Proceedings of the
44th IEEE Conference on Decision and Control, and
the European Control Conference 2005
Seville, Spain, December 12-15, 2005

An Alternative Paradigm for Probabilistic Uncertainty Bounding in Prediction Error Identification

Sippe G. Douma and Paul M.J. Van den Hof

Abstract— **In prediction error identification model uncertainty bounds are generally derived from the statistical properties of the parameter estimator. These statistical properties reflect the variability in the estimated model under repetition of experiments with different realizations of the measured signals. However when the primal interest of the identification is in quantifying the uncertainty in an estimated parameter on the basis of one single experiment, this is not necessarily the best and only approach. In the alternative paradigm that is presented here, not the covariance of the estimator will be used for bounding the model uncertainty, but an a posteriori bound on the error in the estimated parameter will be constructed that is structurally dependent on the particular data sequence. This will allow simpler computations for probabilistic model uncertainty bounds also applicable to the situation of approximate modelling ($\mathcal{S} \notin \mathcal{M}$) and to model structures that are nonlinear in the parameters, such as Output Error (OE) models.**

Index Terms— **model validation, system identification, undermodelling, model uncertainty.**

I. 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 [10] and $\mathcal{H}_\infty$ identification [1] 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., [11].

In the mainstream approach of system identification, i.e. prediction error 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., [9], [12]. This description leads to probabilistic confidence bounds on estimated parameters, from which also probabilistic confidence bounds on estimated frequency responses can be constructed, with any prechosen level of probability.

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 [13], [7]) this holds for finite data length, for ARX models this holds asymptotically. For general model structures, and under the assumption $\mathcal{S} \in \mathcal{M}$, the parameter covariance matrix can be approximated 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. [6], [8] and [7] Chapter 7.

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. Whereas the statistical properties of the parameter estimator reflect the variability in the estimated results upon experiment repetition, the quantification of parameter uncertainty on the basis of only *one experiment* can be done without the full analysis of the parameter estimator. This will be shown to facilitate uncertainty bounding in several ways.

After presenting the principle concepts of the new paradigm, parameter uncertainty regions are derived for ARX and OE models, for the situation that $\mathcal{S} \in \mathcal{M}$. Additionally it is shown how the new procedure can be applied to an OE approximate modelling procedure ($\mathcal{S} \notin \mathcal{M}$). Due to space limitations the discussion is kept brief. For more details the reader is referred to [4].

The authors are with the Delft Center for Systems and Control, Delft University of Technology, Mekelweg 2, 2628 CD Delft, The Netherlands. Tel. +31-15-2784509; Fax: +31-15-2786679; E-mail: p.m.j.vandenhof@dcsc.tudelft.nl.

0-7803-9568-9/05/$20.00 ©2005 IEEE

II. Estimator properties and uncertainty regions

It is standard practice to base the characterization of the quality of a parameter estimate $\hat{\theta}$ on the (statistical) properties of the estimator, where the estimator is defined as a mapping:

$$\boldsymbol{\theta} = g(\mathbf{z})$$

where $\mathbf{z}$ indicates the measurements. In this section boldface symbols are used to distinguish random variables from realizations thereof.

The classical way of arriving at a model uncertainty bound is:

- Assume knowledge of $p_{\boldsymbol{\theta}}(\theta)$, given by prior information and/or by application of the Central Limit Theorem;
- Assume that the estimator is unbiased, i.e. $\theta_0 = \mathbb{E}\boldsymbol{\theta}$;
- Then knowing that every estimate $\hat{\theta}$ is a realization of the random variable with p.d.f. $p_{\boldsymbol{\theta}}(\theta)$, this pdf with expected value θ_0 can be used to statistically bound the difference $|\hat{\theta} - \theta_0|$.

Having in mind this classical way of determining uncertainty bounds let's consider the following example.

Example 1: 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 of which it is only known that they satisfy $|\boldsymbol{x}_2| \leq 2$ and $|\boldsymbol{x}_1 - 3| \leq 1$. We consider the following estimator for θ_0:

$$\boldsymbol{\theta} = \frac{\mathbf{y}}{\boldsymbol{x}_1}. \tag{1}$$

Substituting the expression for $\mathbf{y}$ in the estimator expression, shows that

$$\theta_0 + \min\left\{\frac{\boldsymbol{x}_2}{\boldsymbol{x}_1}\right\} \leq \boldsymbol{\theta} \leq \theta_0 + \max\left\{\frac{\boldsymbol{x}_2}{\boldsymbol{x}_1}\right\}.$$

In other words, the statistical properties of the estimator lead to the expression

$$\boldsymbol{\theta} \in \theta_0 + [-1, 1], \tag{2}$$

and this expression is valid not only for the random variable $\boldsymbol{\theta}$, but then also for every realization $\hat{\theta}$ of $\boldsymbol{\theta}$.

