Assessing the quality of identified models through the asymptotic theory - When is the result reliable ?

S. Garatti [a] M.C. Campi [b] S. Bittanti [a]

[a] *Dipartimento di Elettronica ed Informazione - Politecnico di Milano,*
Piazza Leonardo da Vinci 32, 20133 Milano, Italy -
{bittanti,sgaratti}@elet.polimi.it

[b] *Dipartimento di Elettronica per l'Automazione - University of Brescia,*
Via Branze 38, 25123 Brescia, Italy - campi@ing.unibs.it -
http://bsing.ing.unibs.it/~campi/

Abstract

In this paper, the problem of estimating uncertainty regions for identified models is considered. A typical approach in this context is to resort to the asymptotic theory of Prediction Error Methods (PEM) for system identification, by means of which ellipsoidal uncertainty regions can be constructed for the uncertain parameters.

We show that the uncertainty regions worked out through the asymptotic theory can be unreliable in certain situations, precisely characterized in the paper.

Then, we critically analyze the theoretical conditions for the validity of the asymptotic theory, and prove that the asymptotic theory also applies under new assumptions which are less restrictive than the usually required ones. Thanks to this result, we single out the classes of models among standard ones (ARX, ARMAX, Box Jenkins, etc.) where the asymptotic theory can be safely used in practical applications to assess the quality of the identified model.

These results are of interest in many applications, including iterative controller design schemes.

Key words: system identification, model quality assessment, asymptotic theory, ARMAX, Box-Jenkins

[⋆] This paper was not presented at any IFAC meeting. Corresponding author Marco C. Campi - Tel. +39-030-3715458 - Fax 39-030-380014 - E-mail: campi@ing.unibs.it

1 Introduction

Consider a data-generating dynamical system P and a model $\widehat{P}$ of it estimated from data. It has been fully recognized in the literature that the estimated model $\widehat{P}$ is of little use without a statement on its quality. In other words, it is fundamental to characterize the error model, i.e. the distance between P and $\widehat{P}$ (see e.g [1], [7], [10], [16], [19] and [23]).
The most commonly used tool for evaluating the error model is the asymptotic theory of Prediction Error Methods (PEM) for system identification. It returns ellipsoidal confidence regions in the space of parameters such that the true system parameters belong to this ellipsoid with a specified probability (see e.g. [15] and [22]).
The main advantage of using the asymptotic theory is that the confidence regions can be easily computed from the available data. Moreover, these confidence regions are often reliable and give a tight description of uncertainty.
On the other hand, asymptotic theory has its own drawbacks too.
First, its applicability substantially requires the absence of un-modelled dynamics, while in real applications this assumption never applies (even if in many cases it does approximately). The importance of undermodelling is witnessed by many recent contributions such as [6], [8], [9], [14], [18], [16], [17], [21] and [24]. In order to overcome the problems encountered when the system order is unknown, certain formulas valid for both the model order and the number of data points growing unbounded have been derived, see e.g. [13], [15], [20], [27].
A second drawback is that the asymptotic theory is rigorously correct only when the number of data tends to infinity in such a way that the total amount of information on the system parameters grows unbounded. On the other hand, in real applications it often happens that the amount of excitation is substantial for certain parameters while there is a lack of information on other parameters (poor excitation). As a consequence, the asymptotic theory is used as a heuristic tool for the model quality evaluation.

In this paper, we focus attention on the problems arising when data are not informative enough, and one of our aims is to pinpoint the situations where the asymptotic theory may fail to provide sensible results with poor excitation. In these situations, the estimated parameters are subject to large uncertainty levels and the asymptotic theory can as well provide misleading results. This is quite a severe limitation since assessing the model quality is especially important for large uncertainty levels. Indeed, in the opposite case, the estimated model can be safely used in place of the true system with no particular need for an evaluation of its uncertainty. This leads to our first contribution:

i) by way of an example, we explain why the asymptotic theory may fail for the model quality evaluation in presence of a high level of uncertainty.

We also note that this result is relevant to iterative control schemes where the closed-loop bandwidth is very restricted at the first iterations leading to poorly exciting signals and, in turn, to wide uncertainty in the estimated model (see [2], [5], [11], [25]).

We next move to establish the situations where the asymptotic theory does not suffer from the problem highlighted in point i) above, and it turns out that the asymptotic theory provides sensible results or not depending on the model class in which the data-generating system is identified. Our second contribution can be summarized as follows:

ii) we single out the model classes among standard ones (ARX, ARMAX, Box-Jenkins, etc.) such that the asymptotic theory can be safely used to assess model quality, even in presence of a high level of uncertainty.

This latter result is made possible by a new asymptotic result, valid under relaxed assumptions, also worked out in this paper.

A different approach can be adopted in the analysis of uncertainty in the estimate by explicitly considering the finiteness of the data record. For some recent contributions along this line see [3], [4] and [26].

Structure of the paper

In Section 2, our working assumptions are stated and a brief summary of the classical asymptotic theory is given. This allows us to keep the paper self-contained. Section 3 delivers the example as explained in point i) above. After a mid-paper conclusion section (Section 4), Section 5 contains the new asymptotic result valid under relaxed assumptions. In Section 6, we move to consider the quality assessment with finite data points and show the relevance of the theorem in Section 5 to this purpose. Finally, in Section 7 the classes of models to which the asymptotic theory can be safely applied for model quality estimation are singled out, while some illustrative simulations are given in Section 8.

2 Asymptotic theory of Prediction Error Methods

In this section we provide a compendium of the asymptotic theory of Prediction Error Methods for system identification with the objective of clarifying the context of our results. For a more comprehensive description of the subject, we refer the reader to the literature (see e.g. [15] and [22]).

2.1 Mathematical setting

Let

$$\mathcal{M}_\vartheta = \left\{ \widehat{y}(t,\vartheta) = W_u(z^{-1},\vartheta)u(t) + W_y(z^{-1},\vartheta)y(t),\ \vartheta \in \Theta \subseteq \mathbb{R}^n \right\} \tag{1}$$

be a parameterized set of predictor models, where $W_u(z^{-1},\vartheta)$ and $W_y(z^{-1},\vartheta)$ satisfy the following assumption.

Assumption 1 *$W_u(z^{-1},\vartheta)$ and $W_y(z^{-1},\vartheta)$ are rational strictly proper (as functions in z) transfer functions whose coefficients are functions of a parameter $\vartheta \in \Theta$, where Θ is a nonempty compact set in $\mathbb{R}^n$. The coefficients are four times differentiable with respect to ϑ and the fourth derivatives are continuous. Moreover, $W_u(z^{-1},\vartheta)$ and $W_y(z^{-1},\vartheta)$ are asymptotically stable, $\forall \vartheta \in \Theta$.*

Remark 1 *In the classical asymptotic theory, the coefficients of the transfer functions $W_u(z^{-1},\vartheta)$ and $W_y(z^{-1},\vartheta)$ are usually only required to be twice differentiable with continuous second derivatives. Here, the assumption has been strengthened in view of our further results. It is perhaps worth mentioning that for standard identification model classes (ARX, ARMAX, Box-Jenkins, etc.) the coefficients are the parameters themselves, so that the differentiability assumption is not an issue.*

u and y are respectively the input and output of the system, and are generated according to the following scheme.

Assumption 2 *Processes u and y are given by*

$$u(t) = G_u(z^{-1})r(t) + H_u(z^{-1})e(t) \tag{2}$$
$$y(t) = G_y(z^{-1})r(t) + H_y(z^{-1})e(t), \tag{3}$$

where $G_u(z^{-1})$, $G_y(z^{-1})$, $H_u(z^{-1})$ and $H_y(z^{-1})$ are asymptotically stable rational transfer functions. $e(t)$ is a zero mean independent process with constant variance equal to $\lambda^2 > 0$ and such that $\sup_t \mathbb{E}[|e(t)|^{4+\delta}] < \infty$, for some $\delta > 0$. $r(t)$ is a wide sense stationary, ergodic, stochastic, external input sequence. $e(t)$ and $r(t)$ are independent.

Remark 2 *The results given below can be proved even if $r(t)$ is a bounded deterministic external input sequence. Considering a stationary, ergodic reference as in Assumption 2 has been preferred since it simplifies the presentation.*

Remark 3 *Note that Assumption 2 encompasses closed-loop as well as open-loop configurations. In the latter, $H_u(z^{-1}) = 0$ and $G_u(z^{-1}) = 1$.*

We also require that the data-generating system belongs to the class of models $\mathcal{M}_\vartheta$, that is:

Assumption 3 *There exists a parameter ϑ^o, which is an interior point of Θ, such that*

$$y(t) = W_u(z^{-1}, \vartheta^o)u(t) + W_y(z^{-1}, \vartheta^o)y(t) + e(t). \tag{4}$$

Remark 4 *When the data-generating system does not belong to the assumed class of model $\mathcal{M}_\vartheta$, the system-model mismatch comprises two terms: a variance term and a bias term. In this case the asymptotic theory applies so as to only assess the variance term at the price of a more complicated formulation that accounts for the correlation in the residue due to the bias term. See e.g. [8] and [9].*

Parameter ϑ is estimated by the minimization of the standard quadratic cost:

$$V_N(\vartheta) = \frac{1}{N}\sum_{t=1}^{N} \varepsilon(t, \vartheta)^2,$$

where N is the number of data points and $\varepsilon(t, \vartheta) = y(t) - \widehat{y}(t, \vartheta)$ is the prediction error.
Thus, the estimate is

$$\widehat{\vartheta}_N = \arg\min_{\vartheta \in \Theta} V_N(\vartheta).$$

The asymptotic cost criterion is $\overline{V}(\vartheta) = \mathbb{E}[\varepsilon(t, \vartheta)^2]$, and we will denote by Θ^* the corresponding set of minimizers within the feasible set Θ, that is

$$\Theta^* = \Big\{ \arg\min_{\vartheta \in \Theta} \overline{V}(\vartheta) \Big\}.$$

In the classical asymptotic theory it is assumed that $\overline{V}(\vartheta)$ has a unique minimizer:

