IDENTIFICATION WITH FINITELY MANY DATA POINTS: THE LSCR APPROACH

Marco C. Campi [*,1] **Erik Weyer** [**,2]

[*] *Dip. di Elettronica per l'Automazione*
Universita' di Brescia
Via Branze 38, 25123 Brescia, Italy
campi@ing.unibs.it
[**] *Dept. of Electrical and Electronic Engineering*
The University of Melbourne
Parkville, VIC 3010, Australia
e.weyer@ee.unimelb.edu.au

Abstract: This paper gives an overview of LSCR (Leave-out Sign-dominant Correlation Regions), a general technique for system identification. Under normal conditions, observations contain information corrupted by disturbances and measurement noise so that only an approximate description of the underlying system can at best be obtained from a finite data set. This is similar to describing an object seen through a frosted glass. Differently from standard identification methods that deliver single models, LSCR generates a model set. As information increases, the model set shrinks around the true system and, for any finite sample size, the set is guaranteed to contain the true system with a precise probability chosen by the user. LSCR only assumes a minimum amount of prior information on the noise.

Keywords: System identification, non-asymptotic results, model set, correlation.

1. INTRODUCTION: FROM NOMINAL-MODEL TO MODEL-SET IDENTIFICATION

System identification is the science of deriving a model from data.

In practical applications, the number of data points an identification procedure can rely on is always finite and, at times, scarce. Nevertheless, most results in identification are of asymptotic nature, that is they tell us what happens when the number of data points tends to infinity. While these results are useful in indicating fundamental qualitative links between the identification set-up and the expected properties of the identified model, when it comes to quantitative assessments, they necessarily have to be used on an approximate basis. Moreover, contributions have appeared showing that the asymptotic theory can produce misleading quantitative results in some system identification endeavors, Garatti et al. (2004,2006).

In this paper, an attempt to bridge this gap is made: We here introduce, review and expand the LSCR (Leave-out Sign-dominant Correlation Regions) approach. LSCR is a system identification methodology that provides rigorous results for any finite data set, no matter how small.

[1] Supported by MIUR (Ministero dell'Istruzione, dell'Universita' e della Ricerca) under the project *New methods for Identification and Adaptive Control for Industrial Systems*
[2] and by the Australian Research Council.

1.1 The need for something more than a nominal model

Suppose we want to identify a system S by selecting a model in a model class $\mathcal{M}$. Even when S belongs to the model class, we cannot expect that the model $\hat{M}$ identified from a finite data sample coincides with S due to a number of accidents affecting our data: measurement noise, presence of disturbances acting on the system, etc.

Example 1. Consider the system

$$y_t = \theta^0 u_t + w_t, \tag{1}$$

where u_t is input, y_t output and w_t is an unmeasured disturbance. Let $u_t = 1$ for all t and identify θ by means of least-squares:

$$\hat{\theta} = \left(\sum_{t=1}^{N} u_t^2\right)^{-1} \left(\sum_{t=1}^{N} u_t y_t\right) = \frac{1}{N}\sum_{t=1}^{N} y_t,$$

where N is the number of data points. The estimation error $\hat{\theta} - \theta^0 = \frac{1}{N}\sum_{t=1}^{N} y_t - \theta^0 = \frac{1}{N}\sum_{t=1}^{N}(\theta^0 + w_t) - \theta^0 = \frac{1}{N}\sum_{t=1}^{N} w_t$ does converge to zero when $N \to \infty$ under natural assumptions on w_t. However, for any finite N, $\frac{1}{N}\sum_{t=1}^{N} w_t$ cannot be expected to be zero and a system-model mismatch is present. □

If a probabilistic description of the uncertain elements is adopted, under general circumstances the only probabilistic claim we are in a position to make is that

$$Pr\{\hat{M} = S\} = 0,$$

clearly a useless statement if our intention is that of crediting the model with reliability. Thus, when an identification procedure only returns a nominal model, we certainly cannot trust it to coincide with the true system and - when this model is used for any purpose - it is done in the hope that the system-model mismatch does not affect the final result too badly. While this way of proceeding is practical, it is not grounded on a solid theoretical basis.

A scientific use of an identified nominal model requires instead that this model be complemented with additional information, information able to certify the model accuracy.

1.2 Model-set identification

A way to obtain certified results is to move from nominal-model to model-set identification. An example explains the idea.

Example 2. (continuation of Example 1). In system (1), suppose that w_t is white and Gaussian with zero mean and unitary variance. Then, $\hat{\theta} - \theta^0 = \frac{1}{N}\sum_{t=1}^{N} w_t$ is a Gaussian random variable with variance equal to $1/N$ and we can easily attach a probability to the event $\{|\hat{\theta} - \theta^0| \leq \gamma\}$ for any given finite N by looking up a Gaussian table. The practical use of this result is that - after estimating $\hat{\theta}$ from data - an interval $[\hat{\theta} - \gamma, \hat{\theta} + \gamma]$ can be computed having guaranteed probability to contain the true parameter value. □

In more general terms, the situation can be depicted as in Figure 1: Given any finite N, the parameter estimate $\hat{\theta}$ is affected by random fluctuation, so that it has probability zero to exactly hit the system parameter value. Considering a region around $\hat{\theta}$, however, elevates the probability that θ^0 belongs to the region to a nonzero - and therefore meaningful - value.

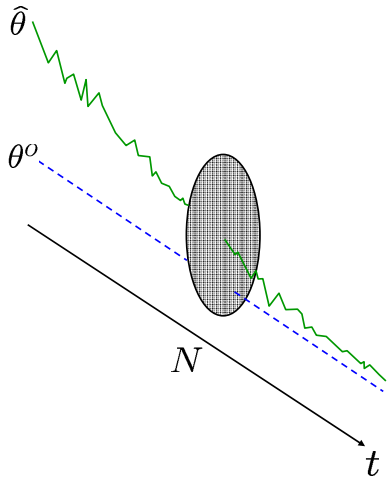

Fig. 1. Random fluctuation in parameter estimate.

This observation points to a simple, but important fact:

– exploiting the information conveyed by a finite number of data points can in standard situations at best generate a model set to which the system belong. Any further attempt to provide sharper estimation results (e.g. one single model) goes beyond the available information level and generates results that cannot be guaranteed.

As a consequence, considering identification procedures returning parameter sets - as opposed to single parameter values - is a sensible approach in the attempt to provide guaranteed results.

Having made this conceptual point, it is also important to observe that this in no way deny the importance of nominal models: Nominal models do play an important practical role in many areas such as simulation, prediction and control. Thus,

constructing nominal models is often a significant step towards finding a solution of a problem. Yet, in order to judge the quality of the solution one should also account for uncertainty, as given by a model uncertainty set.

1.3 The role of a-priori information

Any system identification process is based on two sources of information:

- **(i)** a-priori information, that is we a-priori know that S belongs to a system set $\mathcal{S}$;
- **(ii)** a-posteriori information, that is data collected from the system.

Without any a-priori information we can hardly generate any guaranteed result. In Example 2, we assumed quite a bit of a-priori information: The system structure was known; and the fact that the noise was Gaussian with variance 1 was exploited in the construction of the confidence region.

Having said that a-priori information cannot be totally renounced, it is also important to point out that a good theory should demand as little prior as possible. Indeed:

- stringent prior conditions reduce the theory's applicability;
- in real applications, stringent prior conditions may be difficult to verify even when they are actually satisfied.

Going back to Example 2, let us ask the following question: Can we reduce prior knowledge and still provide guaranteed results? Some answer is provided in the following example continuation.