Now, given one particular estimate $\hat{\theta}$ a relevant question is '*how far is $\hat{\theta}$ removed from θ_0?*' or in practice '*what bound on the difference $|\hat{\theta} - \theta_0|$ can be established based on assumptions and data?*' In standard practice this question of model uncertainty is answered by using the estimator properties, that is, the answer would be

$$(\hat{\theta} - \theta_0) \in [-1, 1]. \tag{3}$$

However, a much more accurate answer is readily available using the fact that $\boldsymbol{x}_1$ is measured as well. Suppose the particular estimate $\hat{\theta}$ was obtained from a measurement y and $x_1 = 4$. Then $\hat{\theta} = \frac{y}{x_1} = \theta_0 + \frac{x_2}{x_1} = \theta_0 + \frac{x_2}{4}$. The exact realization x_2 is unknown but the properties of the random variable $\boldsymbol{x}_2$ can now be used to specify the parameter uncertainty:

$$(\hat{\theta} - \theta_0) \in \left[-\frac{2}{4}, \frac{2}{4}\right]. \tag{4}$$

The example is mainly intended to illustrate that for the construction of parameter uncertainty bounds it is not strictly necessary to employ the full statistical properties of the *estimator*. However, the suggestion is also raised that the alternative model uncertainty bound (4) is always smaller than the one derived from the estimator statistics (3). But this is not true in general.

In this example only hard bounds are used leading necessarily to a worst-case description in the estimator statistics. When another distribution of the random terms is assumed, either of the approaches can lead to the smallest region, since the size of the bound (4) depends on the particular realization.

Note that the uncertainty region (4) would have followed as the result of the estimator statistics if the assumption would have been adopted that the input x_1 is deterministic. This difference in assumptions is apparently essential for the estimator statistics, but is not necessarily crucial for the construction of parameter uncertainty regions on the basis of one single experiment.

To further illustrate the conceptual and computational advantages of the approach considered here, the previous example is slightly extended.

Example 2: Consider the situation of Example 1 but now assume that the random numbers are Gaussian distributed and correlated. Under these conditions the estimator (1) satisfies

$$\boldsymbol{\theta} = \frac{\mathbf{y}}{\boldsymbol{x}_1} = \theta_0 + \frac{\boldsymbol{x}_2}{3 + \frac{1}{2}\boldsymbol{x}_2} \text{ for } \boldsymbol{x}_2 \in \mathcal{N}(0, 2). \tag{5}$$

The probability density function of this estimator is rather complex and will generally not be Gaussian[1]. Therefore, evaluation of parameter uncertainty regions on the basis of $p_{\boldsymbol{\theta}}$ will be cumbersome.

However since $\boldsymbol{x}_1(\boldsymbol{\theta} - \theta_0) = \boldsymbol{x}_2$, and a particular pair $x_1, \hat{\theta}$ is available from the measurement (y, x_1), it can easily be verified that

$$x_1(\hat{\theta} - \theta_0) = x_2 \tag{6}$$

where the term on the right hand side is unknown. Using the prior information that x_2 is a realization of the random variable $\boldsymbol{x}_2$ it simply follows that

$$(\hat{\theta} - \theta_0)x_1^2(\hat{\theta} - \theta_0) \leq \sigma^2_{\boldsymbol{x}_2} c_\chi(\alpha, 1) \quad \text{w.p. } \alpha, \tag{7}$$

where $c_\chi(\alpha, 1)$ corresponds to a probability level α in the Chi-squared distribution with one degree of freedom, i.e. the α probability region under a one-dimensional Gaussian distribution. Expression (7) is easily derived and, more importantly, the probability level associated with the bound is exact.