Assumption 4 *The set Θ^* has cardinality equal to 1.*

Remark 5 *Under Assumption 3, it is easy to demonstrate that the parameter ϑ^o always belongs to the set Θ^*. Therefore, Assumption 4 can be rewritten as $\Theta^* = \{\vartheta^o\}$.*

2.2 Asymptotic theory results

Let

$$Q_N = \frac{\frac{1}{N}\sum_{t=1}^{N}\psi(t,\widehat{\vartheta}_N)\psi(t,\widehat{\vartheta}_N)'}{\frac{1}{N}\sum_{t=1}^{N}\varepsilon(t,\widehat{\vartheta}_N)^2},$$

where $\psi(t,\vartheta)$ denote $\frac{\mathrm{d}}{\mathrm{d}\vartheta}\varepsilon(t,\vartheta)$, and consider the following ellipsoid centered in $\widehat{\vartheta}_N$:

$$\mathcal{E}(r) = \Big\{\vartheta : (\widehat{\vartheta}_N - \vartheta)'Q_N(\widehat{\vartheta}_N - \vartheta) \le r\Big\}, \tag{5}$$

where r is a real positive number called the size of the ellipsoid.
The standard result of the asymptotic theory writes as follows:

Theorem 1 *Let $p \in [0,1)$ and assume that $\frac{\mathrm{d}^2}{\mathrm{d}\vartheta^2}\overline{V}(\vartheta^o) > 0$. Under Assumptions 1, 2, 3 and 4, it follows that*

$$\lim_{N\to\infty}\mathbb{P}\Big\{\vartheta^o \in \mathcal{E}\Big(\frac{\alpha(p)}{N}\Big)\Big\} = p,$$

where $\alpha(p)$ is the inverse of the function $p = \int_0^\alpha f_{\chi^2}(x)\mathrm{d}x$ and $f_{\chi^2}(x)$ is the probability density of a χ^2 random variable with n degrees of freedom.

The above theorem suggests how to select r so as to obtain an ellipsoidal confidence region for ϑ^o of pre-assigned asymptotic probability p. The proof of Theorem 1 can be found in Chapter 9 of [15].
The following result is obtained immediately from Theorem 1.

Theorem 2 *Assume that $\frac{\mathrm{d}^2}{\mathrm{d}\vartheta^2}\overline{V}(\vartheta^o) > 0$. Under Assumptions 1, 2, 3 and 4, for any sequence α_N which tends to ∞ as $N \to \infty$, we have that*

$$\lim_{N\to\infty}\mathbb{P}\Big\{\vartheta^o \in \mathcal{E}\Big(\frac{\alpha_N}{N}\Big)\Big\} = 1.$$

Remark 6 *As a natural choice for α_N, consider $\alpha_N = \alpha(p)(1+\beta_N)$, for some p, with $\beta_N \to \infty$ as $N \to \infty$, that is, the ellipsoid size is inflated by the factor $1+\beta_N$ with respect to Theorem 1. If $\frac{\beta_N}{N} \to 0$, when $N \to \infty$, the ellipsoid size still tends to zero, though with a slower rate than the ellipsoid of Theorem 1. Theorem 2 says that, no matter how slow such an inflation takes place, the true parameter ϑ^o will asymptotically belong to the ellipsoid with confidence* 1. *A good choice of β_N is reliant on the specific problem at hand and its value is dictated by experience.*

In real applications, the asymptotic theory is often used to generate confidence regions for the system parameters, even if, as is obvious, such a theory applies only approximately since the evaluation is based on a finite number of data points. Though it is common experience that the results are still reliable in many cases even for a moderate data sample, it is also true that in other situations the asymptotic theory may fail to provide sensible indications, even for large set of data points.

The goal of the present paper is to give a clearcut view of the situations in which this actually occurs and to pinpoint the model classes for which the asymptotic theory can be safely used. We start in the next section with an example clarifying where the trouble may come from in the use of the asymptotic theory.

3 An example where the asymptotic theory provides misleading results with poorly informative data

Consider the following data-generating system:

$$y(t) = \frac{b^o z^{-1}}{1 + a^o z^{-1}} u(t) + (1 + h^o z^{-1}) e(t), \tag{6}$$

where $a^o = -0.7$, $b^o = 0.3$, $h^o = 0.5$ and $e(t) \sim WGN(0,1)$ ($WGN =$ White Gaussian Noise). In addition, the plant is operated in closed loop as shown in Figure 1. It is a trivial task to verify that the closed loop system is stable.

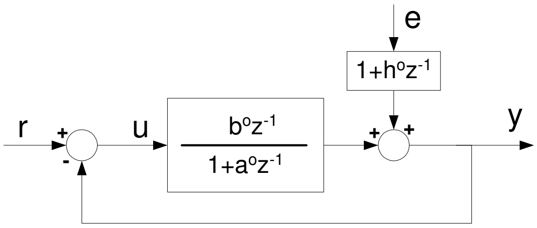

Figure 1. The real plant

$N = 10000$ data points (u, y) have been collected when the system was operated with a reference signal $r(t) = \mathrm{WGN}(0, 10^{-6})$, independent of $e(t)$ (note that the variance of the reference signal is very small as compared to the noise variance - poor excitation). Based on the (u, y) measurements, a full order model for the data-generating system (6) has been identified and a confidence region $\mathcal{E}(\frac{\alpha(p)}{N})$, $p = 0.99$, has also been estimated through the asymptotic Theorem 1.

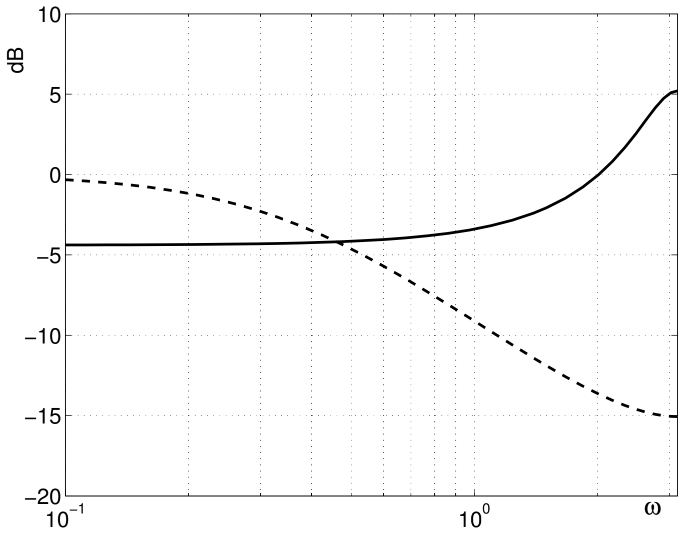

Figure 2. Amplitude Bode plot of the real plant (- -) and of the estimated model (—)

The amplitude Bode diagrams of the identified model and of the real system u to y transfer functions have been plotted in Figure 2.

From the plot, a wide mismatch between the real plant and the identified model is apparent. This is not surprising, since the reference signal is poorly exciting. On the other hand, we would also expect that the uncertainty region supplied by the asymptotic theory is wide.

Figure 3 displays the confidence region $\mathcal{E}(\frac{\alpha(p)}{N})$ in the frequency domain. Surprisingly, the confidence region concentrates around the identified model, showing that the model quality assessment is completely unreliable in this case.

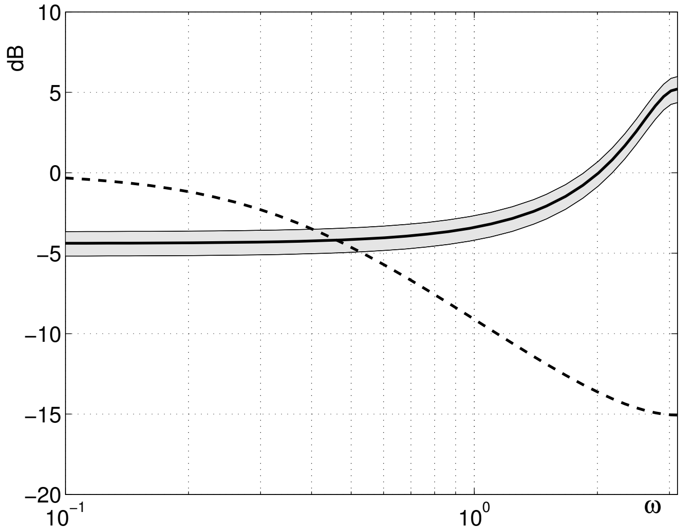

Figure 3. Uncertainty region of the estimated model vs. real plant Bode diagram

It is perhaps interesting to note that the presented situation – though admittedly artificial – is a simplification of what often happens in practical identification, where poor excitation is due to a restricted bandwidth of the closed-loop system. The simplified situation of a poorly exciting external signal $r(t)$ has been adopted here for ease of presentation.

Explanation

Let us briefly explain the mechanism that made the model quality estimation unreliable in the present situation.

The explanation becomes easier if we assume that the reference signal is exactly equal to zero. For this reason we concentrate for a moment on the case $r(t) = 0$ and we return to the case where $r(t)$ has a small variance further below.

For $r(t) = 0$, a simple computation shows that:

$$\overline{V}(\vartheta) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left| \frac{1 + h^o z^{-1}}{1 + h z^{-1}} \cdot \frac{1 + (a+b) z^{-1}}{1 + (a^o + b^o) z^{-1}} \cdot \frac{1 + a^o z^{-1}}{1 + a z^{-1}} \right|^2_{z = e^{j\omega}} \mathrm{d}\omega, \tag{7}$$

where $\vartheta = [a\ b\ h]'$.

The minimal value of $\overline{V}(\vartheta)$ is 1 and it is easy to see that the minimum is achieved if and only if every monomial at the numerator is cancelled by another monomial at the denominator. This happens only in the following two cases:

$$\begin{array}{rcl|rcl} a_1^* + b_1^* & = & a^o + b^o & a_2^* + b_2^* & = & a^o + b^o \\ a_1^* & = & a^o & a_2^* & = & h^o \\ h_1^* & = & h^o & h_2^* & = & a^o. \end{array}$$

Thus, there are just two distinct minima of the asymptotic cost criterion, one of which corresponds to the true system. Figure 4 represents $\overline{V}(\vartheta)$ along the line connecting the two minimizers.