Example 3. (continuation of Example 2). Suppose now that the variance of w_t is σ^2 and that it is unknown and no upper bound on its value is available. We still want to quantify the probability of the event $\{|\hat{\theta} - \theta^0| \leq \gamma\}$ and we also want that this probability to be guaranteed for any situation allowed by our a-priori knowledge, that is for any value of σ^2.

Since $\hat{\theta} - \theta^0$ is Gaussian with variance σ^2/N, for any given $\gamma > 0$ (even very large), $\sup_{\sigma^2} Pr\{|\hat{\theta} - \theta^0| \leq \gamma\} = 0$, so that the only statement valid for all possible σ^2 is

$$Pr\{|\hat{\theta} - \theta^0| \leq \gamma\} \geq 0,$$

which is evidently a void statement.

A natural question is: Is the situation hopeless and do we have to give up on finding guaranteed results or can we attempt some other approach? To answer, let us try to put the problem under a different light: A well known fact from statistics is that

$$\frac{\hat{\theta} - \theta^0}{\sqrt{\frac{1}{N(N-1)} \sum_{t=1}^{N} (y_t - \hat{\theta})^2}}$$

has a Student t-distribution (Richmond (1964)) with $N-1$ degrees of freedom, independently of the value of σ^2. Thus, given a γ, using tables of the Student t-distribution one can determine a δ such that

$$Pr\left\{ \frac{|\hat{\theta} - \theta^0|}{\sqrt{\frac{1}{N(N-1)} \sum_{t=1}^{N} (y_t - \hat{\theta})^2}} \leq \gamma \right\} = \delta,$$

where the important fact is that the result holds no matter what σ^2 is. Hence:

$$Pr\left\{ |\hat{\theta} - \theta^0| \leq \gamma(y) \right\} = \delta \tag{2}$$

with

$$\gamma(y) = \gamma \sqrt{\frac{1}{N(N-1)} \sum_{t=1}^{N} (y_t - \hat{\theta})^2}.$$

The latter is the desired accuracy evaluation result: One selects a γ and computes the corresponding accuracy variable $\gamma(y)$ using data. The table of the Student t-distribution is then used to find the probability δ of the confidence interval. □

Two remarks on Example 3 are in order:

- **(1)** If σ^2 is unknown, we have seen that no accuracy result - save the void one - can be established for a fixed deterministic γ. Yet, by allowing γ to be data dependent (i.e. by substituting γ with $\gamma(y)$) a meaningful conclusion like (2) can be derived. This fact tells us that we need to let the data speak and the uncertainty set size has to depend on observed data.
- **(2)** The random variable $\frac{\hat{\theta} - \theta^0}{\sqrt{\frac{1}{N(N-1)} \sum_{t=1}^{N} (y_t - \hat{\theta})^2}}$ is a function of data, and the distribution of the data themselves depends on the noise variance σ^2. Despite this, the distribution of $\frac{\hat{\theta} - \theta^0}{\sqrt{\frac{1}{N(N-1)} \sum_{t=1}^{N} (y_t - \hat{\theta})^2}}$ is independent of σ^2. In the statistical literature, this is called a 'pivotal' variable because its distribution does not depend on the variable elements of the problem. The existence of a pivotal variable is crucial in this example to establish a result that is guaranteed for any σ^2.

Unfortunately, finding pivotal variables is generally hard even for very simple examples:

Example 4. Consider now the autoregressive system

$$y_t + \theta^0 y_{t-1} = w_t,$$

where again w_t is zero-mean Gaussian with unknown variance σ^2. Finding a pivotal variable for this situation is already a very hard problem. □

Thus, the approach outlined in Example 3 does not appear to be easy to generalize. Nevertheless, the concept of 'pivotal' distribution - or, more generally, of 'pivotal' result - has a general appeal and we shall come back to it at a later stage in this paper.

1.4 Content of the present paper

In this paper, we provide an overview of LSCR as a methodology to address the above described problems of determining guaranteed model sets in system identification. LSCR delivers data-based confidence sets that contain the true parameter value with guaranteed probability for any finite data sample and it requires minimal knowledge on the noise affecting the system.

The main idea behind the LSCR method is to compute empirical correlation functions and to leave out those regions in parameter space where the correlation functions take on positive or negative values too many times. This principle, which is the reason for he name of the method, is based on the fact that for the true parameter value the correlation functions are sums of zero mean random variables and, therefore, it is unlikely that nearly all of them will be positive or nearly all of them will be negative.

Part I of the paper deals with the case where the system S belongs to the model class $\mathcal{M}$.

In many cases, however, the structure of S is only partially known, and - even when S is known to belong to a certain class $\mathcal{S}$ - we at times deliberately look for a model in a restricted model class $\mathcal{M}$ because $\mathcal{S}$ is too large to work with. In these cases, asking for a probability that $S \in \mathcal{M}$ looses any meaning and we should instead ask whether $\mathcal{M}$ contains a suitable approximation or 'projection' of S. Part II of this paper contains results for systems with unmodeled dynamics.

Further generalizations to a nonlinear set-up are given in Part III.

The presentation is mainly based on examples to help readability, while general results are only sketched.

It goes without saying that the perspective of this paper of reviewing LSCR is a matter of choice and other techniques exist to deal with finite sample identification. Without any claim or attempt of completeness, some of these techniques are briefly mentioned in the next section.

2. A SHORT LITERATURE REVIEW OF FINITE SAMPLE SYSTEM IDENTIFICATION

The pioneering work of Vapnik and Chervonenkis (1968,1971) provided the foundations of what became the field of statistical learning theory, e.g. Vapnik (1998), Cherkassky and Mulier (1998), Vidyasagar (2002). By using exponential inequalities such as the Hoeffding inequality (Hoeffding (1963)) and combinatorial arguments, they derived rigorous uniform probabilistic bounds on the difference between expected values and empirical means computed using a finite number of data points. An underlying assumption was that the data points were independent and identically distributed, an assumption not often satisfied in a system identification setting. Since the 1990s results have appeared which extend the uniform convergence results to dependent data sequences such as M-dependent sequences, and α-, β- and ψ-mixing sequences, e.g. Yu (1994), Bosq (1998), Karandikar and Vidyasagar (2002). These ideas were applied to problems in identification and prediction developing non-asymptotic bounds on the estimation and prediction accuracies. See e.g. Mohda and Masry (1996,1998), Campi and Kumar (1998), Goldenshluger (1998), Weyer et al. (1999), Weyer (2000), Meir (2000), Venkatesh and Dahleh (2001), Campi and Weyer (2002), Weyer and Campi (2002), Vidyasagar and Karandikar (2002,2006), Bartlett (2003).

The finite sample results with roots in learning theory gave bounds on the difference between expected and empirical means, and the bounds were depending on the number of data points, but not on the actually seen data. For this reason they could be quite conservative for the particular system at hand. A way around conservatism is to make active use of the data and construct the bounds using data coming from the system under investigation. Data based methods for evaluation of model quality using bootstrap and subsampling (e.g. Efron and Tibshirani (1993), Shao and Tu (1995), Politis (1998), Politis et al. (1999)) have been explored in Tjärnström and Ljung (2002), Bittanti and Lovera (2000) and Dunstan and Bitmead (2003). However, few truly rigorous finite sample results for model quality assessment are available for these techniques. Using subsampling techniques, Campi et al. (2004) obtained some guaranteed results for generalised FIR systems. See also Hjalmarsson and Ninness (2004) and Ninness and Hjalmarsson (2004) for non-asymptotic

variance expressions for frequency function estimates.