Since the distribution of the right hand side of (6) is known, we now consider the test statistic

$$x_1(\hat{\theta} - \tilde{\theta})$$

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

and select all the values of $\tilde{\theta}$ that lead to an $x_1(\hat{\theta}-\tilde{\theta})$ that is within the α probability level of the Gaussian distribution of $\boldsymbol{x}_2$. This set is exactly given by

$$\mathcal{D}(\alpha,\hat{\theta}) = \left\{\theta \ |x_1(\hat{\theta}-\theta)|^2 \leq \sigma^2_{\boldsymbol{x}_2} c_\chi(\alpha,1)\right\} \tag{8}$$

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

III. The conceptual difference

The correctness of expression (7) in Example 2 might seem questionable at first sight: can we make a solid probabilistic expression of parameter uncertainty bounds without knowing the p.d.f of the parameter estimator? However the probabilistic expression is solid and correct.

In the classical approach first an α-probability region $\mathcal{D}(\alpha,\theta_0)$ would be constructed for the estimator $\boldsymbol{\theta}$ by upper-bounding the distance from $\theta_0 = \mathbb{E}\boldsymbol{\theta}$, leading to, e.g.,

$$\mathcal{D}(\alpha,\theta_0) := \left\{\theta \ \mid \ |\theta-\theta_0|^2 \leq c_{p_{\boldsymbol{\theta}}}(\alpha)\right\}$$

with $c_{p_{\boldsymbol{\theta}}}(\alpha)$ such that $\boldsymbol{\theta} \leq c_{p_{\boldsymbol{\theta}}}(\alpha)$ w.p. α.[2] Subsequently, an uncertainty region $\mathcal{D}(\alpha,\hat{\theta})$ for a given estimate $\hat{\theta}$ is constructed using the same upper-bound $c_{p_{\boldsymbol{\theta}}}(\alpha)$ to describe the distance from θ_0, i.e.

$$\mathcal{D}(\alpha,\hat{\theta}) := \left\{\theta \ \mid \ |\theta-\hat{\theta}|^2 \leq c_{p_{\boldsymbol{\theta}}}(\alpha)\right\}.$$

Per definition, however, the upper-bound $c_{p_{\boldsymbol{\theta}}}(\alpha)$ on the distance between θ_0 and $\hat{\theta}$ was only correct for those estimates $\hat{\theta} \in \mathcal{D}(\alpha,\theta_0)$. In other words, the statement holds that

$$\theta_0 \in \mathcal{D}(\alpha,\hat{\theta}) \text{ w.p. } \alpha, \tag{9}$$

which means 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 will contain the true parameter only a number of αn times if $n \to \infty$.

Instead of considering the properties of the estimator $\boldsymbol{\theta}$ to evaluate the estimate $\hat{\theta}$, the alternative form considers the properties of $\boldsymbol{x}_2 \in \mathcal{N}(0,2)$. For any realization x_2 of this random variable it holds that

$$x_2^2 \leq \sigma^2_{x_2} c_\chi(\alpha,1) \quad \text{w.p. } \alpha,$$

which states that the inequality $x_2^2 \leq \sigma^2_{\boldsymbol{x}_2} c_\chi(\alpha,1)$ is true only αn times out of n experiments for $n \to \infty$. Now, since for each realization x_2 expression (6) holds true, it follows that the inequality

$$|x_1(\hat{\theta}-\theta_0)|^2 \leq \sigma^2_{\boldsymbol{x}_2} c_\chi(\alpha,1) \tag{10}$$

will be true w.p. α as well. And again,with

$$\mathcal{D}_{new}(\alpha,\hat{\theta}) := \left\{\theta \mid |x_1(\hat{\theta}-\theta)|^2 \leq \sigma^2_{\boldsymbol{x}_2} c_\chi(\alpha,1)\right\} \tag{11}$$

the statement $\theta_0 \in \mathcal{D}_{new}(\alpha,\hat{\theta})$ w.p. α means that when we construct the uncertainty region $\mathcal{D}_{new}(\alpha,\hat{\theta})$ for n experiments, i.e. n realizations of $\boldsymbol{x}_2$, the constructed region will contain the true parameter only a number of αn times if $n \to \infty$.

Whereas the classical approach considers experiment repetition for analysis of the variation in the estimated parameter $\hat{\theta}$, the new paradigm considers the statistical properties of a data-dependent mapping of the parameter, i.e. $\boldsymbol{x}_1(\hat{\boldsymbol{\theta}}-\theta_0)$.

