the several tests are collected in Table 1. Note that in these tests exact knowledge is used of P_0 (Result 1), and Q_0 (Result 2), while the test for Result 3 can be performed on the basis of measurement data only.

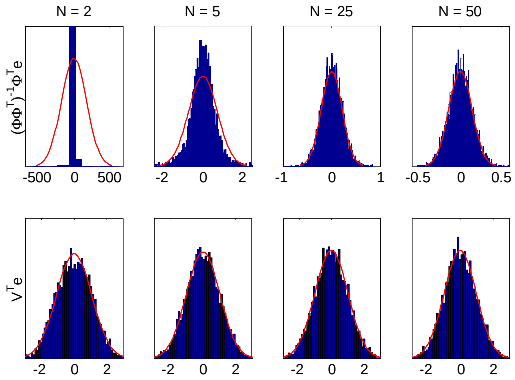

Fig. 2. Distribution of parameters in ARX structure. Top: second element of $(\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\mathbf{e}$ corresponding to $\hat{\theta}_b$ for data length $N = 2, 5, 25, 50$. Bottom: the distribution of the second element of $\mathbf{V}^T\mathbf{e}$ corresponding to $\hat{\theta}_b$. Red solid curves are best fitting Gaussian distributions.

For a further justification and understanding of the finite-sample behaviour of the uncertainty bounds of Result 3 we consider the following Lemma. Its proof is added in the Appendix.

Lemma 1 ([9]) *Consider random vectors $\mathbf{z}, \mathbf{e} \in \mathbb{R}^{N\times 1}$ and*

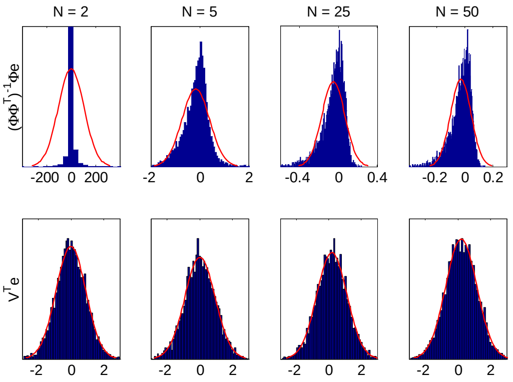

Fig. 3. Similar simulation results as in Figure 2 but then for the first vector elements, i.e. related to the denominator parameter $\hat{\theta}_a$. Red solid curves are best fitting Gaussian distributions.

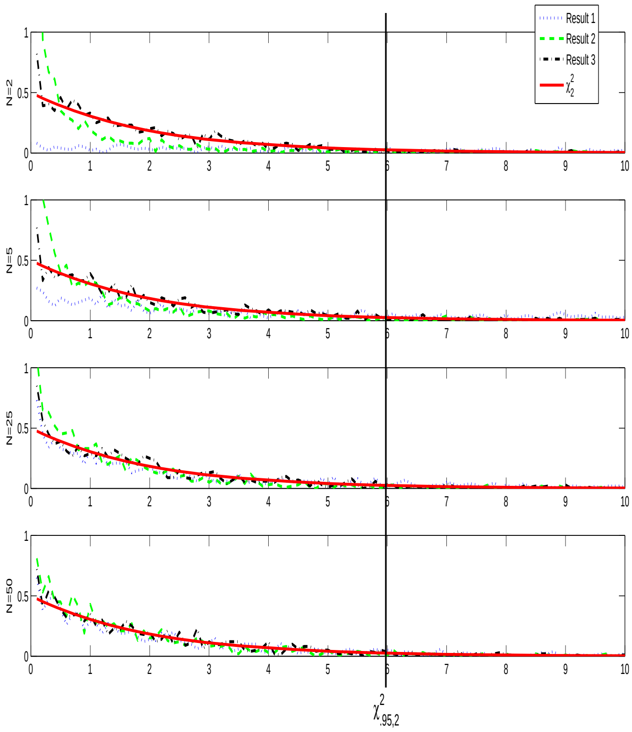

Fig. 4. Monte Carlo evaluation of the χ^2_d tests of Result 1 (blue dotted), Result 2 (green dashed) and Result 3 (black dash-dot), compared with the theoretical χ^2_d test (solid red). Number of data is from top figure to bottom figure: $N = 2, 5, 25, 50$.

a random matrix[5] $\mathbf{V} \in \mathbb{R}^{N\times N}$ *related through* $\mathbf{z} = \mathbf{V}^T\mathbf{e}$. *If the following properties are satisfied:*

(1) $\mathbf{e}$ *has independent identical Gaussian distributed entries,* $\mathcal{N}(0, \sigma^2)$, *and*

	$N = 2$	$N = 5$	$N = 25$	$N = 50$
Result 1	0.1810	0.5530	0.8320	0.8760
Result 2	0.9690	0.9610	0.9550	0.9550
Result 3	0.9610	0.9500	0.9590	0.9550

Table 1
Experimental coverage rates of the χ^2_d tests related to Results (1)-(3) for different values of N; $\alpha = 0.95$.

(2) $\mathbf{e}$ *and* $\mathbf{V}$ *are independent, and*
(3) $\mathbf{V}$ *is unitary, i.e.* $\mathbf{V}^T\mathbf{V} = I$

then the vector elements of $\mathbf{z}$ *are independent identically distributed with Gaussian distribution* $\mathcal{N}(0, \sigma^2)$. □

Note that the result of this Lemma is quite remarkable, in particular because it does not rely on an asymptotic assumption on N. Irrespective of the *pdf* of the elements of matrix $\mathbf{V}$, the resulting random variable $\mathbf{z}$ has a Gaussian distribution. This implies that, would the conditions as formulated in the Lemma hold true, then the uncertainty bound as formulated in Result 3 is exact also for finite values of N. However the condition of statistical independence of $\mathbf{V}$ and $\mathbf{e}$ is formally not met in the situation of ARX models. This is due to the fact that for ARX models $\mathbf{\Phi}$ and $\mathbf{e}$ will have correlation.

However, the results of Example 1 suggest that the existing correlation between $\mathbf{V}$ and $\mathbf{e}$ hardly affects the normality of the test statistic, and therefore the validity of the uncertainty bound for finite N also.

The conclusions to be drawn from sections 5 and 6 are:

- The choice of different test statistics leads to different parameter uncertainty bounds;
- The theoretical (asymptotic) test from PE theory that is used for ARX models requires knowledge of the (unknown) exact covariance matrix P_0;
- Whereas replacing P_0 by a sample estimate is considered to be a compromise in asymptotic PE theory, there exists a test statistic (Result 3) that formally generates the related uncertainty bound;
- This latter data-based uncertainty bound shows improved performance for finite data.