Data based finite sample results have also been derived in the context of model validation by Smith and Dullerud (1996), and the set membership approach to system identification, e.g. Milanese and Vicino (1991), Vicino and Zappa (1996), Giarre' et al. (1997a,b), Garulli et al. (2000, 2002), Milanese and Taragna (2005).

Along a different line of thought, finite sample properties of worst case identification in deterministic frameworks were studied in Dahleh et al. (1993,1995), Poolla and Tikku (1994) and Harrison et al. (1996). Spall (1995) considered uncertainty bounds for M-estimators with small sample sizes. Non-parametric identification methods with finite number of data points have been studied by Welsh and Goodwin (2002) (see also Heath (2001)), while Ding and Chen (2005) have studied the recursive least squares method in a finite sample context.

The LSCR method presented in this paper is making use of subsampling techniques, and in particular it extends the results of Hartigan (1969,1970) to a dynamical setting. Loosely speaking, one could view LSCR as a stochastic set membership approach to system identification, where the setting we consider is the standard stochastic setting for system identification from e.g. Ljung (1999) or Söderström and Stoica (1988), but where the outcomes are sets of models as in set membership identification.

PART I: Known system structure

3. LSCR: A PRELIMINARY EXAMPLE

We start with a preliminary example that readily provides some insight in the LSCR technique. More general results and comments are given in the next section.

Consider again the system of Example 4:

$$y_t + \theta^0 y_{t-1} = w_t. \tag{3}$$

Assume we know that w_t is an independent process and that it has a symmetric distribution around zero. Apart from this, no knowledge on the noise is assumed: It can have any (unknown) distribution: Gaussian; uniform; flat with small-area spikes at high-value locations describing the chance of outliers; etc. Its variance can be any (unknown) number, from very small to very large. We do not even make any stationarity assumption on w_t and allow that its distribution varies with time. The assumption that w_t is independent can be interpreted by saying that we know the system structure: It is an autoregressive system of order 1.

9 data points were generated according to (3) and they are shown in Figure 2. The values of θ^0 and w_t are for the moment not revealed to the reader (they will be disclosed later). Our goal is to form a confidence region for θ^0 from the available data set.

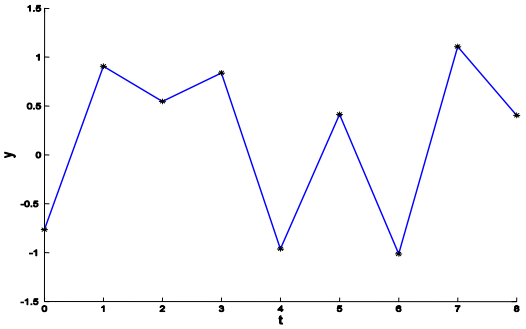

Fig. 2. Data for the preliminary example.

We next adopt a user's point of view and describe the procedure in order to solve this problem with LSCR. Later on comments regarding the obtained results will be provided.

Rewrite the system as a model with generic parameter θ:

$$y_t + \theta y_{t-1} = w_t.$$

The predictor and prediction error associated with the model are

$$\hat{y}_t(\theta) = -\theta y_{t-1}, \quad \epsilon_t(\theta) = y_t - \hat{y}_t(\theta) = y_t + \theta y_{t-1}.$$

Next we compute the prediction errors $\epsilon_t(\theta)$ for $t = 1, \ldots, 8$, and calculate

$$f_{t-1}(\theta) = \epsilon_{t-1}(\theta)\epsilon_t(\theta), \quad t = 2, \ldots, 8.$$

Note that, $f_{t-1}(\theta)$ are functions of θ that can indeed be computed from the available data set. Then, we take the average of some of these functions in many different ways. Precisely, we form 8 averages of the form:

$$g_i(\theta) = \frac{1}{4} \sum_{k \in I_i} f_k(\theta), \quad i = 1, \ldots, 8, \tag{4}$$

where the sets I_i are subsets of $\{1, \ldots, 7\}$ containing the elements highlighted by a bullet in the table below. For instance: $I_1 = \{1, 2, 4, 5\}$, $I_2 = \{1, 3, 4, 6\}$, etc.. The last set, I_8, is an exceptional set: It is empty and we let $g_8(\theta) = 0$. The functions $g_i(\theta)$, $i = 1, \ldots, 7$, can be interpreted

as empirical 1-step correlations of the prediction error.

	1	2	3	4	5	6	7
I_1	•	•		•	•		
I_2	•		•	•		•	
I_3		•	•		•	•	
I_4	•	•				•	•
I_5	•		•		•		•
I_6		•	•	•			•
I_7				•	•	•	•
I_8							

The functions $g_i(\theta)$, $i = 1, \ldots, 7$, obtained for the data in Figure 2 are displayed in Figure 3.

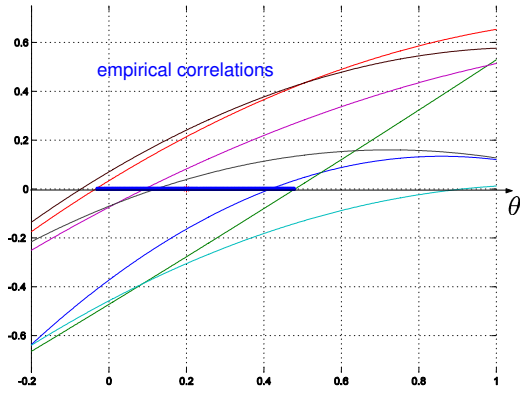

Fig. 3. The $g_i(\theta)$ functions.

Now, a simple reasoning leads us to conclude that these $g_i(\theta)$ functions have a tendency to intersect the θ-axis near θ^0 and that, for $\theta = \theta^0$, they take on positive or negative value with equal probability. Why is it so? Let us re-write one of these functions, say $g_1(\theta)$, as follows:

$$
\begin{aligned}
g_1(\theta) &= \frac{1}{4} \sum_{k\in\{1,2,4,5\}} [y_k + \theta y_{k-1}][y_{k+1} + \theta y_k] \\
&= \frac{1}{4} \sum_{k\in\{1,2,4,5\}} [(y_k + \theta^0 y_{k-1}) + (\theta - \theta^0) y_{k-1}] \\
&\qquad \times [(y_{k+1} + \theta^0 y_k) + (\theta - \theta^0) y_k] \\
&= \frac{1}{4} \sum_{k\in\{1,2,4,5\}} [w_k + (\theta - \theta^0) y_{k-1}] \\
&\qquad \times [w_{k+1} + (\theta - \theta^0) y_k] \\
&= (\theta - \theta^0)^2 \frac{1}{4} \sum_{k\in\{1,2,4,5\}} y_{k-1} y_k \\
&\quad + (\theta - \theta^0) \frac{1}{4} \sum_{k\in\{1,2,4,5\}} w_k y_k \\
&\quad + (\theta - \theta^0) \frac{1}{4} \sum_{k\in\{1,2,4,5\}} y_{k-1} w_{k+1} \\
&\quad + \frac{1}{4} \sum_{k\in\{1,2,4,5\}} w_k w_{k+1}.
\end{aligned}
$$

If $\frac{1}{4}\sum_{k\in\{1,2,4,5\}} w_k w_{k+1} = 0$, the intersection with the θ-axis is obtained for $\theta = \theta^0$. The vertical 'displacement' $\frac{1}{4}\sum_{k\in\{1,2,4,5\}} w_{k-1} w_k$ is a random variable with equal probability of being positive or negative; moreover, due to averaging, vertical dispersion caused by noise is de-emphasized (it would be more de-emphasized with more data). So, we would like to claim that θ^0 will be 'somewhere' near where the average functions intersect the θ-axis. Moreover - following the above reasoning - we recognize that for $\theta = \theta^0$ it is very unlikely that almost all the $g_i(\theta)$, $i = 1, \ldots, 7$, functions have the same sign, and we therefore discard the rightmost and leftmost regions where at most one function is less than zero or greater than zero. The resulting interval $[-0.04, 0.48]$ is the confidence region for θ^0.