Turn now to the case where $r(t)$ is a $WGN(0, 10^{-6})$, that is, to the actual situation. Here, the minimizer of the asymptotic cost criterion $\overline{V}(\vartheta)$ is unique, as the asymptotic theory prescribes, and coincides with ϑ^o. The other minimum becomes a local minimum. Yet, the difference between the values taken by $\overline{V}(\vartheta)$ at the two minimizers will be very small.

When identification is performed in practice, the empirical cost $V_N(\vartheta)$ has to be used in place of $\overline{V}(\vartheta)$. Since a finite number of data points is available, $V_N(\vartheta)$ is only an imprecise replica of $\overline{V}(\vartheta)$ so that the global minimizer of $V_N(\vartheta)$ may as well happen to be near the minimizer of $\overline{V}(\vartheta)$ which does not correspond to the real plant parameter (this is what happened in our simulation results). If so, $\widehat{\vartheta}_N$ gets trapped far from ϑ^o.

It is important to reassert the fact that such a behavior is a consequence of the poorness of the available information. In turn, this is primarily due to the

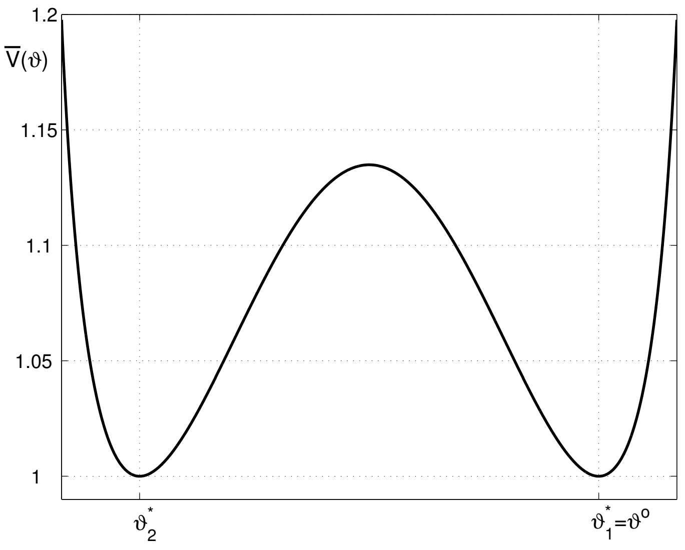

Figure 4. $\overline{V}(\vartheta)$ along the line connecting ϑ_1^* and ϑ_2^*

poor excitation conveyed by each single data point (since $r(t)$ is very small) and secondarily to the finiteness of the number of data points (so that the total amount of information in the data is limited).

In order to explain why the confidence region provided by the asymptotic theory is not reliable, it is, at this point, necessary to recall an aspect of the asymptotic theory which is relevant to the present discussion (see [15] and [22] for details).

Theorems 1 and 2 are both based on the following fundamental expansion:

$$0 = \sqrt{N}\frac{\mathrm{d}}{\mathrm{d}\vartheta}V_N(\widehat{\vartheta}_N) = \sqrt{N}\frac{\mathrm{d}}{\mathrm{d}\vartheta}V_N(\vartheta^o) + \frac{\mathrm{d}^2}{\mathrm{d}\vartheta^2}V_N(\xi_N)\sqrt{N}(\widehat{\vartheta}_N - \vartheta^o). \tag{8}$$

This equation is nothing but the Taylor expansion of $\frac{\mathrm{d}}{\mathrm{d}\vartheta}V_N$ (where all terms are inflated by the coefficient $\sqrt{N}$ and ξ_N is a point between ϑ^o and $\widehat{\vartheta}_N$). The evaluation of the confidence region for $\widehat{\vartheta}_N - \vartheta^o$ is carried out by observing that: first, $\sqrt{N}\frac{\mathrm{d}}{\mathrm{d}\vartheta}V_N(\vartheta^o)$ is asymptotically a zero mean Gaussian random variable; second, $\frac{\mathrm{d}^2}{\mathrm{d}\vartheta^2}V_N(\xi_N)$ converges to $\frac{\mathrm{d}^2}{\mathrm{d}\vartheta^2}\overline{V}(\vartheta^o)$, since $\widehat{\vartheta}_N \to \vartheta^o$ so that ξ_N is squeezed towards ϑ^o. The quantity $\frac{\mathrm{d}^2}{\mathrm{d}\vartheta^2}\overline{V}(\vartheta^o)$ is further approximated by $\frac{\mathrm{d}^2}{\mathrm{d}\vartheta^2}V_N(\widehat{\vartheta}_N)$ leading to the asymptotic Theorems 1 and 2.

If $\widehat{\vartheta}_N$ is sufficiently close to ϑ^o, this last approximation concerning the second derivative has a negligible effect. However, in the previous example this is not so, since the estimate $\widehat{\vartheta}_N$ is trapped far from ϑ^o and this is the reason for the misleading result as shown in Figure 3.

Let us explain more in detail the mechanism through which such a misleading result is generated.

Due to the effect of the inflating coefficient $\sqrt{N}$, $\sqrt{N}(\widehat{\vartheta}_N - \vartheta^o)$ takes on quite a large value. Despite this, equation (8) holds true (equation (8) is always true since it contains no approximation). In fact, in (8) $\frac{\mathrm{d}^2}{\mathrm{d}\vartheta^2}V_N$ is computed in a point ξ_N between ϑ^o and $\widehat{\vartheta}_N$ where $\frac{\mathrm{d}^2}{\mathrm{d}\vartheta^2}V_N(\xi_N)$ is almost singular, leading to a term $\frac{\mathrm{d}^2}{\mathrm{d}\vartheta^2}V_N(\xi_N)\sqrt{N}(\widehat{\vartheta}_N - \vartheta^o)$ of moderate magnitude. Unfortunately, as explained before, ξ_N is not accessible and $\frac{\mathrm{d}^2}{\mathrm{d}\vartheta^2}V_N(\xi_N)$ is substituted by $\frac{\mathrm{d}^2}{\mathrm{d}\vartheta^2}V_N(\widehat{\vartheta}_N)$ which turns out to be well positive definite. This leads to the mistaken conclusion that $\widehat{\vartheta}_N - \vartheta^o$ is small and to the unreliable uncertainty region shown in Figure 3.

Note that a second interpretation of the obtained result is also possible: For $r(t) = 0$ the found region is in fact a confidence region around ϑ_2^*, the spurious minimizer different from ϑ^o. When $r(t) = WGN(0, 10^{-6})$ the found confidence region can be interpreted as a perturbation of the previous one.

4 Mid paper conclusions

The results of the previous sections can be summarized as follows:

i) the classical asymptotic theory requires that the asymptotic cost criterion has a unique minimizer $\vartheta^* = \vartheta^o$; moreover if data are poorly informative so that $\widehat{\vartheta}_N$ is not close enough to ϑ^* (wide uncertainty), then the resulting uncertainty evaluation by means of $\mathcal{E}(\frac{\alpha(p)}{N})$ can be unreliable, i.e. the asymptotic theory results do not hold, even approximately;

ii) due to i), a blind application of the asymptotic theory can lead to misleading results.

In the next sections our goal is to study the situations where the asymptotic theory provides reliable results, even when $\widehat{\vartheta}_N$ is far from ϑ^o. To this purpose, we proceed along the following lines:

iii) we extend the asymptotic theory results so as to encompass the case of multiple minimizers of the asymptotic cost criterion $\overline{V}(\vartheta)$ (Section 5);

iv) thanks to the result of point iii), we show that – if a suitable additional condition on the model class is satisfied – then the asymptotic theory can be safely used even for a high level of uncertainty, namely for $\widehat{\vartheta}_N$ far from ϑ^o (Section 6);

v) we establish which standard model classes (ARMAX, Box-Jenkins, etc.) satisfy the additional condition of point iv) (Section 7).

5 A new asymptotic result

In this section, we provide a new asymptotic result which generalizes the standard asymptotic theory of Section 2. The fact that this result is useful when data are poorly informative is discussed in the next Section 6.
Assumption 4 in Section 2 is here replaced by the following one.

Assumption 4′ *$\Theta^* = \mathcal{S} \cap \Theta$, where $\mathcal{S}$ is an affine subspace of the parameter space $\mathbb{R}^n$.*
Moreover, $\widehat{\vartheta}_N \to \vartheta^*$ (not necessary equal to ϑ^o) almost surely, where $\vartheta^* \in \Theta^*$ is an interior point of Θ.

Remark 7 *Note that Assumptions 1, 2, 3 and 4′ are more general than Assumptions 1, 2, 3 and 4. Indeed, Assumption 4 implies that $\Theta^* = \{\vartheta^o\}$, so that the first part of Assumption 4′ holds with $\mathcal{S} = \{\vartheta^o\}$, which is an affine subspace (it is the origin of $\mathbb{R}^n$ translated).*
As for the second part of Assumption 4′, it holds under Assumptions 1, 2, 3 and 4 with $\vartheta^* = \vartheta^o$.

Remark 8 *In Assumption 4′ the important fact is that Θ^* is* linearly *structured (apart from the fact that it is confined to Θ).*

In the following Theorem 3 we show that the asymptotic Theorem 2 can be preserved in the present setting.

Theorem 3 *Assume that $\frac{\mathrm{d}^2}{\mathrm{d}\vartheta^2}\overline{V}(\vartheta^*)$ is positive definite along the directions of $\mathcal{S}^\perp$ (the subspace orthogonal to $\mathcal{S}$). Under Assumptions 1, 2, 3 and 4′, for any sequence α_N which tends to ∞ as $N \to \infty$, we have that (see (5) for the definition of $\mathcal{E}(\cdot)$)*

$$\lim_{N\to\infty} \mathbb{P}\Big\{\vartheta^o \in \mathcal{E}\Big(\frac{\alpha_N}{N}\Big)\Big\} = 1. \tag{9}$$

Proof: see the Appendix.