7 A likelihood perspective

In order to put the results presented so far in a perspective, we consider the problem also in a likelihood framework. It will allow us to further specify the relations with existing statistical hypothesis tests, as well as point to further alternatives.
In the considered situation of data generating system as described in Section 2, the joint probability distribution of the observations $y^N = \{y(t)\}_{t=1,\cdots,N}$ (conditioned on the

[5] A random vector/matrix is a vector/matrix with random variables as elements. It does not imply that the several elements are uncorrelated.

given deterministic input sequence u^N) is given by:

$$f_y(y^N;\theta_0) = \prod_{t=1}^{N} f_e(\epsilon(t,\theta_0);\theta_0). \tag{25}$$

Assuming a zero-mean Gaussian distributed noise e, i.e.

$$f_e(\varepsilon(t,\theta_0);\theta_0) = \frac{1}{\sqrt{2\pi\sigma_e^2}} \exp\left[-\frac{1}{2\sigma_e^2}\varepsilon^2(t,\theta_0)\right] \tag{26}$$

and taking the logarithm of the joint pdf delivers

$$\log f_y(y^N;\theta_0) = -\frac{N}{2}\log(2\pi) - N\log\sigma_e - \frac{N}{2\sigma_e^2}V_N(\theta_0). \tag{27}$$

If we substitute the available observations y^N for the corresponding indeterminate variables in (25) and regard the resulting expression as a function of the parameter vector θ for fixed observations y^N, this leads to the likelihood function which now is written as $f_y(\theta;y^N)$. The maximum likelihood estimator (MLE) of θ_0 is given by

$$\hat{\theta}_N = \arg\max_{\theta} f_y(\theta;y^N) = \arg\min_{\theta} V_N(\theta). \tag{28}$$

We need three additional notions to specify relevant test statistics. The Fisher score $S_N(\theta)$ is defined as

$$S_N(\theta) := \frac{\partial \log f_y(y^N;\theta)}{\partial\theta} = \frac{-N}{2\sigma^2}\frac{\partial V_N(\theta)}{\partial\theta}, \tag{29}$$

the Fisher information matrix [12] $J_N(\theta_0)$ as

$$J_N(\theta_0) = -\mathbb{E}\left[\left.\frac{\partial^2 \log f_y(y^N;\theta)}{\partial\theta^2}\right|_{\theta=\theta_0}\right], \tag{30}$$

and the generalized likelihood ratio $L_G(\theta)$ ([16]) as

$$L_N(\theta) = \frac{f_y(\theta;y^N)}{\sup_\theta f_y(\theta;y^N)} = \frac{f_y(\theta;y^N)}{f_y(\hat{\theta}_N;y^N)}. \tag{31}$$

The latter function is bound between 0 and 1.

There are a number of (asymptotic) statistical results that can serve as test statistics for the hypothesis test required for uncertainty bounding procedures. They are formulated in the following Proposition

Proposition 1 *Under appropriate conditions on the data generating system as formulated in Section 2, the following distributions hold asymptotically in N:*

(a) The Maximum likelihood estimator has an asymptotic normal distribution ([12]),

$$\hat{\theta}_N \to \mathcal{N}(\theta_0, J_N^{-1}(\theta_0)) \tag{32}$$

(b) According to Wald ([26]), the covariance matrix of the asymptotic ML estimator, can be replaced by an estimated covariance,

$$\hat{\theta}_N \to \mathcal{N}(\theta_0, J_N^{-1}(\hat{\theta}_N)) \tag{33}$$

(c) The Fisher score has an asymptotic normal distribution ([28]),

$$S_N(\theta_0) \to \mathcal{N}(0, J_N(\theta_0)) \tag{34}$$

(d) The log generalized likelihood has an asymptotic χ_d^2 distribution ([16]),

$$-2\log L_N(\theta_0) \to \chi_d^2. \tag{35}$$

The several notions can be specified for the situation of ARX models. This is formalized in the following Proposition.

Proposition 2 *For the ARX models defined before, the following expressions hold:*

$$(1)\ J_N(\theta_0) = \frac{1}{\sigma_e^2}\mathbb{E}[\mathbf{\Phi}^T\mathbf{\Phi}] \tag{36}$$

$$(2)\ S_N(\theta_0) = \frac{1}{\sigma_e^2}\mathbf{\Phi}^T\mathbf{e} \tag{37}$$

$$(3)\ -2\log L_N(\theta_0) = \frac{N}{\sigma_e^2}\left[V_N(\theta_0) - V_N(\hat{\theta}_N)\right] \tag{38}$$

$$= \frac{N}{\sigma_e^2}(\hat{\boldsymbol{\theta}}_N - \theta_0)^T\frac{1}{N}\mathbf{\Phi}^T\mathbf{\Phi}(\hat{\boldsymbol{\theta}}_N - \theta_0) \tag{39}$$

Proof
(1) This follows directly from the fact that $\frac{\partial^2}{\partial\theta^2}V_N(\theta) = \frac{2}{N}\mathbf{\Phi}^T\mathbf{\Phi}$ and utilizing this in the definition of $J_N(\theta_0)$.
(2) By writing $S_N(\theta_0) = \frac{-N}{2\sigma_e^2}\left.\frac{\partial V_N(\theta)}{\partial\theta}\right|_{\theta_0}$, and substituting $\left.\frac{\partial V_N(\theta)}{\partial\theta}\right|_{\theta_0} = \frac{-2}{N}\sum_{t=1}^{N}\varepsilon(t,\theta_0)\varphi(t) = \frac{-2}{N}\mathbf{\Phi}^T\mathbf{e}$ the result (37) follows.
(3) Expression (38) follows directly from applying the definition of $J_N(\theta)$. The step from (38) to (39) is motivated as follows. Writing $V_N(\hat{\boldsymbol{\theta}}_N, Z^N) = \frac{1}{N}(\mathbf{y}-\mathbf{\Phi}\hat{\boldsymbol{\theta}}_N)^T(\mathbf{y}-\mathbf{\Phi}\hat{\boldsymbol{\theta}}_N)$ and substituting $\mathbf{y} = \mathbf{e} + \mathbf{\Phi}\theta_0$ it follows that $V_N(\hat{\boldsymbol{\theta}}_N) = V(\theta_0) - \frac{2}{N}(\hat{\boldsymbol{\theta}}_N - \theta_0)^T\mathbf{\Phi}^T\mathbf{e} + \frac{1}{N}(\hat{\boldsymbol{\theta}}_N - \theta_0)^T\mathbf{\Phi}^T\mathbf{\Phi}(\hat{\boldsymbol{\theta}}_N - \theta_0)$. Since $\mathbf{\Phi}^T\mathbf{e} = \mathbf{\Phi}^T(\mathbf{y} - \mathbf{\Phi}\theta_0) = \mathbf{\Phi}^T\mathbf{\Phi}(\hat{\boldsymbol{\theta}}_N - \theta_0)$ the result (39) follows. □

The several asymptotic results (a)-(d) formulated in Proposition 1 all give rise to test statistics that can be used for the construction of parameter confidence bounds. Their formulation follows directly from the reasoning around hypothesis testing as discussed in Section 3.

Corollary 1 *The following test statistics all follow an asymptotic-in-N χ_d^2 distribution under the hypothesis*

$\theta = \theta_0$:

$$T_J : (\hat{\boldsymbol{\theta}}_N - \theta)^T J_N^{-1}(\theta)(\hat{\boldsymbol{\theta}}_N - \theta) \quad (40)$$

$$T_W : (\hat{\boldsymbol{\theta}}_N - \theta)^T J_N^{-1}(\hat{\theta}_N)(\hat{\boldsymbol{\theta}}_N - \theta) \quad (41)$$

$$T_R : \frac{1}{\sigma_e^2}(\hat{\boldsymbol{\theta}}_N - \theta)^T \boldsymbol{\Phi}^T \boldsymbol{\Phi} J_N^{-1}(\theta) \boldsymbol{\Phi}^T \boldsymbol{\Phi}(\hat{\boldsymbol{\theta}}_N - \theta) \quad (42)$$

$$T_{LR} : \frac{1}{\sigma_e^2}(\hat{\boldsymbol{\theta}}_N - \theta)^T \boldsymbol{\Phi}^T \boldsymbol{\Phi}(\hat{\boldsymbol{\theta}}_N - \theta) \quad (43)$$

As a result, for each of these test statistics there exist corresponding parameter uncertainty sets $\mathcal{D}(\alpha, \hat{\theta}_N)$ for which holds that $\theta_0 \in \mathcal{D}(\alpha, \hat{\theta}_N)$ with probability α, where

$$\mathcal{D}(\alpha, \hat{\theta}_N) := \{\theta \mid T_* \leq \chi^2_{d,\alpha}\}$$

where '' refers to the particular test statistic T_J, T_W, T_R, or T_{LR} and values of T_* are evaluated according to one of the expressions (40)-(43) substituting the particular realization $\hat{\theta}_N$ for $\hat{\boldsymbol{\theta}}_N$.*

It appears that the uncertainty sets based on T_W and T_{LR} are ellipsoidal regions, due to the fact that the "covariance" matrix in between the terms $(\hat{\boldsymbol{\theta}}_N - \theta)^T$ and $(\hat{\boldsymbol{\theta}}_N - \theta)$ is known and determined by $\hat{\theta}_N$. However the uncertainty sets based on the test statistics T_J and T_R are typically non ellipsoidal and are computationally more expensive, requiring function evaluations at many parameter points θ in order to construct a contour of the uncertainty set.