A deeper reasoning (given in Appendix A for not breaking the continuity of discourse here) reveals that a fairly strong claim on the obtained interval can be made:

> **RESULT:** the confidence region constructed this way has exact probability $1 - 2 \cdot 2/8 = 0.5$ to contain the true parameter value θ^0.

A few comments are in order:

(1) The interval is stochastic because it depends on data; the true parameter value θ^0 is not and it has a fixed location that does not depend on any random element. Thus, what the above RESULT says is that the interval is subject to random fluctuation and covers the true parameter value θ^0 in 50% of the cases.

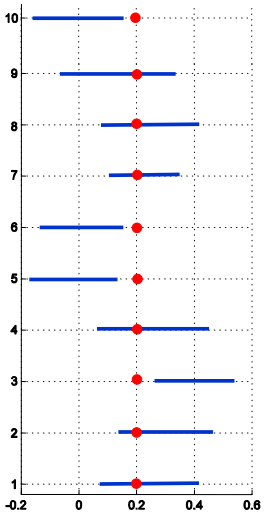

Fig. 4. 10 more trials.

To better understand the nature of the result, we performed 10 more simulation trials obtaining the results in Figure 4. Note that θ^0 and w_t were as follows: $\theta^0 = 0.2$, w_t inde-

pendent with uniform distribution between -1 and $+1$.

(2) In this example, probability is low (50%) and the interval is rather large. With more data, we obtain smaller intervals with higher probability (see also the next section).

(3) The LSCR algorithm was applied with no knowledge on the noise level or distribution and, yet, it returned an interval whose probability was exact, not an upper bound. What is the key here is that the above RESULT is a 'pivotal' result as the probability remains the same no matter what the noise characteristics are.

(4) The result was established along a totally different inference principle from standard Prediction Error Minimization (PEM) methods. In particular - differently from the asymptotic theory of PEM - LSCR does not construct the confidence region by quantifying the variability in an estimate.

(5) We also mention a technical aspect: The pivotal RESULT holds because $\{I_i, i = 1, \ldots, 8\}$ form a group under the symmetric difference operation, that is $(I_j \cup I_k) - (I_j \cap I_k)$ returns another set in $\{I_i, i = 1, \ldots, 8\}$ for any j and k. For instance, $(I_1 \cup I_2) - (I_1 \cap I_2) = I_3$.

4. LSCR FOR GENERAL LINEAR SYSTEMS

4.1 Data generating system

Consider now the general linear system in Figure 5.

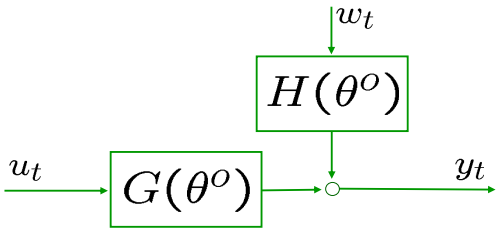

Fig. 5. The system.

We assume that w_t and u_t are independent processes. This does not mean however that we are confining ourselves to an open-loop configuration since closed-loop systems can be reduced to the set-up in Figure 5 by regarding w_t and u_t as external signals, see Figure 6.

$G(\theta^0)$ and $H(\theta^0)$ are stable rational transfer functions. $H(\theta^0)$ is monic and has a stable inverse. w_t is a zero-mean independent sequence (noise). No a-priori knowledge of the noise level is assumed.

The basic assumption we make is that the system structure is known and, correspondingly, we take a full-order model class of the form:

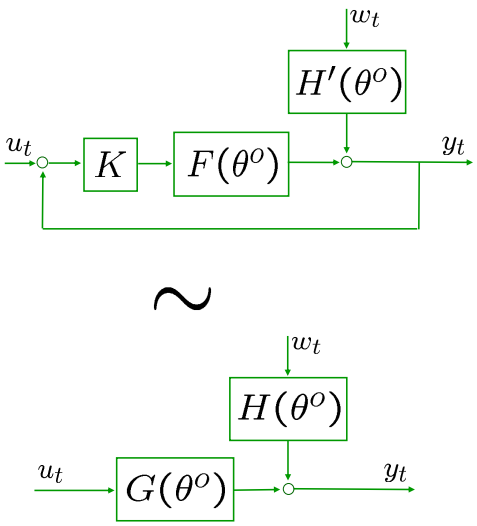

Fig. 6. Closed-loop recast as open-loop.

$$y_t = G(\theta)u_t + H(\theta)w_t,$$

such that the true transfer functions $G(\theta^0)$ and $H(\theta^0)$ are obtained for $\theta = \theta^0$ and for no other parameter than this. We assume that θ is restricted to a set Θ such that $H(\theta)$ is monic and $G(\theta)$, $H(\theta)$ and $H^{-1}(\theta)$ are stable for all $\theta \in \Theta$.

Our goal is to construct an algorithm that works in the following way (see Figure 7): A finite input-output data set is given to the algorithm together with a probability $\bar{p}$. The algorithm is required to return a confidence region that contains the true θ^0 with probability $\bar{p}$ under assumptions on the noise as general as possible.

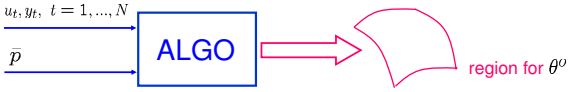

Fig. 7. The algorithm.

4.2 Construction of confidence regions

We start by describing procedures for the determination of confidence sets Θ_r^ϵ based on correlation properties of ϵ (the prediction error) at different time instants (this generalizes the preliminary example in Section 3) and of confidence sets Θ_s^u based on cross-correlation properties of ϵ and u. In general, confidence regions $\hat{\Theta}$ for θ^0 can be constructed by taking the intersection of a few of the Θ_r^ϵ and Θ_s^u sets and this is discussed at the end of this section.

Procedure for the construction of Θ_r^ϵ

(1) Compute the prediction errors

$$\epsilon_t(\theta) = y_t - \hat{y}_t(\theta) = H^{-1}(\theta)y_t - H^{-1}(\theta)G(\theta)u_t$$

for a finite number of values of t, say $t = 1, 2, \ldots, K$;

(2) Select an integer $r \geq 1$. For $t = 1+r, \ldots, N+r = K$, compute

$$f^{\epsilon}_{t-r,r}(\theta) = \epsilon_{t-r}(\theta)\epsilon_t(\theta);$$

(3) Let $I = \{1, \ldots, N\}$ and consider a collection G of subsets $I_i \subseteq I$, $i = 1, \ldots, M$, forming a group under the symmetric difference operation (i.e. $(I_i \cup I_j) - (I_i \cap I_j) \in G$, if $I_i, I_j \in G$). Compute

$$g^{\epsilon}_{i,r}(\theta) = \sum_{k \in I_i} f^{\epsilon}_{k,r}(\theta), \quad i = 1, \ldots, M;$$

(4) Select an integer q in the interval $[1, (M+1)/2)$ and find the region Θ^{ϵ}_r such that at least q of the $g^{\epsilon}_{i,r}(\theta)$ functions are bigger than zero and at least q are smaller than zero. □

The above procedure is the same as the one used for construction of the confidence set in the preliminary example in Section 3. In that example we had $H^{-1}(\theta) = 1 + \theta z^{-1}$, $G(\theta) = 0$, and $K = 8, N = 7, r = 1, M = 8$ and $q = 2$. Normalization $\frac{1}{4}$ in the preliminary example was introduced for the purpose of interpreting the $g_i(\theta)$ functions as empirical averages but it could have been dropped similarly to point 3 in the above procedure without affecting the final result.