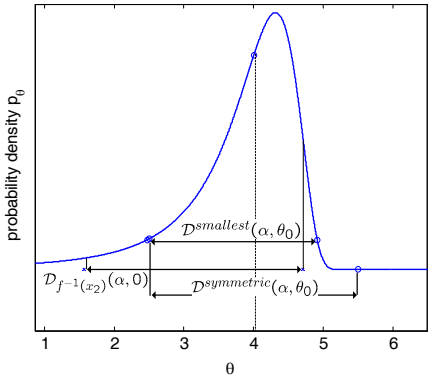

Fig. 1. Probability density function of $\boldsymbol{\theta}$ (example 2) and three uncertainty regions each corresponding to a probability of $\alpha = .9$. The symmetric and smallest 90% regions are tied to the pdf of $\boldsymbol{\theta}$. The computed 90% region $\mathcal{D}_{f^{-1}(x_2)}(\alpha,0)$ corresponds to all $\hat{\theta}$ for which the derived bound (10) is correct. This region is based on a 90% probability region of random variable $\boldsymbol{x}_2$.

IV. ARX modelling

In prediction error identification with ARX models a one-step-ahead predictor is considered of the format

$$\hat{y}(t|t-1;\theta) = \varphi^T(t)\theta \tag{12}$$

with $\varphi^T(t) = [-y(t-1)\cdots-y(t-n_a)\ u(t)\cdots u(t-n_b+1)]$, and $\theta^T = [a_1\cdots a_{n_a}\ b_0\cdots b_{n_b-1}]$, both having dimensions $n = n_a + n_b$. The parameter estimate is obtained by minimizing the quadratic prediction error criterion

$$\hat{\theta}_N = \arg\min_\theta V_N(\theta); \quad V_N(\theta) = \frac{1}{N}\sum_{t=1}^{N}\varepsilon(t,\theta)^2 \tag{13}$$

with $\varepsilon(t,\theta) = y(t) - \hat{y}(t|t-1;\theta)$. By denoting

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

it follows that $\hat{\theta}_N = (\Phi^T\Phi)^{-1}\Phi^T\mathbf{y}$.If the data generating system belongs to the model class ($\mathcal{S} \in \mathcal{M}$) then it holds that $\mathbf{y} = \Phi\theta_0 + \mathbf{e}$ with $\mathbf{e}$ an N-vector of samples from a white noise process, and so

$$\hat{\theta}_N = \theta_0 + (\Phi^T\Phi)^{-1}\Phi^T\mathbf{e}. \tag{14}$$

A. Classical approach

When analyzing the statistical properties of the estimator, it is generally derived[3] that, for $N \to \infty$,

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

[2]For Gaussian distributions such 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.

[3]For this derivation it is required that both the terms $(\Phi^T\Phi)^{-1}$ and $\Phi^T\mathbf{e}$ as well as their product converge almost surely.

with

$$P_{arx} = (\mathbb{E}[\frac{1}{N}\Phi^T\Phi])^{-1} \cdot Q \cdot (\mathbb{E}[\frac{1}{N}\Phi^T\Phi])^{-1} \tag{15}$$

with $Q = \mathbb{E}[\frac{1}{N}\Phi^T\Phi] \cdot \sigma_e^2$ and σ_e^2 the variance of the white noise, leading to

$$P_{arx} = (\mathbb{E}[\frac{1}{N}\Phi^T\Phi])^{-1} \cdot \sigma_e^2. \tag{16}$$

This leads to the expression that, asymptotically in N,

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

$$\mathcal{D}_{arx}(\alpha, \hat{\theta}_N) := \{\theta \mid (\theta - \hat{\theta}_N)^T P_{arx}^{-1}(\theta - \hat{\theta}_N) \leq \frac{c_\chi(\alpha, n)}{N}\}.$$

One of the problems in the latter expression is that P_{arx} cannot be computed, since σ_e^2 and $\mathbb{E}[\frac{1}{N}\Phi^T\Phi]$ are not available directly. Therefore in practice, the exact covariance matrix P_{arx} is commonly replaced by an estimate

$$\hat{P}_{arx} = (\frac{1}{N}\Phi^T\Phi)^{-1} \cdot \hat{\sigma}_e^2, \tag{17}$$

where the estimate $\hat{\sigma}_e^2$ is usually determined on the basis of $\varepsilon^2(t, \hat{\theta}_N)$.