Remark 9 *In contrast to Theorems 1 and 2, in Theorem 3 the positive definiteness of $\frac{\mathrm{d}^2}{\mathrm{d}\vartheta^2}\overline{V}(\vartheta^*)$ is only required in the directions of $\mathcal{S}^\perp$. In this connection, one could note that $\frac{\mathrm{d}^2}{\mathrm{d}\vartheta^2}\overline{V}(\vartheta^*)$ is in fact singular in the direction of $\mathcal{S}$ due to Assumption 4′.*

Remark 10 *Allowing for multiple minimizers of $\overline{V}(\vartheta)$, as is done in Theorem 3, permits to cope with situations where there is a lack of excitation (see Section 6 for further discussion).*

Remark 11 *Similarly to Remark 4 we mention here that the results of the* *new asymptotic Theorem 3 can be extended to the case in which the model class does not contain the true system. Clearly, in full analogy with Remark 4 this allows one to assess the variance term only, so that the ensuing results are perhaps less interesting than in the full order case. Details are omitted as a complete discussion of the matter would lead us too far afield.*

6 Use of Theorem 3 in practice

As we have seen in Section 3, in certain cases applying the asymptotic formulas to assess the quality of the identified model can lead to misleading conclusions. Here, we want to show that, under an additional condition on the model class, the asymptotic formulas can indeed be safely used for such an evaluation. This conclusion is possible in the light of the new asymptotic result stated in the previous section.

Let us go back for a moment to the example of Section 3. There, if $r(t) = 0$, then $\overline{V}(\vartheta)$ has two global isolated minimizers. When we performed the identification of the plant, instead of minimizing $\overline{V}(\vartheta)$ we had of course to resort to its empirical counterpart $V_N(\vartheta)$; moreover, $r(t)$ was small, but not equal to 0. Thus, the actual identification optimization setting can be seen as a perturbed setting with respect to the ideal one where one minimizes $\overline{V}(\vartheta)$ with $r(t) = 0$. As we have seen, $\widehat{\vartheta}_N$ can possibly fall near the minimizer of the ideal setting which does not correspond to the true system. If so, the asymptotic formulas lead to computing a deceivingly small uncertainty region.

We now introduce the following additional condition on the model class:

Condition 1 *Independently of the level of excitation in the signals, the set of the minimizers of $\overline{V}(\vartheta)$ is an affine subspace.*

Remark 12 *Condition 1, as is obvious, can be rewritten as* For every excitation level of the signals, there exists an affine subspace $\mathcal{S}$ such that $\Theta^* = \mathcal{S} \cap \Theta$. *However, the reader should note that this requirement is different from the first part of Assumption 4′ where one requires that $\Theta^* = \mathcal{S} \cap \Theta$ only for Θ^* arising in the particular operating condition, i.e. for a fixed level of excitation of the input signal.*

Now, suppose that a model class fulfilling Condition 1 is used. If we are in an ideal situation with a complete lack of excitation, then $\overline{V}(\vartheta)$ is minimized in an affine subspace, say $\mathcal{S}$, and Theorem 3 can be applied to this situation. If instead we are in a real identification setting where we minimize $V_N(\vartheta)$ and, possibly, some extra degree of excitation is added to the signals, this real setting can be seen as a perturbed setting of the ideal one. Thus, though $\widehat{\vartheta}_N$ is far from ϑ^o, Theorem 3 still holds approximately and formula (9) can be used for the model quality assessment.

As it appears, the asymptotic theory can be safely applied with poorly exciting data to the model classes for which the set of minimizers of $\overline{V}(\vartheta)$ is an affine subspace. Studying these classes is the subject of the next section, while simulation examples illustrating the result are shown in Section 8.

7 Assessment of the model classes for which $\overline{V}(\vartheta)$ is minimized in an affine subspace

We treat separately two different situations, namely open-loop identification and closed-loop identification as these two settings give different results.

7.1 Open-loop identification

By "open-loop identification" we mean that the input signal $u(t)$ and the noise signal $e(t)$ are independent. Technically speaking, this is equivalent to taking $H_u(z^{-1}) = 0$ and $G_u(z^{-1}) = 1$ in Assumption 2.

Theorem 4 *Let $\mathcal{M}_\vartheta$ be the Box-Jenkins (BJ) class of predictor models, i.e.*

$$\begin{aligned}\mathcal{M}_\vartheta = \Big\{ \; \widehat{y}(t,\vartheta) = \quad & (1 - H(z^{-1},\vartheta)^{-1})y(t) \\ & + H(z^{-1},\vartheta)^{-1}G(z^{-1},\vartheta)u(t), \;\; \vartheta \in \Theta \; \Big\},\end{aligned}$$

where G and H are rational transfer functions, $H(0,\vartheta) = 1$, $\forall\vartheta \in \Theta$, and ϑ is a vector containing the numerator and denominator polynomial coefficients of G and H.
Suppose that the identification is performed in open-loop and that Assumptions 1, 2 and 3 are satisfied.
Then, Condition 1 holds true.

Proof: see the Appendix.

Theorem 4 can be applied to Output Error (OE) models as well, since OE is a particular case of BJ. In fact, we remind that the OE predictor model class is

$$\mathcal{M}_\vartheta = \Big\{\widehat{y}(t,\vartheta) = G(z^{-1},\vartheta)u(t), \;\vartheta \in \Theta\Big\},$$

where G is a rational transfer function and ϑ is the vector of the numerator and denominator polynomial coefficients of G.

Even though Theorem 4 does not apply directly, a result similar to Theorem 4 holds for ARX and ARMAX models too. In this case,

$$\mathcal{M}_\vartheta = \Big\{\widehat{y}(t,\vartheta) = \Big(1 - \frac{A(z^{-1},\vartheta)}{C(z^{-1},\vartheta)}\Big)y(t) + \frac{B(z^{-1},\vartheta)}{C(z^{-1},\vartheta)}u(t),\ \vartheta \in \Theta\Big\},$$

where A, B and C are polynomials in z^{-1}, A and C are monic, and ϑ is the vector of the coefficients of these polynomials (the ARX case corresponds to $C(z^{-1},\vartheta) = 1$). One should note that in the ARX and ARMAX structures, $G(z^{-1},\vartheta)$ and $H(z^{-1},\vartheta)$ are not freely parameterized as assumed in Theorem 4. However, the proof of this theorem can be extended with minor amendments to cover the ARX and ARMAX cases.

It is perhaps worth mentioning that not all model structures satisfy Condition 1 even in open-loop. An example is given by the model class

$$A(z^{-1},\vartheta)y(t) = G(z^{-1},\vartheta)u(t) + H(z^{-1},\vartheta)e(t) \tag{10}$$

which corresponds to the predictor model class

$$\begin{aligned}\mathcal{M}_\vartheta = \Big\{\ \widehat{y}(t,\vartheta) &= (1 - A(z^{-1},\vartheta)H(z^{-1},\vartheta)^{-1})y(t) \\ &\qquad +H(z^{-1},\vartheta)^{-1}G(z^{-1},\vartheta)u(t), \quad \vartheta \in \Theta\Big\},\end{aligned}$$

where A is a monic polynomial in z^{-1}, G and H are rational transfer functions, $H(0,\vartheta) = 1$, $\forall\vartheta \in \Theta$, and ϑ is the vector of the coefficients of A and of the numerator and denominator polynomial coefficients of G and H. In Section 8 a simulation example involving this class of models is presented.

7.2 Closed-loop identification

Suppose now that the system is operated in closed-loop with a controller R as in Figure 5.

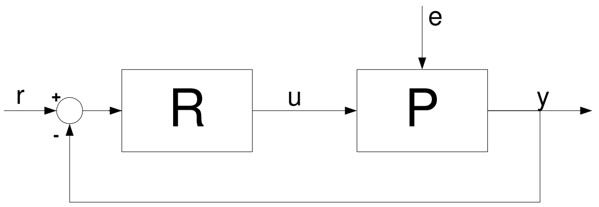

Figure 5. Closed loop system

Theorem 5 *Suppose that the identification is performed in closed-loop and that Assumptions 1, 2 and 3 are satisfied.*
Then, Condition 1 holds true for the ARMAX *and* OE *classes of models.*

Proof: see the Appendix.

It has to be noted that, when identification is performed in closed-loop, the Box-Jenkins structure does not meet Condition 1 in general. In fact, the example presented in Section 3 was based on a Box-Jenkins model.

8 Simulation examples

8.1 Example - BJ model of Section 3 in open-loop

Consider again the data-generating system described in (6), but suppose now that the system is operated in *open-loop* with an input signal $u(t) \sim WGN(0, 10^{-6})$, independent of $e(t)$. A full order model has been identified by means of the BJ model class with $N = 10000$. An ellipsoidal confidence region $\mathcal{E}(\frac{\alpha_N}{N})$, $\alpha_N = \alpha(p)$, $p = 0.99$, has been also estimated.
The identified model is shown in Figure 6. Again, as in Section 3, the model presents a large mismatch with the true system since the noise-to-signal ratio is large.

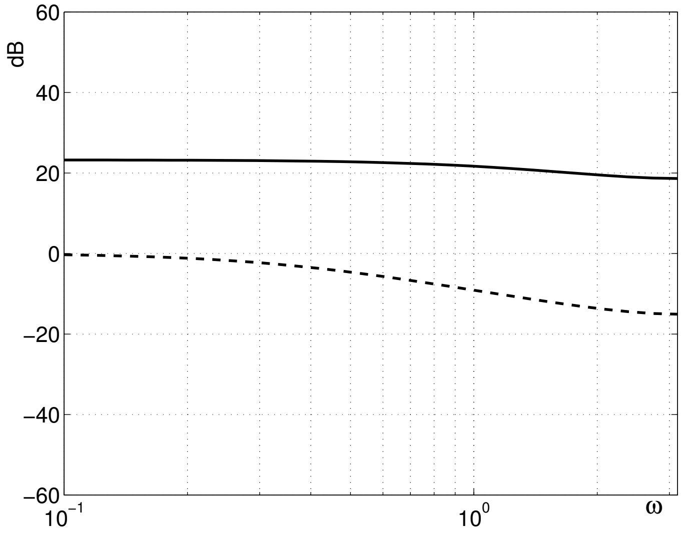

Figure 6. Amplitude Bode plot of the real plant (- -) and of the estimated model (—)

Figure 7 displays $\mathcal{E}(\frac{\alpha_N}{N})$ in the frequency domain.
The uncertainty region is very scattered in this case, and covers the gap be-

tween the identified model and the true plant. Thus, the estimated uncertainty is reliable, in agreement with Theorem 4.