It is now also fairly easy to formulate the relations with the uncertainty sets that were constructed in Section 5. We formalize these relations in the following observations:

Corollary 2

- *The uncertainty set based on test statistic T_J is the finite-time equivalent of the asymptotic Result 1 in section 5; however whereas in the earlier situation the "covariance" matrix is considered a fixed matrix J_N, in the test statistic T_J it is considered a function of θ_0 and therefore it becomes parametrized as a function of θ in (40).*
- *A similar relation exists between the result for test statistic T_R and the uncertainty set related to Result 2 in section 5.*
- *The uncertainty set based on test statistic T_{LR} is equivalent to the one related to Result 3 in section 5.*

In particular the latter observation, stating that there is an exact likelihood ratio interpretation of Result 3, is stressed here.

Finally we close the discussion on handling ARX models by showing the results of a second simulation example, in which we will illustrate the finite-time properties of the several test statistics.

8 ARX Simulation Experiment

A Monte Carlo simulation experiment is performed to evaluate and compare the methods for computing confidence regions described in the preceding sections. For different data lengths N, a number of $K = 50,000$ 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:

$$y(t)+a_1y(t-1)+a_2y(t-2) = b_0u(t-1)+b_1u(t-2)+e(t), \quad (44)$$

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 $\sigma_e^2 = 0.5$. From each data set, the model was identified using a model set $\mathcal{M}$ with the same ARX structure as the data generating system ($\mathcal{S} \in \mathcal{M}$); then for each data set the estimate $\hat{\theta}_N$ was calculated and it was recorded whether or not the several confidence regions described before contained the true value θ_0. Note that the latter action does not require the construction of the full confidence regions. The observed coverage $\gamma_{0.95}$, for a nominal confidence level $\alpha = 0.95$, is defined as the percentage of the total number of data sets K, for which the true parameter values lay within the confidence region. This means that the asymptotical theory predicts an observed coverage of 95%. Figure 5 shows the observed coverage rates $\gamma_{0.95}$ as a function of the number of data points N. The 95% confidence intervals for $\gamma_{0.95}$ 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 according to T_W and Result 1 are unreliable for small N. Note that the region of Result 1 was constructed on the basis of the theoretical, *asymptotically* valid, expression for the covariance matrix P_0 (14), whose calculation is presented in [5].

On the basis of the presented results it can be concluded that uncertainty bounds on the basis of the likelihood ratio test T_{LR} are equivalent to the earlier derived Result 3, and show the best performance for finite time results. Moreover they are simply to calculate as they are formed by closed-form ellipsoidal regions in the parameter space. They outperform the theoretical uncertainty regions that are used in classical prediction error theory, but they are in fact equal to the pragmatically chosen implementations of the theoretic expressions.

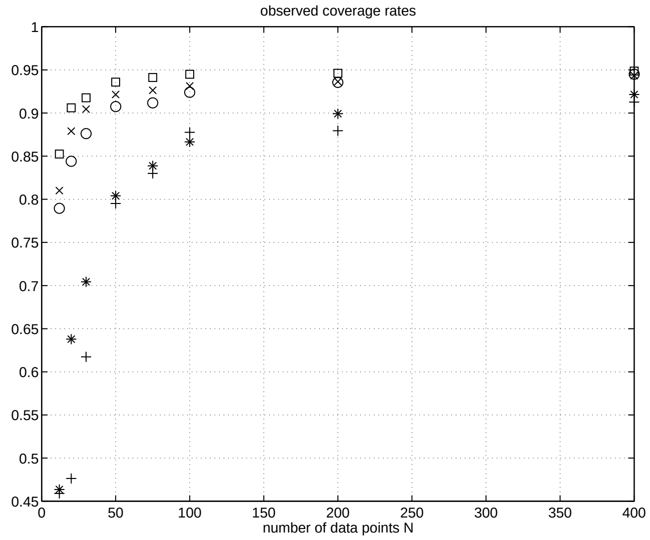

Fig. 5. ARX simulation results for a fixed input input sequence from 50,000 Monte Carlo simulations. Observed coverage rates for parameter uncertainty sets based on the test statistics according to the likelihood ratio test T_{LR} (43) (□), the Rao test T_R (42) (×), the Wald test T_W (41) (○), the test T_J (40) (*) and the uncertainty region according to the classical Result 1 (16) (+). The nominal confidence level is 0.95.

9 Uncertainty bounding in Output Error (OE) models

9.1 Identification setting and standard approach

In the situation of ARX models fruitful use can be made of the fact that there exists a closed form (linear) expression for the parameter estimator (12). The situation becomes more complex if we turn to model structures that are nonlinear in the parameters. An Output Error (OE) model set is determined by

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

with the notation similar to the situation of ARX models, replacing polynomial A by the output error denominator polynomial F, and the dimension of the parameter vector now equals $d = n_b + n_f$. The one-step-ahead predictor becomes

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

For quantifying parameter uncertainty bounds in the classical (PE)- approach, the starting point is to derive a closed-form approximation of $\hat{\boldsymbol{\theta}}_N$ on the basis of a first order Taylor expansion:

$$(\hat{\boldsymbol{\theta}}_N - \theta_0) \approx -[V_N''(\theta_0)]^{-1}[V_N'(\theta_0)] \tag{45}$$

where $V_N'(\theta_0) = \partial V_N(\theta)/\partial\theta|_{\theta=\theta_0}$ and V_N'' is the second derivative of $V_N(\theta)$. Under the common regularity conditions in the PE approach ([18]), $V_N''(\theta)$ in the neighborhood of θ_0 converges uniformly to $\bar{V}''(\theta_0)$, being the second derivative of $\bar{V}'\theta) = \lim_{N\to\infty}\frac{1}{N}\sum_{t=1}^{N}\mathbb{E}\varepsilon(t,\theta)^2$. This implies that $V_N''(\theta_0) \to \bar{V}''(\theta_0)$ with probability 1. The first derivative $V_N'(\theta_0)$ asymptotically reaches a Gaussian distribution:

$$V_N'(\theta_0) \to \mathcal{N}(0,Q), \quad Q = \sigma_e^2 \lim_{N\to\infty}[\frac{1}{N}\Psi(\theta_0)^T\Psi(\theta_0)] \tag{46}$$

with

$$\Psi^T(\theta) = [\psi(1,\theta)\cdots\psi(N,\theta)] \tag{47}$$

and $\psi(t,\theta)$ being the predictor gradient:

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

Additionally $\bar{V}''(\theta_0) = \lim_{N\to\infty}[\frac{1}{N}\Psi^T(\theta_0)\Psi(\theta_0)]$.
Note that due to the fact that the predictor only contains data from a filtered input signal, $\Psi(\theta_0)$ is a deterministic matrix.