In the procedure, the group G can be freely selected. Thus, if e.g. $I = \{1, 2, 3, 4\}$, a suitable group is $G = \{\{1, 2\}, \{3, 4\}, \emptyset, \{1, 2, 3, 4\}\}$; another one is $G = \{\{1\}, \{2, 3, 4\}, \emptyset, \{1, 2, 3, 4\}\}$; yet another one is $G =$ all subsets of I. While the theory presented holds for any choice and the region Θ^{ϵ}_r is guaranteed to be a confidence region in any case (see Theorem 1 below), the feasible choices are limited by computational considerations. For example, the set of all subsets cannot be normally chosen as it is a truly large set. Gordon (1974) discusses how to construct groups of moderate size where the subsets contain approximately half of the elements in I and such a procedure is also summarized in Appendix B. These sets are particularly well suited for use in point 3 of the above procedure.

The intuitive idea behind the construction in the procedure is that, for $\theta = \theta^0$, the functions $g^{\epsilon}_{i,r}(\theta)$ assume positive or negative value at random, so that it is unlikely that almost all of them are positive or that almost all of them are negative. Since point 4 in the construction of Θ^{ϵ}_r discards regions where all $g^{\epsilon}_{i,r}(\theta)$'s but a small fraction (q should be taken to be small compared to M) are of the same sign, we expect that $\theta^0 \in \Theta^{\epsilon}_r$ with high probability. This is put on solid mathematical grounds in Theorem 1 below, showing that the probability that $\theta^0 \in \Theta^{\epsilon}_r$ is actually $1 - 2q/M$. Thus, q is a tuning parameter that has to be selected such that a desired probability of the confidence region is obtained. Moreover, as q increases, we exclude larger and larger regions of Θ and hence Θ^{ϵ}_r shrinks and the probability that $\theta^0 \in \Theta^{\epsilon}_r$ decreases.

The procedure for construction of the sets Θ^u_s is in the same spirit. The only difference being that the empirical auto-correlations in point 2 are replaced by empirical cross-correlations between the input signal and the prediction error.

Procedure for the construction of $\Theta^{\mathbf{u}}_{\mathbf{s}}$

(1) Compute the prediction errors

$$\epsilon_t(\theta) = y_t - \hat{y}_t(\theta) = H^{-1}(\theta)y_t - H^{-1}(\theta)G(\theta)u_t$$

for a finite number of values of t, say $t = 1, 2, \ldots, K$;

(2) Select an integer $s \geq 0$. For $t = 1+s, \ldots, N+s = K$, compute

$$f^u_{t-s,s}(\theta) = u_{t-s}\epsilon_t(\theta);$$

(3) Let $I = \{1, \ldots, N\}$ and consider a collection G of subsets $I_i \subseteq I$, $i = 1, \ldots, M$, forming a group under the symmetric difference operation. Compute

$$g^u_{i,s}(\theta) = \sum_{k \in I_i} f^u_{k,s}(\theta), \quad i = 1, \ldots, M;$$

(4) Select an integer q in the interval $[1, (M+1)/2)$ and find the region Θ^u_s such that at least q of the $g^u_{i,s}(\theta)$ functions are bigger than zero and at least q are smaller than zero. □

The next theorem gives the exact probability that the true parameter θ^0 belongs to one particular of the above constructed sets. The proof of this theorem - as well as comments on the technical assumption on densities - can be found in Campi and Weyer (2005).

Theorem 1. Assume that the variables w_t and $w_t u_\tau$ admit densities and that w_t is symmetrically distributed around zero. Then, the sets Θ^{ϵ}_r and Θ^u_s constructed above are such that:

$$Pr\{\theta^0 \in \Theta^{\epsilon}_r\} = 1 - 2q/M, \tag{5}$$

$$Pr\{\theta^0 \in \Theta^u_s\} = 1 - 2q/M. \tag{6}$$

The following comments pinpoint some important aspects of this result:

(1) The procedures return regions of guaranteed probability despite that no a-priori knowledge on the noise level is assumed: The noise level enters the procedures through data only. This could be phrased by saying that the procedures let the data speak, without a-priori assuming what they have to tell us.

(2) As expected, noise level does impact the final result as the shape and size of the region depend on noise via the data.

(3) Evaluations (5) and (6) are nonconservative in the sense that $1 - 2q/M$ is the exact probability, not a lower bound of it.

Each one of the sets Θ_r^ϵ and Θ_s^u is a non-asymptotic confidence set for θ^0. However, each one of these sets is based on one correlation only and will usually be unbounded in some directions of the parameter space, and therefore not particularly useful. A general practically useful confidence set $\hat{\Theta}$ can be obtained by intersecting a number of the sets Θ_r^ϵ and Θ_s^u, i.e.

$$\hat{\Theta} = \cap_{r=1}^{n_\epsilon} \Theta_r^\epsilon \cap_{s=1}^{n_u} \Theta_s^u. \tag{7}$$

An obvious question is how to choose n_ϵ and n_u in order to obtained well shaped confidence sets that are bounded and concentrated around the true parameter θ^0. It turns out that the answer depends on the particular model class under consideration and this issue will be further discussed in Section 6.

We conclude this section with a fact which is immediate from Theorem 1.

Theorem 2. Under the assumptions of Theorem 1,

$$Pr\{\theta^0 \in \hat{\Theta}\} \geq 1 - (n_\epsilon + n_u)2q/M,$$

where $\hat{\Theta}$ is given by (7).

The inequality in the theorem is due to that the events $\{\theta^0 \notin \Theta_r^\epsilon\}$, $\{\theta^0 \notin \Theta_s^u\}$, $r = 1, \ldots, n_\epsilon$, $s = 1, \ldots, n_u$, may be overlapping.

Theorem 2 can be used in connection with robust design procedures: If a problem solution is robust with respect to $\hat{\Theta}$ in the sense that a certain property is achieved for any $\theta \in \hat{\Theta}$, then such a property is also guaranteed for the true system with the selected probability $1 - (n_\epsilon + n_u)2q/M$.

5. EXAMPLES

Two examples illustrate the developed methodology. The first one is simple and permits an easy illustration of the method. The second is more challenging.