B. New approach

Alternatively we can use expression (14) to analyze for the particular data sequence $\{u(t), y(t)\}_{t=0,\cdots N-1}$ that is available, the expression

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

The unknown term on the right hand side of the equation is known to satisfy

$$\beta = \frac{1}{\sqrt{N}}\Phi^T\mathbf{e} \in \mathcal{N}(0, Q), \quad Q = \mathbb{E}[\frac{1}{N}\Phi^T\Phi]\sigma_e^2$$

where the Gaussian distribution is reached asymptotically in N as a result of the Central Limit Theorem.
For the random variable $\beta = \frac{1}{\sqrt{N}}\Phi^T\mathbf{e}$, the following uncertainty bound can be specified asymptotically in N:

$$\beta \in \mathcal{D}_\beta(\alpha, 0) \text{ w.p.}\alpha, \text{ with}$$

$$\mathcal{D}_\beta(\alpha, 0) := \left\{\beta \mid \beta^T Q^{-1}\beta \leq c_\chi(\alpha, n)\right\}$$

As this probabilistic expression is also valid for the particular estimate $\beta = \frac{1}{\sqrt{N}}\Phi^T\Phi(\hat{\theta}_N - \theta_0)$, it follows that

$$(\hat{\theta}_N - \theta_0)^T \frac{1}{N}\Phi^T\Phi Q^{-1} \frac{1}{N}\Phi^T\Phi(\hat{\theta}_N - \theta_0) \leq \frac{c_\chi(\alpha, n)}{N} \text{ w.p. } \alpha$$

and consequently $\theta_0 \in \mathcal{D}_{arx}(\alpha, \hat{\theta}_N)$ w.p. α, with

$$\mathcal{D}_{arx}(\alpha, \hat{\theta}_N) := \{\theta \mid (\theta - \hat{\theta}_N)^T P_{arx,n}^{-1}(\theta - \hat{\theta}_N) \leq \frac{c_\chi(\alpha, n)}{N}\} \tag{18}$$

with

$$P_{arx,n} = (\frac{1}{N}\Phi^T\Phi)^{-1} Q (\frac{1}{N}\Phi^T\Phi)^{-1}. \tag{19}$$

Note that this expression is very close to the classical expression (15). However instead of the three asymptotes that were required to be satisfied in the classical case, the current expression only requires the a.s. convergence of $\frac{1}{\sqrt{N}}\Phi^T e$. Again, as in the classical case, since σ_e^2 and $\mathbb{E}[\frac{1}{N}\Phi^T\Phi]$ are unknown, they are replaced by their estimates $\hat{\sigma}_e^2$ and $\frac{1}{N}\Phi^T\Phi$, leading to (18).

$$\hat{P}_{arx,n} = (\frac{1}{N}\Phi^T\Phi)^{-1}\sigma_e^2. \tag{20}$$

Whereas in the classical approach P_{arx} has the interpretation of covariance matrix of the parameter estimator, this interpretation is not applicable to the matrix $P_{arx,n}$. The latter expression only serves as a basis for the parameter uncertainty region.

C. Evaluation

In its implementable form (20) the alternative approach is seen to result in exactly the same uncertainty region as is practically used in the classical approach (17), based on the theoretical result (15). However, when comparing the two theoretical expressions (15) and (19) it appears that, besides the replacement of σ_e^2 by an estimate, the latter approach requires only the replacement of Q by a computable estimate, while the former requires three substitutions to be made. Summarizing, through the new paradigm the commonly used uncertainty region based on (17) has a stronger theoretical support than is generally acknowledged.
Moreover, the Gaussian distribution in the standard approach requires the distribution of $(\Phi^T\Phi)^{-1}\Phi\mathbf{e}$ to be Gaussian, whereas the alternative approach only requires $\frac{1}{\sqrt{N}}\Phi\mathbf{e}$ to be Gaussian distributed. Monte Carlo simulations show that the term $\frac{1}{\sqrt{N}}\Phi\mathbf{e}$ becomes Gaussian even for very small data length N. The term $(\Phi^T\Phi)^{-1}\Phi\mathbf{e}$ generally requires a longer data length to approximate the Gaussian distribution.
The results presented here are applicable both to open-loop and closed-loop data. In the latter situation the input signal will also be correlated to the noise. This will influence the convergence properties of the terms $\frac{1}{N}\Phi^T\Phi$ and $\frac{1}{\sqrt{N}}\Phi^T\mathbf{e}$. Note that also in this situation the new approach has to deal with the second term only.