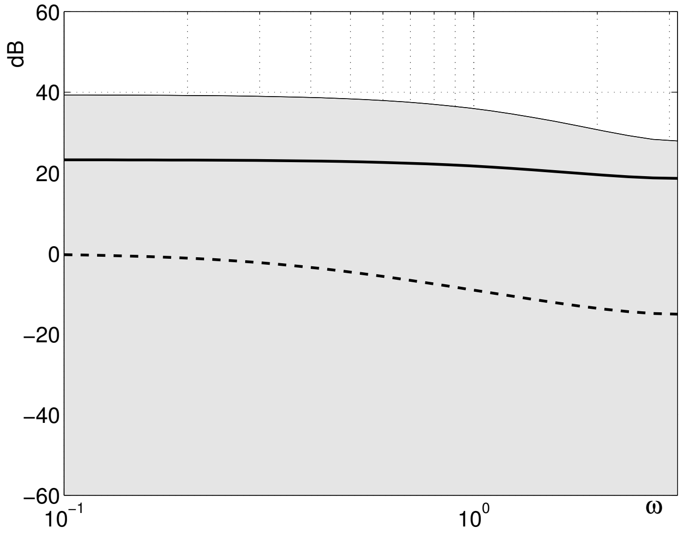

Figure 7. Uncertainty region of the estimated model

8.2 Example - a model class which does not meet Condition 1 in open-loop

Consider now the following data-generating system:

$$(1 + a^o z^{-1})y(t) = b^o z^{-1}u(t) + \frac{1}{1 + h^o z^{-1}}e(t), \tag{11}$$

where $a^o = -0.7$, $b^o = 0.3$, $h^o = 0.5$ and $e(t) \sim WGN(0, 1)$.
We have identified a full order model when the plant is operated in open-loop with a constant (poorly exciting) input signal $u(t) = 1$, $\forall t$, and $N = 10000$. An ellipsoidal confidence region $\mathcal{E}(\frac{\alpha_N}{N})$, $\alpha_N = \alpha(p)$, $p = 0.99$, has been also estimated.
System (11) belongs to the model class (10) and falls outside the realm of applicability of Theorem 4. The computed uncertainty region is displayed in Figure 8, showing that the asymptotic theory provides unreliable results. As a matter of fact it is not difficult to see that Condition 1 is violated in this case. Indeed, a simple computation shows that:

$$\overline{V}(\vartheta) = |1 + h|^2 \left| \frac{b^o}{1 + a^o} - \frac{b}{1 + a} \right|^2 + \frac{1}{2\pi} \int_{-\pi}^{\pi} \left| \frac{1 + hz^{-1}}{1 + h^o z^{-1}} \cdot \frac{1 + az^{-1}}{1 + a^o z^{-1}} \right|^2_{z=e^{j\omega}} \mathrm{d}\omega,$$

where $\vartheta = [a\ b\ h]'$.
The minimal value of $\overline{V}(\vartheta)$ is achieved only in the following two points:

$$\begin{array}{ll|ll} b_1^* & = 1 + a_1^* & b_2^* & = 1 + a_2^* \\ a_1^* = & \quad a^o & a_2^* = & \quad h^o \\ h_1^* = & \quad h^o & h_2^* = & \quad a^o, \end{array}$$

and, therefore, Condition 1 does not hold.

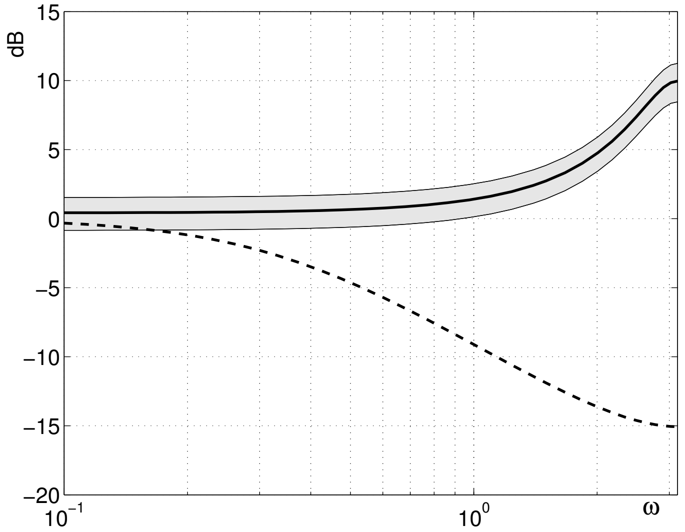

Figure 8. Uncertainty region of the estimated model

9 Concluding remarks

In this paper we have considered the problem of assessing the quality of identified models in a “Prediction Error” framework. Two main facts have been pointed out:

- in case of large uncertainty, the confidence regions supplied by the asymptotic theory may be unreliable;
- in spite of the presence of large uncertainty, the same confidence regions can be safely used if an extra condition holds true for the model class used in the identification procedure.

Moreover, we have provided a classification of the standard model classes (ARX, ARMAX, Box-Jenkins, etc.) which satisfy the extra condition.

The results of this paper can possibly be extended to new directions so as to cover other settings of interest in system identification. In particular, one could consider correlation approaches (e.g. instrumental variable methods), that play an important role in a number of applications.

Acknowledgment
Paper supported by MIUR under the National Research Project ”New methods for Identification and Adaptive Control for Industrial Systems”.

10 Appendix: proofs of the results

10.1 Proof of Theorem 3

We need a preliminary result.

Lemma 1 *Let $\overline{\vartheta}$ be a minimizer of $\overline{V}(\vartheta)$. Then, under Assumptions 1, 2, 3, it holds that*

$$\varepsilon(t,\overline{\vartheta}) = e(t) \textit{ almost surely.}$$

Proof
Since $\widehat{y}(t,\vartheta)$ depends on data up to time $t-1$ only (see Assumption 1), predictor $\widehat{y}(t,\vartheta)$ and $e(t)$ are independent for any ϑ.
Therefore, thanks to Assumption 3, we obtain that

$$\overline{V}(\vartheta) = \mathbb{E}\Big[\Big(e(t) + \widehat{y}(t,\vartheta^o) - \widehat{y}(t,\vartheta)\Big)^2\Big] = \mathbb{E}\Big[e(t)^2\Big] + \mathbb{E}\Big[\Big(\widehat{y}(t,\vartheta^o) - \widehat{y}(t,\vartheta)\Big)^2\Big].$$

Since $\overline{\vartheta}$ minimizes $\overline{V}(\vartheta)$, the term $\mathbb{E}\Big[\Big(\widehat{y}(t,\vartheta^o) - \widehat{y}(t,\overline{\vartheta})\Big)^2\Big]$ must be equal to 0 and this implies that $\widehat{y}(t,\vartheta^o) - \widehat{y}(t,\overline{\vartheta}) = 0$ almost surely.
Finally, $\varepsilon(t,\overline{\vartheta}) = y(t) - \widehat{y}(t,\overline{\vartheta}) = y(t) - \widehat{y}(t,\vartheta^o) = e(t)$ almost surely. □

Proof of Theorem 3
Recall that, by the definition (5) of $\mathcal{E}(\,\cdot\,)$, the condition

$$\vartheta^o \in \mathcal{E}\Big(\frac{\alpha_N}{N}\Big)$$

is equivalent to

$$(\widehat{\vartheta}_N - \vartheta^o)' Q_N (\widehat{\vartheta}_N - \vartheta^o) \leq \frac{\alpha_N}{N}.$$

As a consequence, the theorem can be proven by showing that

$$\lim_{N\to\infty} \frac{N}{\alpha_N} \cdot (\widehat{\vartheta}_N - \vartheta^o)' Q_N (\widehat{\vartheta}_N - \vartheta^o) = 0 \quad \text{in probability.} \tag{12}$$

Let d be the dimension of the affine subspace $\mathcal{S}$. Then, let $x \in \mathbb{R}^d$ $[z \in \mathbb{R}^{n-d}]$ be the first d [the remaining $n-d$] coordinates of ϑ, that is $\vartheta = [x' \; z']'$. Thus, $\vartheta^* = [(x^*)' \; (z^*)']'$, $\vartheta^o = [(x^o)' \; (z^o)']'$ and $\widehat{\vartheta}_N = [(\widehat{x}_N)' \; (\widehat{z}_N)']'$.
Without loss of generality we assume that $\mathcal{S}$ is parallel to the hyperplane determined by the x coordinates (this can be always achieved by a rotation of the axes). See Figure 9 for a graphical representation of the parameter space when $\Theta \subset \mathbb{R}^2$ and $\mathcal{S}$ is a straight line.
We now prove equation (12).