Substituting these asymptotic results in (45) leads to the asymptotic distribution

$$\sqrt{N}(\hat{\boldsymbol{\theta}}_N - \theta_0) \to \mathcal{N}(0, P_{oe})$$

with

$$P_{oe} = \sigma_e^2[\lim_{N\to\infty}\frac{1}{N}\Psi(\theta_0)^T\Psi(\theta_0)]^{-1}. \tag{48}$$

A confidence bound for θ_0 is obtained by a hypothesis test on the basis of the test statistic

$$\frac{1}{N}(\hat{\boldsymbol{\theta}}_N - \theta)^T P_{oe}^{-1}(\hat{\boldsymbol{\theta}}_N - \theta) \tag{49}$$

which under the hypothesis $\theta = \theta_0$ is known to have a χ_d^2-distribution. This leads to the following -standard- result [18]:

Result 4 (OE-standard) *On the basis of the test statistic (49), it follows that asymptotically in N, $\theta_0 \in \mathcal{D}(\alpha,\hat{\theta}_N)$ w.p. α, with*

$$\mathcal{D}(\alpha,\hat{\theta}_N) := \{\theta \mid N(\hat{\theta}_N - \theta)^T P_{oe}^{-1}(\hat{\theta}_N - \theta) \leq \chi_{d,\alpha}^2\}. \tag{50}$$

This result is built on the Taylor approximation (45) and the asymptotic normality of the term $V_N''(\theta_0)^{-1}V_N'(\theta_0)$, the latter being equivalent to the asymptotic normality of $[\boldsymbol{\Psi}^T(\theta_0)\boldsymbol{\Psi}(\theta_0)]^{-1}\boldsymbol{\Psi}^T(\theta_0)\mathbf{e}$.

It has to be noted that in the reasoning leading to (49) the covariance matrix P_{oe} is considered to be fixed, and not a function of θ_0. If P_{oe} would be written as $P_{oe}(\theta_0)$ this would give rise to an expression $P_{oe}(\theta)$ in (49) and as a result the ellipsoidal shape of (50) would be lost.

In standard practice the theoretical covariance matrix P_{oe} in 50 which of course is unknown is replaced by a sample estimate

$$\hat{P}_{oe} = \sigma_e^2 \frac{1}{N} \mathbf{\Psi}^T(\hat{\theta}_N)\mathbf{\Psi}(\hat{\theta}_N). \tag{51}$$

Since $\mathbf{\Psi}(\theta_0)$ is a deterministic matrix, the asymptotic normality condition of the above result is no problem. The biggest problem is the fact that P_{oe} depends on the exact system parameter θ_0, which is unknown.

In the sequel we will present several alternatives to this uncertainty set in attempts to obtain results that avoid the Taylor approximation and/or the dependence on θ_0.

9.2 *Alternative 1a - Ellipsoid without Taylor approximation*

We start the analysis of the parameter estimate with the derivative of the identification criterion: $V_N'(\hat{\boldsymbol{\theta}}_N) = 0$ or equivalently

$$\frac{1}{N}\sum_{t=1}^{N}[y(t) - \frac{B(q,\hat{\boldsymbol{\theta}}_N)}{F(q,\hat{\boldsymbol{\theta}}_N)}u(t)] \cdot \psi(t,\hat{\boldsymbol{\theta}}_N) = 0. \tag{52}$$

By defining

$$y_F(t) = F(q,\hat{\boldsymbol{\theta}}_N)^{-1}y(t); \quad u_F(t) = F(q,\hat{\boldsymbol{\theta}}_N)^{-1}u(t) \tag{53}$$

equation (52) can be rewritten as

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

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

$$\hat{\boldsymbol{\theta}}_N = (\mathbf{\Psi}^T\mathbf{\Phi}_F)^{-1}\mathbf{\Psi}^T\mathbf{y}_F \tag{54}$$

with $\mathbf{\Phi}_F^T = \Big[\varphi_F(1,\hat{\boldsymbol{\theta}}_N)\cdots\varphi_F(N,\hat{\boldsymbol{\theta}}_N)\Big]$,

$$\varphi_F^T(t,\hat{\boldsymbol{\theta}}_N) = [-y_F(t{-}1) \cdot\cdot {-}y_F(t{-}n_f)\; u_F(t{-}1) \cdot\cdot u_F(t{-}n_b)]$$

being a vector with dimension $n = n_b + n_f$, and $\mathbf{y}_F = [y_F(1)\cdots y_F(N)]^T$. Here and in the sequel we are using shorthand notation $\mathbf{\Psi} = \mathbf{\Psi}(\hat{\boldsymbol{\theta}}_N)$ and $\mathbf{\Phi}_F = \mathbf{\Phi}_F(\hat{\boldsymbol{\theta}}_N)$.

Note that (54) is an equation that characterizes $\hat{\boldsymbol{\theta}}_N$; however it cannot be used to *calculate* an actual estimate $\hat{\theta}_N$, as the right hand side of the equation is also dependent on $\hat{\boldsymbol{\theta}}_N$. Nevertheless the equation can fruitfully be used to characterize the parameter uncertainty on $\hat{\boldsymbol{\theta}}_N$.
To this end we write the system's relations as:

$$y(t) = \frac{B_0(q)}{F_0(q)}u(t) + e(t), \tag{55}$$

which by filtering through the filter $F_0(q)/F(q,\hat{\boldsymbol{\theta}}_N)$ becomes

$$F_0(q)y_F(t) = B_0(q)u_F(t) + \frac{F_0(q)}{F(q,\hat{\boldsymbol{\theta}}_N)}e(t),$$

that can be written in the regression form:

$$\mathbf{y}_F = \mathbf{\Phi}_F\theta_0 + \mathbf{e}_F, \tag{56}$$

where $\mathbf{e}_F = \frac{F_0(q)}{F(q,\hat{\boldsymbol{\theta}}_N)}[e(1)\ \cdots e(N)]^T$.

Substituting (56) into (54) now delivers:

$$\hat{\boldsymbol{\theta}}_N - \theta_0 = (\mathbf{\Psi}^T\mathbf{\Phi}_F)^{-1}\mathbf{\Psi}^T\mathbf{e}_F$$

which with an svd of $\mathbf{\Psi}^T$: $\mathbf{\Psi}^T = \mathbf{U\Sigma V}^T$ can be written as

$$\mathbf{V}^T\mathbf{\Phi}_F(\hat{\boldsymbol{\theta}}_N - \theta_0) = \mathbf{V}^T\mathbf{e}_F. \tag{57}$$

Similar to the situation of ARX models we can now employ the test statistic

$$\frac{1}{\sigma_e^2}(\hat{\boldsymbol{\theta}}_N - \theta)^T\Phi_F^T V V^T \Phi_F(\hat{\boldsymbol{\theta}}_N - \theta) \tag{58}$$

which under hypothesis $\theta = \theta_0$ will have an asymptotic χ_d^2 distribution, under the conjecture that $\mathbf{V}^T\mathbf{e}_F$ is asymptotically normally distributed.

This leads to the following conjecture:

Result 5 (OE-Alt-1a) *On the basis of the test statistic (58), it follows that asymptotically in N, $\theta_0 \in \mathcal{D}(\alpha,\hat{\theta}_N)$ w.p. α, with*

$$\mathcal{D}(\alpha,\hat{\theta}_N) := \{\theta \mid \frac{1}{\sigma_e^2}(\hat{\theta}_N - \theta)^T\Phi_F^T V V^T\Phi_F(\hat{\theta}_N - \theta) \leq \chi_{d,\alpha}^2\}. \tag{59}$$

This result is built on the presumed asymptotic normality of the term $\mathbf{V}^T\mathbf{e}_F$.

The ellipsoidal uncertainty set (50) does not contain any unknown variables; it is completely determined by the measurement data and the estimated parameter $\hat{\theta}_N$. At the same time it does not rely on a Taylor approximation, and therefore from an analysis point of view it seems to be more powerful.