V. Output Error modelling

In an Output Error (OE) model structure we consider the one-step ahead predictor

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

and we denote the predictor derivative:

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

A. Classical approach

For quantifying parameter uncertainty bounds in the classical approach, the starting point is a first order Taylor expansion:

$$(\hat{\theta}_N - \theta_0) \sim -[\bar{V}''(\theta_0)]^{-1}[V_N'(\theta_0)] \tag{21}$$

where $V_N'(\theta) = \partial V_N(\theta)/\partial\theta$ and $\bar{V}''$ is the second derivative of $\bar{V} = \bar{\mathbb{E}}\varepsilon(t,\theta)^2$ ([9]). In the first order Taylor approximation, the asymptotic expressions

$$V_N'(\theta_0) \to \mathcal{N}(0,Q), \quad Q = \sigma_e^2 \mathbb{E}[\tfrac{1}{N}\Psi(\theta_0)^T\Psi(\theta_0)] \quad (22)$$

$$\bar{V}''(\theta_0) = \mathbb{E}[\tfrac{1}{N}\Psi^T(\theta_0)\Psi(\theta_0)] \quad (23)$$

with

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

are then substituted to arrive at the asymptotic covariance matrix of $(\hat{\theta}_N - \theta_0)$ given by

$$P_{oe} = \sigma_e^2[\mathbb{E}\frac{1}{N}\Psi(\theta_0)^T\Psi(\theta_0)]^{-1} \quad (24)$$

and the asymptotic model uncertainty region

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

$$\mathcal{D}_{oe}(\alpha,\hat{\theta}_N) := \{\theta \mid (\theta-\hat{\theta}_N)^T P_{oe}^{-1}(\theta-\hat{\theta}_N) \le \frac{c_\chi(\alpha,n)}{N}\}.$$

Note that the approximations and assumptions that are involved in this analysis include:

- Validity of the first order Taylor approximation;
- Convergence to (asymptotic) normality of the product term on the right hand side of (21), and
- Convergence of the covariance of this term to the product of the separate terms, as reflected in (22) and (23).

For arriving at a computable expression for the parameter uncertainty, the unknown terms σ_e^2 and $[\mathbb{E}\frac{1}{N}\Psi(\theta_0)^T\Psi(\theta_0)]$ that appear in the covariance matrix are replaced by estimates, to arrive at

$$\hat{P}_{oe} = \hat{\sigma}_e^2[\frac{1}{N}\Psi^T\Psi]^{-1}. \quad (25)$$

B. New approach

In our new paradigm, the starting point for the analysis of the parameter estimate is the derivative of the identification criterion: $V_N'(\hat{\theta}_N) = 0$ or equivalently

$$\frac{1}{N}\sum_{t=1}^{N}[y(t) - \frac{B(q,\hat{\theta}_N)}{F(q,\hat{\theta}_N)}u(t)]\cdot\psi(t,\hat{\theta}_N) = 0. \quad (26)$$

By defining

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

equation (26) can be rewritten as

$$\frac{1}{N}\sum_{t=1}^{N}[F(q,\hat{\theta}_N)y_F(t) - B(q,\hat{\theta}_N)u_F(t)]\cdot\psi(t,\hat{\theta}_N) = 0. \quad (27)$$

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

$$\hat{\theta}_N = (\Psi^T\Phi)^{-1}\Psi^T\mathbf{y}_F \quad (28)$$

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

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

being a vector with dimension $n = n_b + n_f$, and $\mathbf{y}_F = [y_F(1)\cdots y_F(N)]^T$.

Note that (28) is an equation that characterizes $\hat{\theta}_N$; however it cannot be used to *calculate* $\hat{\theta}_N$, as the right hand side of the equation is also dependent on $\hat{\theta}_N$. Nevertheless the equation can fruitfully be used to characterize the parameter uncertainty on $\hat{\theta}_N$.