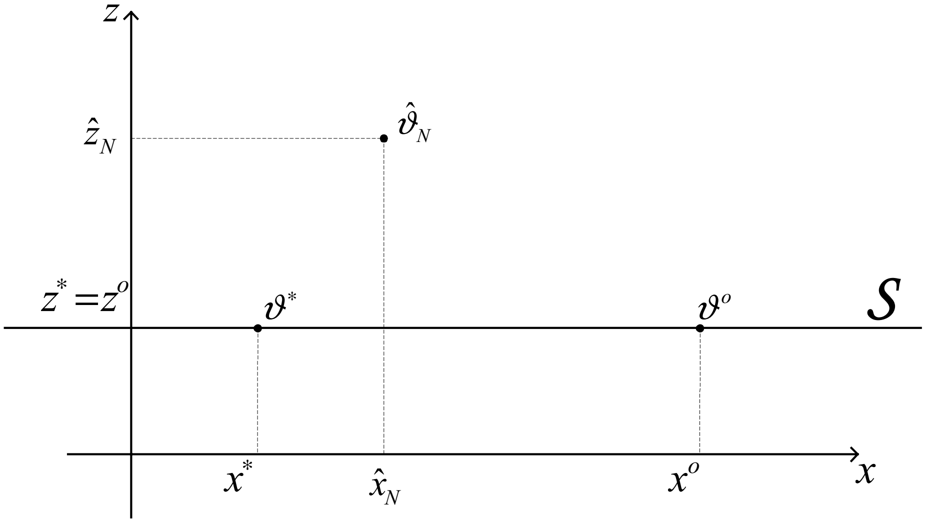

Figure 9. The parameter space

In order to avoid notational cluttering, throughout we omit the t-dependence, e.g. $\psi(\widehat{\vartheta}_N)$ stands for $\psi(t, \widehat{\vartheta}_N)$. Moreover, $\sum$ is used for $\sum_{t=1}^N$.
Since

$$Q_N = \frac{\frac{1}{N}\sum \psi(\widehat{\vartheta}_N)\psi(\widehat{\vartheta}_N)'}{\frac{1}{N}\sum \varepsilon(\widehat{\vartheta}_N)^2} = \frac{\frac{1}{N}\sum \begin{bmatrix} \varepsilon_x(\widehat{\vartheta}_N)\varepsilon_x(\widehat{\vartheta}_N)' & \varepsilon_x(\widehat{\vartheta}_N)\varepsilon_z(\widehat{\vartheta}_N)' \\ \varepsilon_z(\widehat{\vartheta}_N)\varepsilon_x(\widehat{\vartheta}_N)' & \varepsilon_z(\widehat{\vartheta}_N)\varepsilon_z(\widehat{\vartheta}_N)' \end{bmatrix}}{\frac{1}{N}\sum \varepsilon(\widehat{\vartheta}_N)^2},$$

where ε_x and ε_z denote the vector of derivatives of ε with respect to x and z coordinates, we have that

$$\begin{aligned}
&\frac{N}{\alpha_N} \cdot (\widehat{\vartheta}_N - \vartheta^o)' Q_N (\widehat{\vartheta}_N - \vartheta^o) \\
&= \frac{N}{\alpha_N \cdot \frac{1}{N} \sum \varepsilon(\widehat{\vartheta}_N)^2} \frac{1}{N} \sum \Big(\big((\widehat{x}_N - x^o)' \varepsilon_x(\widehat{\vartheta}_N) \big)^2 + \\
&\quad +2(\widehat{x}_N - x^o)' \varepsilon_x(\widehat{\vartheta}_N)(\widehat{z}_N - z^o)' \varepsilon_z(\widehat{\vartheta}_N) + \big((\widehat{z}_N - z^o)' \varepsilon_z(\widehat{\vartheta}_N) \big)^2 \Big) \\
&\leq \frac{N}{\alpha_N \cdot \frac{1}{N} \sum \varepsilon(\widehat{\vartheta}_N)^2} \frac{1}{N} \sum \Big(2\big((\widehat{x}_N - x^o)' \varepsilon_x(\widehat{\vartheta}_N) \big)^2 + 2\big((\widehat{z}_N - z^o)' \varepsilon_z(\widehat{\vartheta}_N) \big)^2 \Big) \\
&= \frac{N}{\alpha_N \cdot \frac{1}{N} \sum \varepsilon(\widehat{\vartheta}_N)^2} \Big((\widehat{x}_N - x^o)' \frac{2}{N} \sum \varepsilon_x(\widehat{\vartheta}_N) \varepsilon_x(\widehat{\vartheta}_N)' (\widehat{x}_N - x^o) \\
&\quad +(\widehat{z}_N - z^o)' \frac{2}{N} \sum \varepsilon_z(\widehat{\vartheta}_N) \varepsilon_z(\widehat{\vartheta}_N)' (\widehat{z}_N - z^o) \Big),
\end{aligned}$$

where in the second last step we have used the inequality $a^2 + 2ab + b^2 \leq 2a^2 + 2b^2$.

The term $\frac{1}{N} \sum \varepsilon(\widehat{\vartheta}_N)^2$ converges almost surely to $\lambda^2 = \mathbb{E}[e(t)^2] > 0$ (see [15] and [12]). Thus, all we need to show is that:

$$\frac{N}{\alpha_N \cdot \lambda^2} (\widehat{x}_N - x^o)' \frac{1}{N} \sum \varepsilon_x(\widehat{\vartheta}_N) \varepsilon_x(\widehat{\vartheta}_N)' (\widehat{x}_N - x^o) \to 0 \text{ in probability,} \quad (13)$$

$$\frac{N}{\alpha_N \cdot \lambda^2} (\widehat{z}_N - z^o)' \frac{1}{N} \sum \varepsilon_z(\widehat{\vartheta}_N) \varepsilon_z(\widehat{\vartheta}_N)' (\widehat{z}_N - z^o) \to 0 \text{ in probability.} \quad (14)$$

Let us first prove equation (14).

We first consider the term $\frac{1}{N} \sum \varepsilon_z(\widehat{\vartheta}_N) \varepsilon_z(\widehat{\vartheta}_N)'$ and prove that:

$$\frac{1}{N} \sum \varepsilon_z(\widehat{\vartheta}_N) \varepsilon_z(\widehat{\vartheta}_N)' \to \frac{\overline{V}_{zz}(x^*, z^*)}{2} \quad \text{almost surely} \quad (15)$$

Note that $\frac{1}{N} \sum \varepsilon_z(\widehat{x}_N, \widehat{z}_N) \varepsilon_z(\widehat{x}_N, \widehat{z}_N)' \to \mathbb{E}[\varepsilon_z(x^*, z^*) \varepsilon_z(x^*, z^*)']$ almost surely, as it follows from Assumptions 1, 2, 3 and 4′ (in fact, this result is a consequence of the uniform convergence of empirical means for linear predictors – see [12] and Theorem $2B.1$ in [15]). Thus, all we need to prove is:

$$\overline{V}_{zz}(x^*, z^*) = 2\mathbb{E}[\varepsilon_z(x^*, z^*) \varepsilon_z(x^*, z^*)']. \quad (16)$$

We have that

$$\frac{\mathrm{d}^2}{\mathrm{d}\vartheta^2} \overline{V}(\vartheta^*) = \mathbb{E}[2\psi(\vartheta^*)\psi(\vartheta^*)'] + \mathbb{E}[2\varepsilon(\vartheta^*) \frac{\mathrm{d}}{\mathrm{d}\vartheta} \psi(\vartheta^*)].$$

Lemma 1 says that $\varepsilon(t, \vartheta^*) = e(t)$, which in turn gives ($\psi$ depends on past data only)

$$\mathbb{E}\Big[2\varepsilon(\vartheta^*)\frac{\mathrm{d}}{\mathrm{d}\vartheta}\psi(\vartheta^*)\Big] = 2\mathbb{E}[e(t)]\mathbb{E}\Big[\frac{\mathrm{d}}{\mathrm{d}\vartheta}\psi(\vartheta^*)\Big] = 0.$$

Thus, $\frac{\mathrm{d}^2}{\mathrm{d}\vartheta^2}\overline{V}(\vartheta^*) = 2\mathbb{E}[\psi(\vartheta^*)\psi(\vartheta^*)']$, and, by specializing this latter expression to the z component, we obtain equation (16) which implies equation (15) as we have shown before.

Turn now to consider the term $\sqrt{N}(\widehat{z}_N - z^o)$ and note that it is equal to $\sqrt{N}(\widehat{z}_N - z^*)$ (in fact $z^o = z^*$).

We show that:

$$\sqrt{N}(\widehat{z}_N - z^*) \sim as\mathcal{G}\Big(0, 2\lambda^2\overline{V}_{zz}(x^*, z^*)^{-1}\Big). \tag{17}$$

As a matter of fact, consider the following Taylor expansion (which holds almost surely thanks to the differentiability properties of transfer function coefficients in Assumption 1):

$$\begin{aligned} 0 &= \sqrt{N}\tfrac{\partial}{\partial z}V_N(\widehat{x}_N, \widehat{z}_N) \\ &= \sqrt{N}\tfrac{\partial}{\partial z}V_N(\widehat{x}_N, z^*) \ + \tfrac{\partial^2}{\partial z^2}V_N(\widehat{x}_N, \xi_N)\sqrt{N}(\widehat{z}_N - z^*), \end{aligned} \tag{18}$$

where ξ_N is a point between $\widehat{z}_N$ and z^* and the first equality follows from the fact that $\widehat{\vartheta}_N = [(\widehat{x}_N)' \ (\widehat{z}_N)']'$ is a minimizer of V_N.

Then, we can follow the same rationale as in [15], chapter 9, to conclude that:

- $\sqrt{N}\frac{\partial}{\partial z}V_N(\widehat{x}_N, z^*) \sim as\mathcal{G}\Big(0, 2\lambda^2\overline{V}_{zz}(x^*, z^*)\Big)$ (this results follows along the same lines as Theorem 9.1 in [15])
- $\frac{\partial^2}{\partial z^2}V_N(\widehat{x}_N, \xi_N) \to \overline{V}_{zz}(x^*, z^*)$ almost surely (again, this result follows from Theorem $2B$.1 in [15]).

These two facts imply (17) (see [15], chapter 9, for details).

Equation (14) now follows from (17) and (15). Indeed, the left hand side of (14) can be rewritten as (note that $z^* = z^o$)

$$\Big[\frac{1}{\alpha_N}\Big]\Big[\sqrt{N}(\widehat{z}_N - z^*)'\frac{\frac{1}{N}\sum\varepsilon_z(\widehat{\vartheta}_N)\varepsilon_z(\widehat{\vartheta}_N)'}{\lambda^2}\sqrt{N}(\widehat{z}_N - z^*)\Big] \tag{19}$$

where the first term goes to zero (recall that $\alpha_N \to \infty$) and the second one converges to a χ^2 distributed random variable.