9.3 *Alternative 1b - Without Taylor approximation*

When substituting (55) into (52) it follows that

$$\frac{1}{N}\sum_{t=1}^{N}[\frac{B_0(q)}{F_0(q)}u(t) - \frac{B(q,\hat{\boldsymbol{\theta}}_N)}{F(q,\hat{\boldsymbol{\theta}}_N)}u(t) + e(t)] \cdot \psi(t,\hat{\boldsymbol{\theta}}_N) = 0, \tag{60}$$

which can be rewritten as

$$\frac{1}{N}\sum_{t=1}^{N}[F(q,\hat{\boldsymbol{\theta}}_N)G_0(q)u_F(t)- \\ B(q,\hat{\boldsymbol{\theta}}_N)u_F(t)+e(t)]\cdot\psi(t,\hat{\boldsymbol{\theta}}_N)=0.$$

By denoting $\varphi_{oe}^T(t,\theta_0)=$

$$[-G_0u_F(t-1)\cdots-G_0u_F(t-n_f)\;u_F(t-1)\cdots u_F(t-n_b)]$$

the equation reduces to

$$\frac{1}{N}\sum_{t=1}^{N}[e(t)-\varphi_{oe}^T(t,\theta_0)(\hat{\boldsymbol{\theta}}_N-\theta_0)]\cdot\psi(t,\hat{\boldsymbol{\theta}}_N)=0.$$

In matrix notation this is formulated as

$$\mathbf{\Psi}^T\mathbf{\Phi}_{oe}(\theta_0)(\hat{\boldsymbol{\theta}}_N-\theta_0)=\mathbf{\Psi}^T\mathbf{e}$$

with $\mathbf{\Phi}_{oe}^T(\theta_0)=[\varphi_{oe}(1,\theta_0)\cdots\varphi_{oe}(N,\theta_0)]$.

When similar to the approach in the previous section, we apply the svd of $\mathbf{\Psi}^T$ it follows that

$$\mathbf{V}^T\mathbf{\Phi}_{oe}(\theta_0)(\hat{\boldsymbol{\theta}}_N-\theta_0)=\mathbf{V}^T\mathbf{e}. \tag{61}$$

The resulting test statistic

$$\frac{1}{\sigma_e^2}(\hat{\boldsymbol{\theta}}_N-\theta)^T\Phi_{oe}(\theta)^TVV^T\Phi_{oe}(\theta)(\hat{\boldsymbol{\theta}}_N-\theta) \tag{62}$$

which under hypothesis $\theta=\theta_0$ is known to have an asymptotic χ_d^2 distribution.

This leads to the following result:

Result 6 (OE-Alt-1b) *On the basis of the test statistic (62), it follows that asymptotically in N, $\theta_0\in\mathcal{D}(\alpha,\hat\theta_N)$ w.p. α, with*

$$\mathcal{D}(\alpha,\hat\theta_N):= \\ \{\theta\mid\frac{1}{\sigma_e^2}(\hat\theta_N-\theta)^T\Phi_{oe}^T(\theta)VV^T\Phi_{oe}(\theta)(\hat\theta_N-\theta)\le\chi_{d,\alpha}^2\}. \tag{63}$$

This result is built on the asymptotic normality of the term $\mathbf{V}^T\mathbf{e}$.

Note that unlike alternative 1a in the previous section the weighting matrix in the quadratic expression (62) is dependent on θ. Therefore the construction of (63) is generally computationally expensive, requiring the evaluation of (62) at a sufficient number of points to produce contours. Unlike all earlier presented uncertainty sets, the confidence region (63) generally is not ellipsoidal.

9.4 Alternative 2 - Ellipsoid with alternative Taylor approximation

In a second alternative, an alternative Taylor approximation is employed. Starting with a first order approximation of $\varepsilon(t,\theta)$ around $\hat\theta_N$, we can write

$$\varepsilon(t,\theta)\approx\varepsilon(t,\hat\theta_N)+\left.\frac{\partial\varepsilon(t,\theta)}{\partial\theta}\right|_{\theta=\hat\theta_N}^T(\theta-\hat\theta_N) \\ =\varepsilon(t,\hat\theta_N)-\psi(t,\hat\theta_N)^T(\theta-\hat\theta_N).$$

Then for $\hat{\boldsymbol{\theta}}_N$ in the neighborhood of θ_0, we can substitute $\theta=\theta_0$ which leads to

$$\varepsilon(t,\hat{\boldsymbol{\theta}}_N)\approx\mathbf{e}(t)-\psi^T(t,\hat{\boldsymbol{\theta}}_N)(\hat{\boldsymbol{\theta}}_N-\theta_0). \tag{64}$$

When substituting this relation in the expression for the cost function gradient:

$$V_N'(\hat{\boldsymbol{\theta}}_N)=\frac{1}{N}\sum_{t=1}^{N}\varepsilon(t,\hat{\boldsymbol{\theta}}_N)\cdot\psi(t,\hat{\boldsymbol{\theta}}_N)=0$$

it follows that

$$\hat{\boldsymbol{\theta}}_N-\theta_0=(\mathbf{\Psi}^T\mathbf{\Psi})^{-1}\mathbf{\Psi}^T\mathbf{e}, \tag{65}$$

which with applying the svd $\mathbf{\Psi}^T=\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ leads to

$$\mathbf{V}^T\mathbf{\Psi}(\hat{\boldsymbol{\theta}}_N-\theta_0)=\mathbf{V}^T\mathbf{e}. \tag{66}$$

Similar to the situation of ARX models and of the previous section, we can employ the test statistic

$$\frac{1}{\sigma_e^2}(\hat{\boldsymbol{\theta}}_N-\theta)^T\Psi^T\Psi(\hat{\boldsymbol{\theta}}_N-\theta) \tag{67}$$

which under hypothesis $\theta=\theta_0$ is known to have an asymptotic χ_d^2 distribution. The following result is now immediate.

Result 7 (OE-Alt-2) *On the basis of the test statistic (67), it follows that asymptotically in N, $\theta_0\in\mathcal{D}(\alpha,\hat\theta_N)$ w.p. α, with*

$$\mathcal{D}(\alpha,\hat\theta_N):=\{\theta\mid\frac{1}{\sigma_e^2}(\hat\theta_N-\theta)^T\Psi^T\Psi(\hat\theta_N-\theta)\le\chi_{d,\alpha}^2\}. \tag{68}$$

This result is built on the Taylor approximation (64) and asymptotic normality of the term $\mathbf{V}^T\mathbf{e}$.

The uncertainty set presented here is also completely determined from data and the estimated parameter $\hat\theta_N$. It does not rely on unknown quantities.

For OE models the matrix $\mathbf{\Psi}$ is composed of filtered input samples (and no output samples). As a result in an open-loop experimental set-up, $\mathbf{V}$ and $\mathbf{e}$ will only be correlated through the fact that $\psi(t, \hat{\boldsymbol{\theta}}_N)$ depends on $\hat{\boldsymbol{\theta}}_N$ which in turn is correlated to the noise in $\mathbf{e}$. This correlation can be considered a secondary effect.

Note that when in the standard uncertainty set (50) the unknown covariance matrix P_{oe} is replaced by the sample estimate (51) then the two uncertainty sets (50) and (68) coincide. Note however that in the latter alternative case we do not have to make the compromise of a sample estimate replacement. Therefore the conclusion is, similar to the ARX case, that the standard method that is commonly used in practice, has a stronger theoretical support than is usually recognized.

Example 2 (Example 1 continued) *A similar simulation experiment as shown in Example 1 is performed but now formulated in an output error model structure. This implies that the data generating system has an additive unit variance output white noise, and a first order model structure is chosen according to*

$$\varepsilon(t, \theta) = y(t) - \frac{\theta_b}{1 + \theta_f q^{-1}} u(t).$$

The parameters θ_f and θ_b are estimated with a least-squares identification criterion.