5.1 First order ARMA system

Consider the ARMA system

$$y_t + a^0 y_{t-1} = w_t + c^0 w_{t-1}, \tag{8}$$

where $a^0 = -0.5, c^0 = 0.2$ and w_t is an independent sequence of zero mean Gaussian random variables with variance 1. 1025 data points were generated according to (8). As a model class we used $y_t + ay_{t-1} = w_t + cw_{t-1}$, $|a| < 1$, $|c| < 1$, with associated predictor and prediction error given by

$$\hat{y}_t(a,c) = -c\hat{y}_{t-1}(a,c) + (c-a)y_{t-1},$$
$$\epsilon_t(a,c) = y_t - \hat{y}_t(a,c) = y_t + ay_{t-1} - c\epsilon_{t-1}(a,c).$$

In order to form a confidence region for $\theta^0 = (a^0, c^0)$ we calculated

$$f_{t-1,1}^\epsilon(a,c) = \epsilon_{t-1}(a,c)\epsilon_t(a,c), \quad t = 2, \ldots, 1024,$$
$$f_{t-2,2}^\epsilon(a,c) = \epsilon_{t-2}(a,c)\epsilon_t(a,c), \quad t = 3, \ldots, 1025,$$

and then computed

$$g_{i,1}^\epsilon(a,c) = \sum_{k \in I_i} f_{k,1}^\epsilon(a,c), \quad i = 1, \ldots, 1024,$$
$$g_{i,2}^\epsilon(a,c) = \sum_{k \in I_i} f_{k,2}^\epsilon(a,c), \quad i = 1, \ldots, 1024,$$

using the group in Appendix B. Next we discarded those values of a and c for which zero was among the 12 largest and smallest values of $g_{i,1}^\epsilon(a,c)$ and $g_{i,2}^\epsilon(a,c)$. Then, according to Theorem 2, (a^0, c^0) belongs to the constructed region with probability at least $1 - 2 \cdot 2 \cdot 12/1024 = 0.9531$. The obtained confidence region is the blank area in Figure 8. The area marked with `x` is where 0 is among the 12 smallest values of $g_{i,1}^\epsilon$, the area marked with `+` is where 0 is among the 12 largest values of $g_{i,1}^\epsilon$. Likewise for $g_{i,2}^\epsilon$ with the squares representing when 0 belongs to the 12 largest elements and the circles the 12 smallest. The true value (a^0, c^0) is marked with a star. As we can see, each step in the construction of the confidence region excludes a particular region.

Using the algorithm for the construction of $\hat{\Theta}$ we have obtained a bounded confidence set with a guaranteed probability based on a finite number of data points. As no asymptotic theory is involved this is a rigorous finite sample result. For comparison, we have in Figure 8 also plotted the 95% confidence ellipsoid obtained using the asymptotic theory (Ljung (1999), Chapter 9). The two confidence regions are of similar shape and size, confirming that the non-asymptotic confidence sets are practically useful, and - unlike the asymptotic confidence ellipsoids - they do have guaranteed probability for a finite sample size.

5.2 A closed-loop system

The following example was originally introduced in Garatti et al. (2004) to demonstrate that the

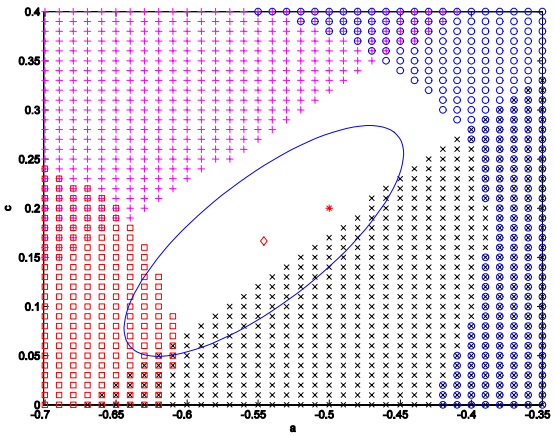

Fig. 8. Non-asymptotic confidence region for (a^0, c^0) (blank region) and asymptotic confidence ellipsoid. $\star$ = true parameter, $\diamond$ = estimated parameter using a prediction error method.

asymptotic theory of PEM can at times deliver misleading results even with a large amount of data points. It is reconsidered here to show how LSCR works in this challenging situation.

Consider the system of Figure 9 where

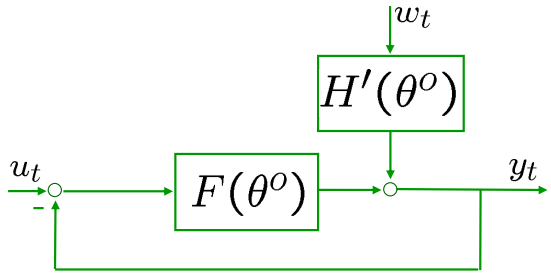

Fig. 9. The closed-loop system.

$$F(\theta^0) = \frac{b^0 z^{-1}}{1 + a^0 z^{-1}}, \qquad a^0 = -0.7, \quad b^0 = 0.3$$
$$H'(\theta^0) = 1 + h^0 z^{-1}, \qquad h^0 = 0.5,$$

w_t is white Gaussian noise with variance 1 and the reference u_t is also white Gaussian, with variance 10^{-6}. Note that the variance of the reference signal is very small as compared to the noise variance, that is there is poor excitation. It is perhaps interesting to note that the present situation - though admittedly artificial - is a simplification of what often happens in practical identification, where poor excitation is due to the closed-loop operation of the system. 2050 measurements of u and y were generated to be used in identification.

We first describe what we obtained using PEM identification.

A full order model was identified. The amplitude Bode diagrams of the transfer function from u to y of the identified model and of the real system are plotted in Figure 10. From the plot, a big mismatch between the real plant and the identified model is apparent, a fact that does not come too much of a surprise considering that the reference signal is poorly exciting. An analysis conducted in Garatti et al. (2004) shows that, when $u_t = 0$, the asymptotic PEM identification cost has two isolated global minimizers, one is θ^0 and a second one is a spurious parameter, say θ^*; when $u_t \neq 0$ but small as is the case in our actual experiment, θ^* does not minimize the asymptotic cost anymore, but random fluctuations in the identification cost due to the finiteness of the data points may as well result in that the estimate gets trapped near the spurious θ^*, generating a totally wrong identified model.

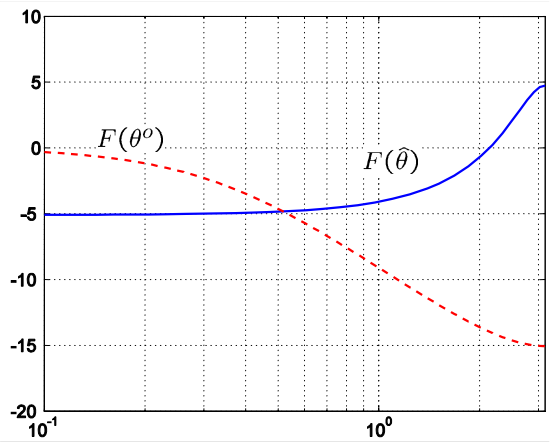

Fig. 10. The identified $u - y$ transfer function.

But, let us now see what we obtained as a 90% confidence region with the asymptotic theory. Figure 11 displays the confidence region in the frequency domain: Surprisingly, it concentrates around the identified model, so that in a real identification procedure where the true transfer function is not known we would conclude that the estimated model is reliable, a totally misleading result. We will come back to this point later and discuss a bit further the theoretical reason for such a bad behavior.

Return now to the LSCR approach. LSCR was used in a totally 'blind' manner, that is with no concern at all for the identification set-up characteristics; in particular, we did not pay any attention to the existence of local minima: The method is guaranteed by the theory and it will work in all possible situations covered by the theory.