We next prove equation (13).
Note that the proof of (13) is substantially different from the one of (14) since $x^o \neq x^*$ and, in contrast to $\widehat{z}_N - z^o$, $\widehat{x}_N - x^o$ does not tend to zero.
We commence by observing that, since $\overline{V}(x, z^*)$ has a constant value in the x direction – recall that $\{[x' \ (z^*)']', \ x : [x' \ (z^*)']' \in \Theta\}$ is the set of minimizers of $\overline{V}(\vartheta)$ – it holds that $\overline{V}_{xx}(x, z^*) = 0$, $\forall x : [x' \ (z^*)']'$ is an interior point of Θ, and, in particular, $\overline{V}_{xx}(x^*, z^*) = 0$. On the other hand, proceeding as for (15), it can be proved that $\frac{1}{N}\sum \varepsilon_x(\widehat{x}_N, \widehat{z}_N)\varepsilon_x(\widehat{x}_N, \widehat{z}_N)' \to \frac{1}{2}\overline{V}_{xx}(x^*, z^*)$ almost surely, and, thus,

$$\frac{1}{N}\sum \varepsilon_x(\widehat{x}_N, \widehat{z}_N)\varepsilon_x(\widehat{x}_N, \widehat{z}_N)' \to 0 \quad \text{almost surely}$$

This last equation suggests that equation (13) can be proved by characterizing the rate of convergence to 0 of $\frac{1}{N}\sum \varepsilon_x(\widehat{x}_N, \widehat{z}_N)\varepsilon_x(\widehat{x}_N, \widehat{z}_N)'$.
Consider the following Taylor expansion:

$$\begin{aligned}&\Big((\widehat{x}_N - x^o)'\varepsilon_x(\widehat{x}_N, \widehat{z}_N)\Big)^2 \\ &= \Big((\widehat{x}_N - x^o)'\varepsilon_x(\widehat{x}_N, z^*)\Big)^2 + (\widehat{z}_N - z^*)'\frac{\partial}{\partial z}\Big((\widehat{x}_N - x^o)'\varepsilon_x(\widehat{x}_N, z)\Big)^2\Big|_{z=z^*} \\ &\quad + (\widehat{z}_N - z^*)'\frac{\partial^2}{\partial z^2}\Big((\widehat{x}_N - x^o)'\varepsilon_x(\widehat{x}_N, z)\Big)^2\Big|_{z=\zeta_N}(\widehat{z}_N - z^*), \end{aligned} \qquad (20)$$

where ζ_N is a point on the segment connecting $\widehat{z}_N$ and z^*. Derivatives are well defined almost surely thanks to the four times differentiability of the transfer functions coefficients, as required in Assumption 1.
We want to show that the first and second terms in the right hand side of (20) are null so that (20) reduces to

$$\Big((\widehat{x}_N - x^o)'\varepsilon_x(\widehat{x}_N, \widehat{z}_N)\Big)^2 = (\widehat{z}_N - z^*)'\frac{\partial^2}{\partial z^2}\Big((\widehat{x}_N - x^o)'\varepsilon_x(\widehat{x}_N, z)\Big)^2\Big|_{z=\zeta_N}(\widehat{z}_N - z^*). (21)$$

To prove (21), let us start by observing that, similarly to equation (16), it can be proved that $\overline{V}_{xx}(x, z^*) = 2\mathbb{E}\Big[\varepsilon_x(x, z^*)\varepsilon_x(x, z^*)'\Big]$, $\forall x : [x' \ (z^*)']'$ is an interior point of Θ (name X^* such a set of points x). Recalling that $\overline{V}_{xx}(x, z^*) = 0$, $\forall x \in X^*$, we then have $\mathbb{E}\Big[\varepsilon_x(x, z^*)\varepsilon_x(x, z^*)'\Big] = 0$, or equivalently $\mathbb{E}\Big[\|\varepsilon_x(x, z^*)\|\Big] = 0$, $\forall x \in X^*$.
Now, from the latter expression we obtain $0 = \int_{X^*}\mathbb{E}\Big[\|\varepsilon_x(x, z^*)\|\Big]\mathrm{d}x = \mathbb{E}\Big[\int_{X^*}\|\varepsilon_x(x, z^*)\|\mathrm{d}x\Big]$, (the last equality is an application of Fubini's theorem), which entails

$$\int\limits_{X^*}\|\varepsilon_x(x, z^*)\|\mathrm{d}x = 0, \quad \text{almost surely.} \qquad (22)$$

Since $\|\varepsilon_x(x, z^*)\|$ is an almost surely continuous function in x, this finally implies that the following relation holds true almost surely:

$$\varepsilon_x(x, z^*) = 0, \quad \forall x \in X^* \tag{23}$$

(indeed if $\varepsilon_x(x, z^*) \neq 0$ for some $x \in X^*$, by continuity $\int_{X^*} \|\varepsilon_x(x, z^*)\| \mathrm{d}x \neq 0$ which can happen on a zero probability set only – see (22)).
By specializing (23) to $x = \widehat{x}_N$, we have $\varepsilon_x(\widehat{x}_N, z^*) = 0$ almost surely, showing that the first term in the right hand side of (20) is null. The fact that the second term is null too follows by observing that

$$\frac{\partial}{\partial z}\Big((\widehat{x}_N - x^o)'\varepsilon_x(\widehat{x}_N, z)\Big)^2\Big|_{z=z^*} = 2\Big((\widehat{x}_N - x^o)'\varepsilon_x(\widehat{x}_N, z^*)\Big)\Big[\varepsilon_{zx}(\widehat{x}_N, z^*)(\widehat{x}_N - x^o)\Big].$$

This proves (21).
Now, the left hand side of (13) can be rewritten as:

$$\frac{N}{\alpha_N \cdot \lambda^2}\frac{1}{N}\sum\Big((\widehat{x}_N - x^o)'\varepsilon_x(\widehat{\vartheta}_N)\Big)^2,$$

which, using (21), is equal to

$$\Big[\frac{1}{\alpha_N}\Big]\Big[\sqrt{N}(\widehat{z}_N - z^*)'\frac{\frac{1}{N}\sum\frac{\partial^2}{\partial z^2}\Big((\widehat{x}_N - x^o)'\varepsilon_x(\widehat{x}_N, z)\Big)^2\Big|_{z=\zeta_N}}{\lambda^2}\sqrt{N}(\widehat{z}_N - z^*)\Big]. \tag{24}$$

The convergence to zero in probability of (24) now follows similarly to the convergence to zero in probability of (19). As a matter of fact, the only difference between (24) and (19) stays in their kernel, where the kernel of (19) $\frac{\frac{1}{N}\sum\varepsilon_z(\widehat{\vartheta}_N)\varepsilon_z(\widehat{\vartheta}_N)'}{\lambda^2}$ tends almost surely to the positive definite matrix $\frac{\overline{V}_{zz}(x^*,z^*)}{2\lambda^2}$ while the kernel of (24) converges to $\frac{2}{\lambda^2}\mathbb{E}[\varepsilon_{zx}(x^*, z^*)(x^*-x^o)(x^*-x^o)'\varepsilon_{xz}(x^*, z^*)]$ almost surely, as it follows from Theorem $2B.1$ in [15].
This concludes the proof. □

10.2 Proof of Theorem 4

The asymptotic cost criterion can be rewritten through Parseval identity as

$$\overline{V}(\vartheta) = \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{|G(e^{-j\omega}, \vartheta) - G(e^{-j\omega}, \vartheta^o)|^2}{|H(e^{-j\omega}, \vartheta)|^2}F_u(\mathrm{d}\omega) + \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{|H(e^{-j\omega}, \vartheta^o)|^2}{|H(e^{-j\omega}, \vartheta)|^2}\lambda^2 \,\mathrm{d}\omega,$$

where F_u is the spectral measure of $u(t)$.
Let ϑ^* be a minimizer of $\overline{V}(\vartheta)$. Since also ϑ^o minimizes $\overline{V}(\vartheta)$, we have that

$$\overline{V}(\vartheta^*) = \overline{V}(\vartheta^o) = \lambda^2.$$

Thus,

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{|G(e^{-j\omega}, \vartheta^*) - G(e^{-j\omega}, \vartheta^o)|^2}{|H(e^{-j\omega}, \vartheta^*)|^2} F_u(\mathrm{d}\omega) = 0, \tag{25}$$

and

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{|H(e^{-j\omega}, \vartheta^o)|^2}{|H(e^{-j\omega}, \vartheta^*)|^2} \lambda^2 \, \mathrm{d}\omega = \lambda^2. \tag{26}$$

Equation (26) implies that

$$H(e^{-j\omega}, \vartheta^*) = H(e^{-j\omega}, \vartheta^o), \ \ \forall \omega \in [0, \pi]. \tag{27}$$

On the other hand, from equation (25) it follows that $G(e^{-j\omega}, \vartheta^*)$ must be equal to $G(e^{-j\omega}, \vartheta^o)$ at every frequency where $u(t)$ is exciting. That is

$$\begin{aligned} &G(e^{-j\omega}, \vartheta^*) = G(e^{-j\omega}, \vartheta^o), \\ &\qquad \forall \omega : F_u(A) > 0, \ \text{for any open } A \text{ containing } \omega. \end{aligned} \tag{28}$$

Now, letting $H(e^{-j\omega}, \vartheta) = \frac{N_H(e^{-j\omega}, \vartheta)}{D_H(e^{-j\omega}, \vartheta)}$ and $G(e^{-j\omega}, \vartheta) = \frac{N_G(e^{-j\omega}, \vartheta)}{D_G(e^{-j\omega}, \vartheta)}$, equations (27) and (28) can be rewritten as

$$N_H(e^{-j\omega}, \vartheta^*) D_H(e^{-j\omega}, \vartheta^o) = D_H(e^{-j\omega}, \vartheta^*) N_H(e^{-j\omega}, \vartheta^o), \ \ \forall \omega \in [0, \pi], \tag{29}$$

and

$$\begin{aligned} &N_G(e^{-j\omega}, \vartheta^*) D_G(e^{-j\omega}, \vartheta^o) = D_G(e^{-j\omega}, \vartheta^*) N_G(e^{-j\omega}, \vartheta^o), \\ &\qquad \forall \omega : F_u(A) > 0, \ \text{for any open } A \text{ containing } \omega. \end{aligned} \tag{30}$$