We illustrate the difference in the distributions of the random variables (65) and (66) for different values of N. Figures 6 and 7 show the corresponding results for the numerator parameter θ_b and the denominator parameter θ_f respectively. The top rows show the histogram of the second element of $(\mathbf{\Psi}^T\mathbf{\Psi})^{-1}\mathbf{\Psi}^T\mathbf{e}$ corresponding with $\hat{\theta}_b$ as a function of data length N and for 5000 Monte Carlo simulations. The bottom rows depict the distribution of the second element of $\mathbf{V}^T\mathbf{e}$, related to Result 6 (OE-Alt-2) in (68). The red solid curves indicate closest Gaussian distributions to the results. Clearly, the bottom rows are nearly indistinguishable from the Gaussian distribution, while the top rows show that a Gaussian distribution is approximated for $N = 50$ (θ_b), or not even achieved for the considered values of N (θ_f). The Output Error results are based on 2000 *Monte Carlo simulations.*

The results of the example show the relevance of the step from (65) to (66) in the analysis leading to Result 6. Similar to the ARX case, the results of the example suggest that the existing correlation between $\mathbf{V}$ and $\mathbf{e}$ hardly affects the normality of the test statistic. This would allow to apply the Gaussian distribution to finite time signals also, even in the case of nonlinearly parametrized model structures as OE.

9.5 Likelihood alternatives for OE models

In Section 7 asymptotic results were presented for test statistics in a likelihood perspective. The statistical results of Proposition 1 are not dependent on any particular model structure, and so they are valid for Output Error models as well. We only need to specify the particular expressions for the variables in the case of Output Error models.

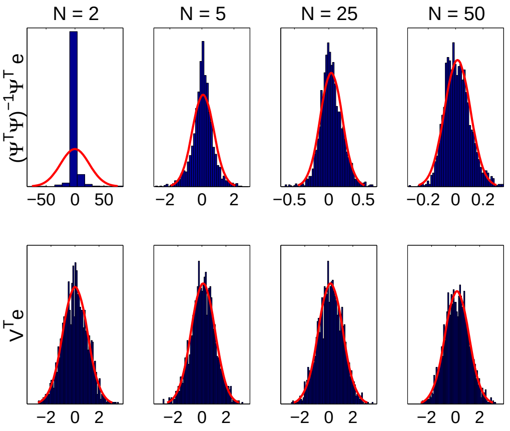

Fig. 6. Distribution of parameters in OE structure. Top: second element of $(\mathbf{\Psi}^T\mathbf{\Psi})^{-1}\mathbf{\Psi}^T\mathbf{e}$ corresponding to $\hat{\theta}_b$ for data length $N = 2, 5, 25, 50$. Bottom: the distribution of the second element of $\mathbf{V}^T\mathbf{e}$ corresponding to $\hat{\theta}_b$. Red solid curves are best fitting Gaussian distributions.

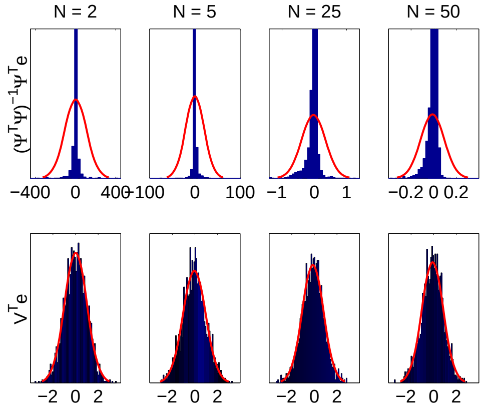

Fig. 7. Similar simulation results as in Figure 6 but then for the test statistic related to the denominator parameter $\hat{\theta}_f$. Red solid curves are best fitting Gaussian distributions.

Proposition 3 *For OE models defined before, the following expressions hold:*

$$(1)\ J_N(\theta_0) = \frac{1}{\sigma_e^2}[\mathbf{\Psi}^T(\theta_0)\mathbf{\Psi}(\theta_0)] \tag{69}$$

$$(2)\ S_N(\theta_0) = \frac{1}{\sigma_e^2}\mathbf{\Psi}^T(\theta_0)\mathbf{e} \tag{70}$$

$$(3)\ -2\log L_N(\theta_0) = \frac{N}{\sigma_e^2}\left[V_N(\theta_0) - V_N(\hat{\theta}_N)\right] \tag{71}$$

Proof
The proof follows along similar lines as the proof of Proposition 2, however now with the ARX predictor gradient $\varphi(t, \theta)$ present in matrix $\mathbf{\Phi}$ being replaced by the OE predictor gradient $\psi(t, \theta)$ in matrix $\mathbf{\Psi}$. □

As a result of the above Proposition, we can again formulate several test statistics that are valid for OE models.

Proposition 4 *The following test statistics all follow an asymptotic-in-N χ^2_d distribution under the hypothesis $\theta = \theta_0$:*

$$T_J : (\hat{\boldsymbol{\theta}}_N - \theta)^T J_N^{-1}(\theta)(\hat{\boldsymbol{\theta}}_N - \theta) \tag{72}$$

$$T_W : (\hat{\boldsymbol{\theta}}_N - \theta)^T J_N^{-1}(\hat{\theta}_N)(\hat{\boldsymbol{\theta}}_N - \theta) \tag{73}$$

$$T_R : \frac{1}{\sigma_e^2}\boldsymbol{\varepsilon}^T(\theta)\boldsymbol{\Psi}(\theta)(\boldsymbol{\Psi}^T(\theta)\boldsymbol{\Psi}(\theta))^{-1}\boldsymbol{\Psi}^T(\theta)\boldsymbol{\varepsilon}(\theta) \tag{74}$$

$$T_{LR} : \frac{N}{\sigma_e^2}[V_N(\theta) - V_N(\hat{\theta}_N)] \tag{75}$$

where $\boldsymbol{\varepsilon}(\theta) = [\varepsilon(1,\theta)\cdots\varepsilon(N,\theta)]^T$.

Proof
The expressions for T_J and T_W are immediate from Proposition 1. The expressions for T_{LR} follows directly from (38). For obtaining T_R we need to write $S(\theta)$ as $S(\theta) = \frac{1}{\sigma_e^2}\boldsymbol{\Psi}(\theta)\boldsymbol{\varepsilon}(\theta)$. Based on the expression $S^T(\theta_0)J_N(\theta_0)^{-1}S(\theta_0) \rightarrow \chi^2_d$, the related test statistic follows. □

Similar to the situation of ARX models, the uncertainty sets related to the different test statistics are formalized as

$$\mathcal{D}(\alpha, \hat{\theta}_N) := \{\theta \mid T_* \leq \chi^2_{d,\alpha}\} \tag{76}$$

where '*' refers to the particular test statistic T_J, T_W, T_R, or T_{LR}.

When evaluating the resulting uncertainty sets we can make the following observations:

- Due to their structure the uncertainty sets based on T_J, T_R and T_{LR} are not ellipsoidal. Only the uncertainty set based on T_W is. Therefore this latter form can be calculated analytically, whereas the other three sets have to be computed by evaluating all possible parameter values in a gridded parameter space. This is of course computationally less attractive.
- The uncertainty set based on T_W is equal to the uncertainty set of Result 6 (OE-Alt-2), and equal to the set of Result 4 (OE-standard) if in this latter set the covariance matrix P_{oe} is replaced by the sample estimate (51). This is the test that is commonly used and that is implemented in Matlab's Identification Toolbox.
- The uncertainty set based on T_R is the only set that does not depend on an estimated parameter $\hat{\theta}_N$.
- The uncertainty set based on T_{LR} can be interpreted as a level set of the cost function V_N. Since V_N is not necessarily convex it is easily understandable that the resulting uncertainty set might contain disconnected regions in the parameter space.