In the present setting, the prediction error is given by

$$\epsilon_t(\theta) = \frac{1}{1 + hz^{-1}} y_t$$

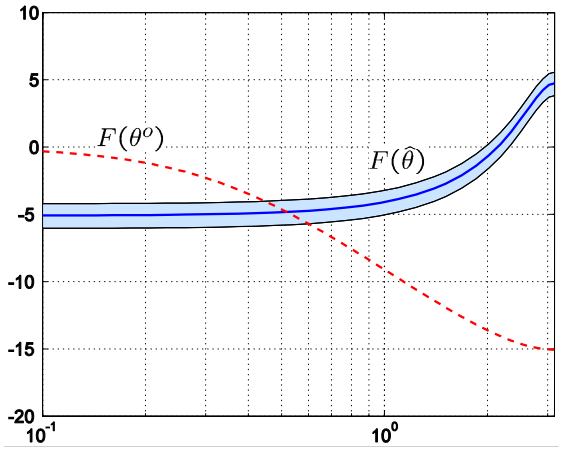

Fig. 11. 90% confidence region for the identified $u - y$ transfer function obtained with the asymptotic theory.

$$-\frac{bz^{-1}}{(1+az^{-1})(1+hz^{-1})}(u_t - y_t) = \frac{1+(a+b)z^{-1}}{(1+az^{-1})(1+hz^{-1})}y_t - \frac{bz^{-1}}{(1+az^{-1})(1+hz^{-1})}u_t.$$

The group was constructed as in the Appendix B ($2^l = 2048$), and we computed

$$g^{\epsilon}_{i,r}(\theta) = \sum_{k \in I_i} \epsilon_{k-r}(\theta)\epsilon_k(\theta), \quad r = 1, 2, 3,$$

in the parameter space, making the standard assumptions that $G(\theta)$ and $H(\theta)$ (i.e. the u to y and w to y closed-loop transfer functions) were stable ($|a + b| < 1$) and that $H(\theta)$ has a stable inverse ($|a| < 1$, $|h| < 1$). We excluded the regions in the parameter space where 0 was among the 34 smallest or largest values of any of the three correlations above to obtain a $1-3\cdot2\cdot34/2048 = 0.9004$ confidence set. The confidence set is shown in Figure 12. The set consists of two separate regions, one around the true parameter θ^0 and one around θ^*, the spurious minimizer. This illustrates the global features of the approach: LSCR produces two separate regions as the overall confidence set because information in the data is intrinsically ineffective in telling us which one of the two regions contain the true parameter.

Figures 13 and 14 show the close-ups of the two regions. The ellipsoid in Figure 13 is the 90% confidence set with the asymptotic PEM theory: When the PEM estimate gets trapped near θ^*, the confidence ellipsoid all concentrates around this spurious θ^* because the PEM asymptotic theory is local in nature (it is based on a Taylor expansion) and is therefore unable to explore locations far from the identified model. This is the reason why in Figure 11 we obtained a frequency domain

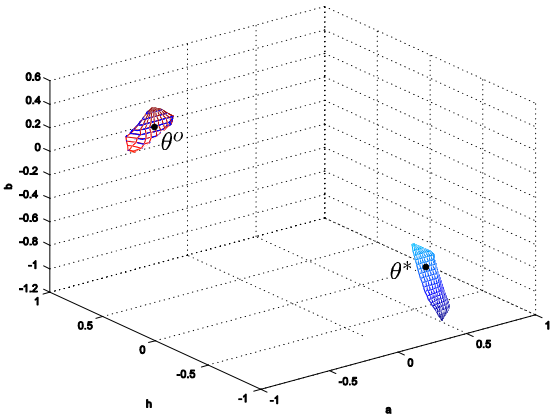

Fig. 12. 90% confidence set.

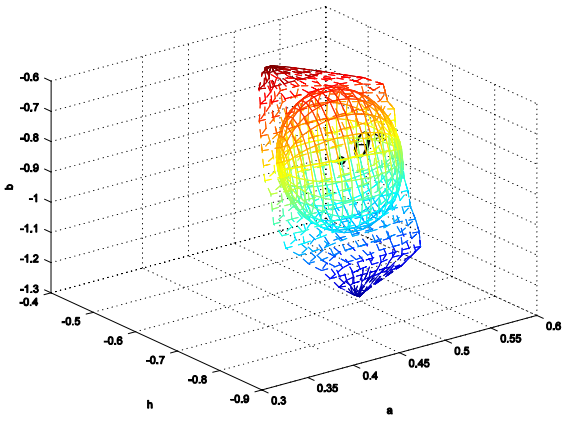

Fig. 13. Asymptotic confidence 90% ellipsoid (-), and the part of the non-asymptotic confidence set around θ^* (- -).

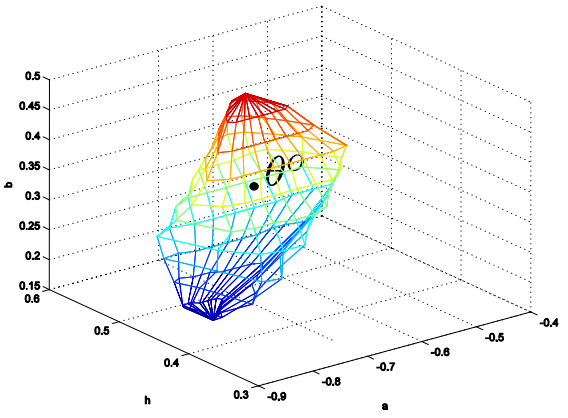

Fig. 14. Close-up of the non-asymptotic confidence region around θ^0.

confidence region unable to capture the real model uncertainty. The reader is referred to Garatti et al. (2004) for more details.

6. LSCR PROPERTIES

As we have seen in Section 4, Theorems 1 and 2 quantify the probability that θ^0 belongs to the

constructed regions. However, this theorem deals only with one side of the story. In fact, a good evaluation method must have two properties:

- the provided region must have guaranteed probability (and this is what Theorems 1 and 2 deliver);
- the region must be bounded, and, in particular, it should concentrate around θ^0 as the number of data points increases.

We next discuss how this second property can be achieved by choosing n_ϵ and n_u in (7). It turns out that the choice depends on the model class, and we here consider ARMA and ARMAX models, while general linear model classes are dealt with in Campi and Weyer (2005).

6.1 ARMA models

Data generating system and model class

The data generating system is given by

$$y_t = \frac{C(\theta^0)}{A(\theta^0)} w_t,$$

where

$$\begin{aligned} A(\theta^0) &= 1 + a_1^0 z^{-1} + \cdots + a_n^0 z^{-n}, \\ C(\theta^0) &= 1 + c_1^0 z^{-1} + \cdots + c_p^0 z^{-p}, \end{aligned}$$

and $\theta^0 = [a_1^0 \ \cdots \ a_n^0 \ c_1^0 \ \cdots \ c_p^0]^T$. In addition to the assumptions in Section 4.1 and in Theorem 1, we assume that $A(\theta^0)$ and $C(\theta^0)$ have no common factors and that w_t is wide-sense stationary with spectral density $\Phi_w(\omega) = \lambda_w^2 > 0$.

The model class is

$$y_t = \frac{C(\theta)}{A(\theta)} w_t,$$

where

$$\begin{aligned} A(\theta) &= 1 + a_1 z^{-1} + \cdots + a_n z^{-n}, \\ C(\theta) &= 1 + c_1 z^{-1} + \cdots + c_p z^{-p}, \end{aligned}$$

$\theta = [a_1 \ \cdots \ a_n \ c_1 \ \cdots \ c_p]^T$, and the assumptions in Section 4.1 are in place.

Confidence regions for ARMA models

We next give a result taken from Campi and Weyer (2005) which shows how a confidence region which concentrates around the true parameter as the number of data points increases can be obtained for ARMA systems.