For any fixed ω, these equations are linear in ϑ^*, so defining an affine subspace. Since the intersection of affine subspaces is an affine subspace, the set of ϑ^* satisfying equations (29) and (30) is still an affine subspace. This concludes the proof. □

10.3 Proof of Theorems 5

Let us consider the ARMAX case first.
Define:
$G^o(z^{-1}) = \frac{B^o(z^{-1})}{A^o(z^{-1})}$, $H^o(z^{-1}) = \frac{C^o(z^{-1})}{A^o(z^{-1})}$, $G(z^{-1},\vartheta) = \frac{B(z^{-1},\vartheta)}{A(z^{-1},\vartheta)}$ and $H(z^{-1},\vartheta) = \frac{C(z^{-1},\vartheta)}{A(z^{-1},\vartheta)}$, where $A^o(z^{-1})$, $B^o(z^{-1})$ and $C^o(z^{-1})$ stand for $A(z^{-1},\vartheta^o)$, $B(z^{-1},\vartheta^o)$ and $C(z^{-1},\vartheta^o)$, respectively.
Similarly to the proof of Theorem 4, the asymptotic cost criterion can be rewritten through Parseval identity as

$$\overline{V}(\vartheta) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{|G(e^{-j\omega},\vartheta) - G^o(e^{-j\omega})|^2}{|1+R(e^{-j\omega})G^o(e^{-j\omega})|^2} \cdot \frac{|R(e^{-j\omega})|^2}{|H(e^{-j\omega},\vartheta)|^2} F_r(\mathrm{d}\omega) + \\ + \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{|H^o(e^{-j\omega})|^2}{|H(e^{-j\omega},\vartheta)|^2} \cdot \frac{|1+R(e^{-j\omega})G(e^{-j\omega},\vartheta)|^2}{|1+R(e^{-j\omega})G^o(e^{-j\omega})|^2} \lambda^2 \mathrm{d}\omega,$$

where F_r is the spectral measure of $r(t)$.
Now, following the same rationale as in the proof of Theorem 4, we obtain that ϑ^* is a minimizer of $\overline{V}(\vartheta)$ if and only if

$$G(e^{-j\omega},\vartheta^*) - G^o(e^{-j\omega}) = 0, \\ \forall\omega : F_r(A) > 0, \text{ for any open } A \text{ containing } \omega \tag{31}$$

and

$$\frac{H^o(e^{-j\omega})}{H(e^{-j\omega},\vartheta^*)} \cdot \frac{1+R(e^{-j\omega})G(e^{-j\omega},\vartheta^*)}{1+R(e^{-j\omega})G^o(e^{-j\omega})} = 1, \ \forall\omega \in [0,\pi]. \tag{32}$$

Then, by the definition of G, G^o, H and H^o we have that (the dependencies on ϑ and $e^{-j\omega}$ have been omitted to ease the notation)

$$G - G^o = \frac{B}{A} - \frac{B^o}{A^o},$$

and

$$\frac{H^o}{H} \cdot \frac{1+RG}{1+RG^o} = \frac{C^oA}{A^oC} \cdot \frac{D_RA + N_RB}{D_RA} \cdot \frac{D_RA^o}{D_RA^o + N_RB^o} = \frac{C^o}{C} \cdot \frac{D_RA + N_RB}{D_RA^o + N_RB^o},$$

where N_R and D_R are, respectively, the numerator and the denominator of R.
As a consequence, equations (31) and (32) can be rewritten as

$$\begin{aligned}&B(e^{-j\omega},\vartheta^*)A^o(e^{-j\omega}) = B^o(e^{-j\omega})A(e^{-j\omega},\vartheta^*),\\ &\qquad \forall\omega : F_r(A) > 0, \text{ for any open } A \text{ containing } \omega,\end{aligned} \tag{33}$$

and

$$\begin{aligned}&C^o(e^{-j\omega})\Big(D_R(e^{-j\omega})A(e^{-j\omega},\vartheta^*) + N_R(e^{-j\omega})B(e^{-j\omega},\vartheta^*)\Big)\\ &\quad = C(e^{-j\omega},\vartheta^*)\Big(D_R(e^{-j\omega})A^o(e^{-j\omega}) + N_R(e^{-j\omega})B^o(e^{-j\omega})\Big) \;\forall\omega \in [0,\pi].\end{aligned} \tag{34}$$

As in Theorem 4, for any fixed ω these equations are linear in ϑ^* and, therefore, the set of ϑ^* satisfying equations (33) and (34) is an affine subspace.

The same proof applies also for OE models considering $G(z^{-1},\vartheta) = \frac{B(z^{-1},\vartheta)}{A(z^{-1},\vartheta)}$, $G^o(z^{-1}) = \frac{B^o(z^{-1})}{A^o(z^{-1})}$ and $H^o(z^{-1}) = H(z^{-1},\vartheta) = 1$ in this case. □

References

[1] S. Bittanti, G. Picci (eds), *Identification, Adaptation, Learning - The science of learning models from data*, Springer Verlag, 1996

[2] S. Bittanti, M.C. Campi, S. Garatti, *An iterative controller design scheme based on average robust control*, Proc. 15th World IFAC Congress, Barcelona, 2002

[3] M.C. Campi, E. Weyer, *Finite sample properties of system identification methods*, IEEE Trans. on Automatic Control, 2002, vol. 47, 8, pp. 1329-1334

[4] M.C. Campi, E. Weyer and Su Ki Ooi, *Nonasymptotic quality assessment of identified models*, Proc. 15th World IFAC Congress, Barcelona, 2002

[5] M. Gevers, *A decade of progress in iterative control design: from theory to practice*, Proc. Int. Symp. on Advanced Control of Chemical Processes, Pisa, Italy, Vol. II, pp. 677-694 (keynote lecture), June 2000

[6] G.C. Goodwin, M. Gevers, B. Ninness, *Quantifying the error in estimated transfer functions with application to model order selection*, IEEE Trans. Automatic Control, 1992, vol. 37, 7, pp. 913-928

[7] G.C. Goodwin *Identification and robust control: bridging the gap*, Proc. of the 7th IEEE Mediterranian conference on control and automation, 1999, Haifa, Israel

[8] R. Hakvoort, P. M. J. Van den Hof, *Identification of probabilistic system uncertainty by explicit evaluation of bias and variance errors*, IEEE Trans. Automatic Control, 1997, vol. 42, 11, pp. 1516-1528

[9] H. Hjalmarsson, L. Ljung, *Estimating model variance in the case of undermodeling*, IEEE Trans. Automatic Control, 1992, vol. 37, 7, pp. 1004-1008

[10] R. L. Kosut, G. C. Goodwin, M. P. Polis (eds), *Special issue on system identification for robust control design*, IEEE Trans. on Automatic Control, 1992, vol. 37, 7

[11] W. S. Lee, B. D. O. Anderson, R. L. Kosut, I. M. Y. Mareels, *On robust performance improvement through the windsurfer approach to adaptive robust control*, Proc. 32nd IEEE Conf. Decision and Control, San Antonio, TX, 1993, pp. 2821-2827

[12] L. Ljung, *Convergence analysis of parametric identification methods*, IEEE Trans. on Automatic Control, 1978, vol. 23, pp. 779-783

[13] L. Ljung, *Asymptotic variance expressions for identified black-box transfer function models*, IEEE Trans. on Automatic Control, 1985, vol. 30, pp834-844

[14] L. Ljung, L. Guo *The role of model validation for assessing the size of the unmodeled dynamics*, IEEE Trans. on Automatic Control, 1997, vol. 42, 9, pp. 1230-1239

[15] L. Ljung, *System Identification: theory for the user*, Prentice-Hall, Upper Saddle River, NJ, 1999

[16] L. Ljung, *Model validation and model error modeling*, report from the Åström Symposium on Control, Lund, Sweden, 1999

[17] L. Ljung, *Model error modeling and control design*, Proc. of the IFAC symposium SYSID, 2000, Santa Barbara, CA, USA

[18] L. Ljung, *Estimating linear time-invariant models of nonlinear time-varying systems*, European Journal of Control, 2001, 7, pp. 203-219

[19] B. Ninness, G. Goodwin, *Estimation of model quality*, Automatica, 1995, vol. 31, 12, pp. 1771-1795

[20] B. Ninness, H. Hjalmarsson, F. Gustafsson, *The foundamental role of orthonormal bases in system identification*, IEEE Trans. Automatic Control, 1999, vol. 44, 7, pp. 1384-1406

[21] W. Reinelt, A. Garulli, L. Ljung, *Comparing different approaches to model error modeling in robust identification*, Automatica, 2002, 38, pp. 787-803

[22] T. Söderström, P. Stoica, *System Identification*, Prentice-Hall, Englewood Cliffs, NJ, 1989

[23] T. Söderström, K. J. Åström (eds), *Special issue on trends in system identification*, Automatica, 1995, vol.31, 12

[24] F. Tjarnstrom, L. Ljung, *Estimating the variance in case of undermodeling using bootstrap*, Proc. of the 38th Conference on Decision and Control, Phoenix, Arizona, 1999

[25] P. M. J. Van den Hof, R. J. P. Schrama, *Identification and control - closed-loop issues*, Automatica, Vol. 31, 12, pp. 1751-1770, December, 1995

[26] E. Weyer, M.C. Campi, *Non-asymptotic confidence ellipsoids for the least squares estimate*, Automatica, 2002, vol. 38, pp. 1539-1547

[27] L. Xie, L. Ljung, *Asymptotic variance expressions for estimated frequency functions*, IEEE Trans. Automatic Control, 2001, vol. 46, 12, pp. 1887-1899