The most important result is formulated in the next Proposition.

Proposition 5 *If* e *is Gaussian distributed then the uncertainty set based on T_R is valid for every value of N.*

Proof Note that with (70) and the fact that $\boldsymbol{\Psi}(\theta_0)$ is deterministic for OE models, the result is immediate. □

This Proposition is actually very strong. There exists a non-asymptotic exact uncertainty bound that is valid for OE models. Its principal disadvantage is, that it is not easily computable, as it is not formalized in a closed form, i.e. it is not ellipsoidal. However for every individual parameter it can simply be verified if it belongs to the set, by simply calculating 74 and verifying whether it satisfies (76). An overview of the several results for Output Error models is provided in Table 2.

10 Simulation Experiment

A simulation experiment has been performed in the same setting as done for the case of ARX models in Section 8. A data generating system has been chosen to be given by

$$G_0(q) = \frac{q^{-1}(b_0 + b_1 q^{-1})}{1 + f_1 q^{-1} + f_2 q^{-2}}, \quad H_0(q) = 1, \tag{77}$$

with $b_0 = 0.1047, b_1 = 0.0872, f_1 = -1.5578$ and $f_2 = 0.5769$, generating data sets of length N. 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 $\sigma^2 = 1$. From each data set, an Output Error model was identified, and and it was recorded whether or not the confidence regions described in the previous sections contained the true value θ_0. Figure 8 shows the observed coverage rates $\gamma_{0.95}$ as a function of the number of data points N, on the basis of 50,000 Monte Carlo simulations.

With these numbers, the 95% confidence intervals for $\gamma_{0.95}$, obtained from the binomial distribution, is 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 result of 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 (73) and the alternative (59) that provide ellipsoidal confidence regions turn out to be unreliable for small N. The reliability of the LR based confidence region (75) turns out to be relatively high, but suboptimal when compared to the Rao

	Taylor approximation	Ellipsoidal	Convergence condition
OE-standard	yes	yes	$V_N''(\theta_0)^{-1}V_N'(\theta_0)$
OE-alt-1a	no	yes	$\mathbf{V}^T\mathbf{e}_F$
OE-alt-1b	no	no	$\mathbf{V}^T\mathbf{e}$
OE-alt-2	yes, alternative	yes	$\mathbf{V}^T\mathbf{e}$
T_J	no	no	$\hat{\boldsymbol{\theta}}_N$
T_W = OE-impl	no	yes	$\hat{\boldsymbol{\theta}}_N$
T_R	no	no	–
T_{LR}	no	no	$V_N(\hat{\boldsymbol{\theta}}_N)$

Table 2
Overview of properties of OE model uncertainty sets based on different test statistics. Indication of the involvement of a Taylor approximation, an ellipsoidal structure of the parameter uncertainty set, and the random variable on the basis of which the asymptotic distribution is derived. OE-impl is the algorithm that is typically implemented in current practice.

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.

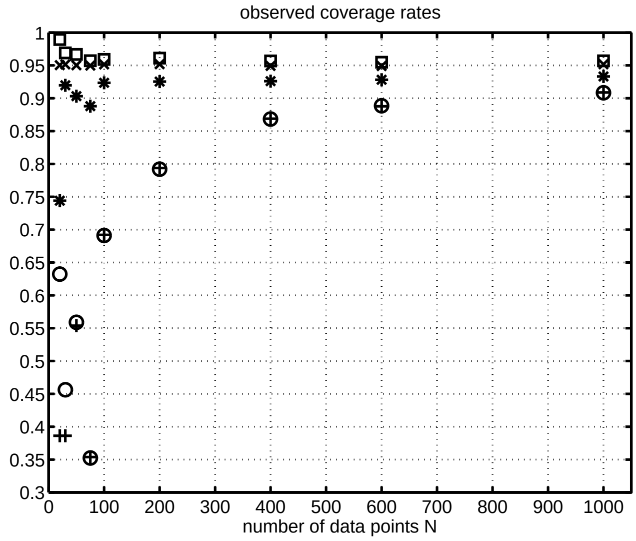

Fig. 8. Observed coverage rates of the confidence regions based on the ”classical” approach of Result 6 (OE-Alt-2) (68) and T_W (73) (○), the alternative approaches OE-Alt-1a (59)(+) and OE-Alt-1b (63)(*), the LR based approach based on T_{LR} (75)(□), and the Rao test based on T_R (74)(×), as a function of N. The data generating system is given by (77) and an OE model set is used with the same structure. The nominal confidence level is 0.95. All results are obtained from 50000 realizations.

11 Discussion and Extensions

In the commonly applied uncertainty bounding algorithms often use is made of parameter linearizations, e.g. through a Taylor approximation. This also typically holds for the uncertainty analysis of variables that are nonlinear functions of the estimated parameters as e.g. zero/pole locations. Undesired effects of these linearizations can be avoided by considering alternatives based on level sets of the cost function (see e.g. [22] pages 326-327, [23],[15]). This appealing alternative is equivalent to the likelihood ratio test statistic discussed before, but comes at the cost of a loss of the ellipsoidal parameter bounding structure, and therefore requires more computationally involved algorithms.

In the setting of this paper, the uncertainty bounding procedure as presented in [3] can be positioned as yet another choice of test statistic.

The alternative way of bounding parameter uncertainty as presented here has also been used fruitfully in the analysis of cross-correlation tests for model validation, see [11].

The presented approach is also directly applicable to instrumental variable estimators. In this case the parameter error will be given by

$$\hat{\theta}_N - \theta_0 = (Z^T\mathbf{\Phi})^{-1}Z^T\mathbf{e}$$

where Z is a (non-random) instrument matrix having elements that are correlated to the data, but uncorrelated to the noise. The equation above is rewritten as

$$\frac{1}{\sqrt{N}}(Z^T\mathbf{\Phi})(\hat{\theta}_N - \theta_0) = \frac{1}{\sqrt{N}}Z^T\mathbf{e}.$$

If the instruments in Z are constructed as filtered versions of the input signal with a pre-determined filter, (including a filter constructed on the basis of parameters that are estimated from a data set that is different than the one used for estimating $\hat{\theta}_N$), Z^T and $\mathbf{e}$ will be statistically independent and the presented finite-time results will fully apply, leading to the parameter uncertainty bound: $\theta_0 \in \mathcal{D}_{iv}(\alpha, \hat{\theta}_N)$, w.p. α,

$$\text{with } \mathcal{D}_{iv}(\alpha, \hat{\theta}_N) := \{\theta \mid N(\theta - \hat{\theta}_N)^T P_{iv}^{-1}(\theta - \hat{\theta}_N) \leq \chi^2_{d,\alpha}\} \quad (78)$$

$$\text{and } P_{iv} = \sigma_e^2 (\frac{1}{N} Z^T \Phi)^{-1} Z^T Z (Z^T \Phi)^{-1}. \quad (79)$$

being valid for any finite N, provided that the noise disturbance signal e is Gaussian distributed.

The presented approach has also good opportunities to be applied to the situation of considering asymptotic bias errors also (situation $\mathcal{S} \notin \mathcal{M}$), see e.g. [7] and [10], and to Box-Jenkins models, see [7].

12 Conclusions

In this paper alternative methods are presented for formulating probabilistic parameter uncertainty intervals for parameters that are identified in the prediction error identification framework. By exploiting the freedom to choose particular test statistics in the underlying hypothesis tests, alternative formulations result. For both ARX and OE models it follows that the standard *implemented* results have a stronger theoretical support than originally suggested, even for finite-length data sets. For OE models several different test statistics are analyzed. Uncertainty sets that are based on linearization (Taylor approximation) leading to ellipsoidal uncertainty sets, show a worse performance than tests that avoid this linearization. However this comes at the cost of computationally more involved algorithms for constructing the sets. It is shown that there exist OE uncertainty sets that are exact for finite-length data sets. The presented theory directly extends to instrumental variable estimators.