To this end we write the system's relations as:

$$y(t) = \frac{B_0(q)}{F_0(q)}u(t) + e(t), \quad \text{or} \quad (29)$$

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

which can be written in the regression form:

$$\mathbf{y}_F = \Phi\theta_0 + \mathbf{e}_F, \quad (30)$$

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

Substituting (30) into (28) now delivers:

$$\hat{\theta}_N - \theta_0 = (\Psi^T\Phi)^{-1}\Psi^T\mathbf{e}_F, \text{ or} \quad (31)$$

$$(\frac{1}{\sqrt{N}}\Psi^T\Phi)(\hat{\theta}_N - \theta_0) = \frac{1}{\sqrt{N}}\Psi^T\mathbf{e}_F. \quad (32)$$

To bound the parameter uncertainty in $\hat{\theta}_N$, now the same procedure can be followed as applied to the ARX case.

The random variable $\boldsymbol{\beta} = \frac{1}{\sqrt{N}}\Psi^T\mathbf{e}_F$ is asymptotically Gaussian distributed with zero mean and covariance matrix

$$Q = \sigma_{\mathbf{e}_F}^2\mathbb{E}[\frac{1}{N}\Psi^T\Psi]$$

Mapping the uncertainty region of $\boldsymbol{\beta}$ to $\hat{\theta}_N - \theta_0$ now leads to the asymptotic uncertainty region

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

$$\mathcal{D}_{oe,n}(\alpha,\hat{\theta}_N) := \{\theta \mid (\theta-\hat{\theta}_N)^T P_{oe,n}^{-1}(\theta-\hat{\theta}_N) \le \frac{c_\chi(\alpha,n)}{N}\}$$

with

$$P_{oe,n} = (\frac{1}{N}\Psi^T\Phi)^{-1}\cdot Q\cdot(\frac{1}{N}\Phi^T\Psi)^{-1}. \quad (33)$$

In order to arrive at a computable expression for the parameter uncertainty, $\sigma_{\mathbf{e}_F}^2$ is replaced by an estimate $\hat{\sigma}_e^2$ and $\mathbb{E}[\frac{1}{N}\Psi^T\Psi]$ is replaced by $[\frac{1}{N}\Psi^T\Psi]$ leading to the estimate

$$\hat{P}_{oe,n} = (\frac{1}{N}\Psi^T\Phi)^{-1}\cdot\frac{1}{N}\Psi^T\Psi\cdot(\frac{1}{N}\Phi^T\Psi)^{-1}\hat{\sigma}_e^2. \quad (34)$$

Note again, as in the ARX situation, that $P_{oe,n}$ does not have the interpretation of a covariance matrix. It only serves to specify the parameter uncertainty region.

C. Evaluation

Unlike the situation for ARX models, the actual expressions for the classical and the new approach now are different. In the new analysis, e.g., no first order Taylor expansion is involved, and asymptotic normality of the term $\frac{1}{\sqrt{N}}\Psi^T\mathbf{e}$ is the only stochastic convergence issue that is involved.

VI. Output error modelling, $\mathcal{S} \notin \mathcal{M}$

So far it has not been possible to quantify parameter uncertainty bounds in situations of approximate models ($\mathcal{S} \notin \mathcal{M}$), except for the situation of linear regression structures with deterministic regressors. See e.g. [5], [2], [6], [7]. The new paradigm presented here shows that also for models that are not linear in the parameters (as e.g. OE models), uncertainty bounds can be derived in a way that relates strongly to the handling of linear regression models. This is reflected by expression (28) that is used to specify the parameter uncertainty bounds. For incorporating unmodelled dynamics the system's equations, as given in (29) are rewritten as

$$y(t) = \frac{B_0(q)}{F_0(q)}u(t) + G_\Delta(q)u(t) + e(t)$$

where the system's coefficient vector θ_0 that is composed of the coefficients of the polynomials B_0 and F_0 is defined as the minimizing argument: $\theta_0 = \arg\min_\theta \bar{\mathbb{E}}\varepsilon(t,\theta)^2$ for an identification experiment with sufficiently exciting input signal. This leads to a system's equation

$$\mathbf{y}_F = \Phi\theta_0 + \mathbf{e}_F + \beta$$

where β is the undermodelling term induced by $G_\Delta(q)u(t)$. Substituting this expression into (28) now delivers

$$\hat{\theta}_N - \theta_0 = (\Psi^T\Phi)^{-1}\Psi^T[\mathbf{e}_F + \beta] \quad \text{or equivalently}$$
$$\frac{1}{\sqrt{N}}\Psi^T\Phi(\hat{\theta}_N - \theta_0) = \frac{1}{\sqrt{N}}\Psi^T[\mathbf{e}_F + \beta].$$