Acknowledgment

The authors acknowledge fruitful discussions with and contributions from Robert Bos and Asbjorn Klomp. Additionally Johan Schoukens, Rik Pintelon, Lennart Ljung and Michel Gevers are acknowledged for providing feedback on the presentation of this material.

References

[1] A. Azzalini. *Statistical Inference - Based on the Likelihood*. Chapman & Hall, London, UK, 1996.

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

[3] M.C. Campi and E. Weyer. Identification with finitely many data points: the LSCR approach. In *Prepr. 14th IFAC Symposium System Identification*, pages 46–64, Newcastle, NSW, Australia, March 2006.

[4] J. Chen and G. Gu. *Control Oriented System Identification*. Wiley Interscience, 2000.

[5] 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. 2007 European Control Conf.*, pages 2879–2886, Kos, Greece, 2-5 July 2007.

[6] 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 M.J. Chung, P. Misra, and H. Shim, editors, *Proc. 17th IFAC World Congress*, pages 5024–5029, Seoul, Korea, 6-11 July 2008.

[7] S. G. Douma. *From Data to Performance - System Identification Uncertainty and Robust Control Design*. PhD thesis, Delft Univ. Technology, Delft, The Netherlands, 2006.

[8] 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 Conf., CDC-ECC'05*, pages 4970–4975, Seville, Spain, December 2005.

[9] S. G. Douma and P. M. J. Van den Hof. Probabilistic model uncertainty bounding: an approach with finite-time perspectives. In *Prepr. 14th IFAC Symp. System Identification*, pages 1021–1026, Newcastle, NSW, Australia, March 2006.

[10] S. G. Douma and P. M. J. Van den Hof. Probabilistic uncertainty bounding in output error models with unmodelled dynamics. In *Proc. 2006 American Control Conf.*, pages 1677–1682, Minneapolis, MA, USA, June 2006.

[11] S.G. Douma, P.M.J. Van den Hof, and X. Bombois. Validity of the standard cross-correlation test for model structure validation. *Automatica*, 44(5):1285–1294, May 2008.

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

[13] R. G. Hakvoort and P. M. J. Van den Hof. Identification of probabilistic uncertainty regions by explicit evaluation of bias and variance errors. *IEEE Trans. Automatic Control*, 42(11):1516–1528, 1997.

[14] P. S. C. Heuberger, P. M. J. Van den Hof, and B. Wahlberg. *Modelling and Identification with Rational Orthogonal Basis Functions*. Springer Verlag, 2005.

[15] H. Hjalmarsson. From experiment design to closed-loop control. *Automatica*, 41(3):393–438, 2005.

[16] S.M. Kay. *Fundamentals of Statistical Signal Processing, Volume II Detection Theory*. Prentice Hall, Upper Saddle River, NJ, 1998.

[17] L. Ljung. Model validation and model error modeling. In B. Wittenmark and A. Rantzer, editors, *The Åström Symposium on Control*, pages 15–42, Lund, Sweden, August 1999. Studentlitteratur.

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

[19] M. Milanese, J. Norton, H. Piet-Lahanier, and E. Walter. *Bounding Approaches to System Identification*. Plenum Press, New York, 1996.

[20] M. Milanese and M. Taragna. H_∞ set membership identification: a survey. *Automatica*, 41(12):2019–2032, 2005.

[21] B.M. Ninness and G.C. Goodwin. Estimation of model quality. *Automatica*, 31(12):1771–1797, December 1995.

[22] R. Pintelon and J. Schoukens. *System Identification: A Frequency Domain Approach*. IEEE Press, Piscataway, NJ, 2001.

[23] S.L. Quinn, T.J. Harris, and D.W. Bacon. Accounting for uncertainty in control-relevant statistics. *J. Process Control*, 15:675–690, 2005.

[24] T. Söderström and P. Stoica. *System Identification*. Prentice-Hall International, Hemel Hempstead, UK, 1989.

[25] P.M.J. Van den Hof, P.S.C. Heuberger, and J. Bokor. System identification with generalized orthonormal basis functions. *Automatica*, 31(12):1821–1834, 1995.

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

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

[28] S.S. Wilks. *Mathematical Statistics*. John Wiley & Sons, Inc., New York, USA, 3rd edition, 1974.

Appendix

Proof of (22)

The random variable $\mathbf{V}^T\mathbf{e}$ can be interpreted as a weighted sum of independent identically distributed random variables with variance σ_e^2. In the considered situation the weights are also random. By using an appropriate central limit theorem (see e.g. [?], page 569) this expression asymptotically converges to a normal distribution provided that $\lim_{N\to\infty}\mathbf{V}^T\mathbf{V}$ exists and is nonsingular. Since for every realization V it follows that $V^TV = I$, the limit expression simply equals the identity matrix, leading to the result (22).

Proof of Lemma 1

Define the vector valued function
$g(\cdot) : \mathbb{R}^{(n+n^2)\times 1} \to \mathbb{R}^{(n+n^2)\times 1}$, defined by

$$g(\mathbf{z}, \mathbf{v}) := \begin{bmatrix} \mathbf{e} \\ \mathbf{v} \end{bmatrix} = \begin{bmatrix} \mathbf{V} & 0 \\ 0 & I^{n^2\times n^2} \end{bmatrix} \begin{bmatrix} \mathbf{z} \\ \mathbf{v} \end{bmatrix}, \qquad (.1)$$

with $\mathbf{v} = col(\mathbf{V}^T)$ a vector containing all elements of $\mathbf{V}^T$. When denoting $\mathbf{e}' := [\mathbf{e}^T \;\; \mathbf{v}^T]^T$ and $\mathbf{z}' := [\mathbf{z}^T \;\; \mathbf{v}^T]^T$ it follows that

$$p_{\mathbf{z}}(z) = \int_v p_{\mathbf{z}'}(z')dv.$$

Using the mapping from $\mathbf{z}'$ to $\mathbf{e}'$ it follows from standard theory on the transformation of random variables [**?**] that

$$p_{\mathbf{z}'}(z') = p_{\mathbf{e}'}(g(z')) \cdot \det(J(g(z')))$$

with the Jacobian given by

$$J(g(z')) = \begin{bmatrix} V & Z \\ 0 & I^{n^2\times n^2} \end{bmatrix},$$

and Z containing the partial derivatives of $V\mathbf{z}$ to $\mathbf{v}$. Consequently

$$\begin{aligned} f_{\mathbf{z}}(z) &= \int_v f_{\mathbf{e}'}(g(z')) \cdot \det(J)dv \\ &= \int_v f_{\mathbf{e}}(Vz)f_{\mathbf{v}}(v) \cdot \det(J)dv \end{aligned}$$

where the latter equation follows from the fact that $\mathbf{e}$ and $\mathbf{V}$ are independent. Using the Gaussian distribution of $\mathbf{e}$ and the fact that $\det(J) = \det(V) = 1$ it follows that

$$\begin{aligned} f_{\mathbf{z}}(z) &= \int_v \frac{1}{\sigma(\sqrt{2\pi})^n} e^{-\frac{1}{2}\sigma^{-2}z^TV^TVz} \; f_{\mathbf{v}}(v)dv \\ &= \int_v \frac{1}{\sigma(\sqrt{2\pi})^n} e^{-\frac{1}{2}\sigma^{-2}z^Tz} \; f_{\mathbf{v}}(v)dv \\ &= f_{\mathbf{e}}(z)\int_v f_{\mathbf{v}}(v)dv = f_{\mathbf{e}}(z). \end{aligned}$$

□