There are several options now to proceed. The term on the right hand side can be bounded in a stochastic way, using the stochastic embedding approach [5]. As an alternative, the right hand side term $\frac{1}{\sqrt{N}}\Psi^T\beta$ can be bounded in a deterministic way, while the remaining term $\frac{1}{\sqrt{N}}\Psi^T\mathbf{e}_F$ can be treated stochastically as in the earlier OE analysis. This second approach ([6]) then comes down to writing

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

Since the right hand side term is asymptotically Gaussian distributed:

$$\frac{1}{\sqrt{N}}\Psi^T\mathbf{e}_F \to \mathcal{N}(0,Q), \quad Q = \sigma^2_{\mathbf{e}_F}[\mathbb{E}\frac{1}{N}\Psi^T\Psi]$$

it follows that asymptotically

$$(\frac{1}{\sqrt{N}}\Psi^T\Phi(\hat{\theta}_N-\theta_0)-\frac{1}{\sqrt{N}}\Psi^T\beta)^T\cdot Q^{-1}\cdot(\cdot) \leq c_\chi(\alpha,n) \text{ w.p. } \alpha.$$

By defining the square root term Γ by

$$\Gamma^T\Gamma = \frac{1}{N}\Phi^T\Psi Q^{-1}\frac{1}{N}\Psi^T\Phi$$

the expression for the uncertainty region can be rewritten as

$$\|\Gamma[(\hat{\theta}_N - \theta_0) - (\Psi^T\Phi)^{-1}\Psi^T\beta\|_2^2 \leq \frac{c_\chi(\alpha,n)}{N} \text{ w.p. } \alpha.$$

Now applying the triangle inequality to the norm expression delivers

$$\|\Gamma(\hat{\theta}_N - \theta_0)\|_2 \leq \sqrt{\frac{c_\chi(\alpha,n)}{N}} + \|\Gamma(\Psi^T\Phi)^{-1}\Psi^T\beta\|_2 \text{ w.p. } \alpha.$$

The second term on the right hand side can be bounded through an upper bound on the unmodelled dynamics. As a result, an ellipsoidal parameter uncertainty region becomes available that quantifies the uncertainty in $\hat{\theta}_N$ with a prespecified level of probability $\geq \alpha$.

VII. Conclusions

The standard approach of formulating probabilistic parameter bounds on the basis of the statistical properties of the parameter estimator is discussed. In this paper an alternative approach is followed for deriving parameter uncertainty regions based on the analysis of data-dependent mappings of the parameter estimator. Many of the approximations and asymptotic assumptions in the standard results can be avoided in this way. For ARX models it follows that the standard *implemented* results have a stronger theoretical support than originally suggested. For OE models the new paradigm leads to new parameter uncertainty regions that, in terms of their computation, are similar in complexity to the linear regression situation. The new paradigm also allows the application to the situation where unmodelled dynamics is present ($\mathcal{S} \notin \mathcal{M}$). BJ model structures are treated in [4]. It seems that the new paradigm bears resemblance to the so-called likelihood method of determining confidence regions [3]; this relation is subject of ongoing research.

Acknowledgements

The authors acknowledge fruitful discussions with Robert Bos, Arjan den Dekker and Xavier Bombois.

References

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

[2] D.K. de Vries and P.M.J. Van den Hof. Quantification of uncertainty in transfer function estimation: A mixed probabilistic-worst-case approach. *Automatica*, 31(4):543–557, 1995.

[3] J.R. Donaldson and R.B. Schnabel. Computational experience with confidence regions and confidence intervals for nonlinear least squares. *Technometrics*, 29(1):67–82, 1987.

[4] S.G. Douma. *From Data to Robust Control - System Identification Uncertainty and Robust Control Design*. PhD thesis, Delft Center for Systems and Control, Delft Univ. Technology, 2006.

[5] G.C. Goodwin, M. Gevers, and B.M. Ninness. Quantifying the error in estimated transfer functions with application to model order selection. *IEEE Trans. Autom. Control*, 37:913–928, 1992.

[6] 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.

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

[8] 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.

[9] L. Ljung. *System Identification — Theory for the User*. Prentice-Hall, Upper Saddle River, NJ, second edition, 1999.

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

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

[12] T. Söderström and P.G. Stoica. *System Identification*. Prentice-Hall International, Hemel Hempstead, U.K., 1989.

[13] 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.