Theorem 3. Let $\epsilon_t(\theta) = \frac{A(\theta)}{C(\theta)} y_t$ be the prediction error associated with the ARMA model class. Then, $\theta = \theta^0$ is the unique solution to the set of equations:

$$E[\epsilon_{t-r}(\theta)\epsilon_t(\theta)] = 0, \quad r = 1, \ldots, n+p.$$

Theorem 3 shows that if we simultaneously impose $n+p$ correlation conditions, where n and p are the orders of the $A(\theta)$ and $C(\theta)$ polynomials, then the only solution is the true θ^0. Guided by this idea, we consider $n+p$ sample correlation conditions, and let $n_\epsilon = n+p$ in (7):

$$\hat{\Theta} = \cap_{r=1}^{n+p} \Theta_r^\epsilon.$$

Theorem 2 guarantees that this set contains θ^0 with probability $1 - (n+p)q/M$, and Theorem 3 entails that the confidence set concentrates around θ^0.

6.2 ARMAX models

Data generating system and model class

Consider now system

$$y_t = \frac{B(\theta^0)}{A(\theta^0)} u_t + \frac{C(\theta^0)}{A(\theta^0)} w_t,$$

where

$$\begin{aligned} A(\theta^0) &= 1 + a_1^0 z^{-1} + \cdots + a_n^0 z^{-n}, \\ B(\theta^0) &= b_1^0 z^{-1} + \cdots + b_m^0 z^{-m}, \\ C(\theta^0) &= 1 + c_1^0 z^{-1} + \cdots + c_p^0 z^{-p}, \end{aligned}$$

and $\theta^0 = [a_1^0 \ \cdots \ a_n^0 \ b_1^0 \ \cdots \ b_m^0 \ c_1^0 \ \cdots \ c_p^0]^T$. In addition to the assumptions in Section 4.1 and in Theorem 1, we assume that $A(\theta^0)$ and $B(\theta^0)$ have no common factors and - similarly to the ARMA case - we assume a stationary environment. Precisely, w_t is wide-sense stationary with spectral density $\Phi_w(\omega) = \lambda_w^2 > 0$ and u_t is wide-sense stationary too and independent of w_t.

The model class is

$$y_t = \frac{B(\theta)}{A(\theta)} u_t + \frac{C(\theta)}{A(\theta)} w_t,$$

where $A(\theta)$, $B(\theta)$, $C(\theta)$ have the same structure as for the true system.

Confidence regions for ARMAX models

The next theorem taken from Campi and Weyer (2005) shows that we can choose correlation equations such that the solution is unique and equal to θ^0, provided that the input signal u_t is white.

Theorem 4. Let $\epsilon_t(\theta) = \frac{A(\theta)}{C(\theta)} y_t - \frac{B(\theta)}{C(\theta)} u_t$ be the prediction error associated with the ARMAX

model class. If u_t is white with spectral density $\Phi_u(\omega) = \lambda_u^2 > 0$, then $\theta = \theta^0$ is the unique solution to the set of equations:

$$E[u_{t-s}\epsilon_t(\theta)] = 0, \quad s = 1, \dots, n+m,$$
$$E[\epsilon_{t-r}(\theta)\epsilon_t(\theta)] = 0, \quad r = 1, \dots, p.$$

Guided by this result, we choose $n_\epsilon = p$ and $n_u = n + m$ in (7) to arrive at the following confidence region for ARMAX models:

$$\hat{\Theta} = \cap_{r=1}^{p} \Theta_r^\epsilon \cap_{s=1}^{n+m} \Theta_s^u.$$

Interestingly enough, the conclusion of Theorem 4 does not hold true for colored input sequences, see Campi and Garatti (2003). On the other hand, assuming that u_t is white is often unrealistic. This impasse can be circumvented by resorting to suitable prefiltering actions, as indicated in Campi and Weyer (2005).

6.3 Properties of LSCR

To summarize, LSCR has the following properties:

- for suitable selections of the correlations, the region shrinks around θ^0;
- for any sample size, θ^0 belongs to the constructed region with given probability, despite that no assumption on the level of noise is made.

7. COMPLEMENTS

In this Part I, the only restrictive assumption on noise was that it had symmetric distribution around zero. This assumption can be relaxed as briefly discussed here.

(i) Suppose that w_t in Section 4 has median 0, i.e. $Pr\{w_t \geq 0\} = Pr\{w_t < 0\} = 0.5$ (note that this is a relaxation of the symmetric distribution condition). Then, theory goes through by considering everywhere $sign(\epsilon_t(\theta))$ instead of $\epsilon_t(\theta)$, where 'sign' is signum function: $sign(x) = 1$ if $x > 0$, $sign(x) = -1$ if $x < 0$ and $sign(x) = 0$ if $x = 0$.

(ii) When w_t is independent and identically but not symmetrically distributed, we can obtain symmetrically distributed data by considering the difference between two subsequent data points, that is $(y_t - y_{t-1}) = G(\theta)(u_t - u_{t-1}) + H(\theta)(w_t - w_{t-1})$; here, $w_t - w_{t-1}$, $t = 2, 4, 6, \dots$ are independent and symmetrically distributed around 0 and we can refer to this 'difference' system to construct confidence regions.

PART II: Presence of unmodeled dynamics

In this second part, we discuss the possibility to deal with unmodeled dynamics within the LSCR framework: Despite that the true system is not within the model class, we would like to derive guaranteed results for some parts of the system.

We start by describing the problem of identifying a full-order u to y transfer function, without deriving a model for the noise. Then, we turn to also consider unmodeled dynamics in the u to y transfer function. General ideas are only discussed by means of simple examples.

8. IDENTIFICATION WITHOUT NOISE DESCRIPTION: AN EXAMPLE

Consider the system

$$y_t = \theta^0 u_t + n_t.$$

Suppose that the structure of the u to y transfer function is known. Instead, the noise n_t describes all other sources of variation in y_t apart from u_t and we do not want to make any assumption on how n_t is generated. Correspondingly, we want that our results regarding the value of θ^0 are valid for any (unknown) deterministic noise sequence n_t with no constraints whatsoever. When the noise is stochastic, the result will then hold for any realization of the noise, that is surely.

In this section, we assume that we have access to the system for experiment: We are allowed to generate a finite number, say 7, of input data and - based on the collected outputs - we are asked to construct a confidence interval $\hat{\Theta}$ for θ^0 of guaranteed probability.

The problem looks very challenging indeed: Since the noise can be whatever, it seems that the observed data are unable to give us a hand in constructing a confidence region. In fact, for any given θ^0 and u_t, a suitable choice of the noise sequence can lead to any observed output signal! Let us see how this problem can be circumvented.

Before proceeding, we feel advisable to make clear what is meant here by 'guaranteed probability'. We said that n_t is regarded as a deterministic sequence, and the result is required to hold true for any n_t, that is uniformly in n_t. The stochastic element is instead the input sequence: We will

select u_t according to a random generation mechanism and we require that $\theta^0 \in \hat{\Theta}$ with a given probability value, where the probability is with respect to the random choice of u_t.

We first indicate input design and then the procedure for construction of the confidence interval $\hat{\Theta}$.

Input design

Let u_t, $t = 1, \ldots, 7$, be independent and identically distributed with distribution

$$u_t = \begin{cases} 1, & \text{with probability } 0.5 \\ -1, & \text{with probability } 0.5. \end{cases}$$

Procedure for construction of the confidence interval $\hat{\Theta}$

Rewrite the system as a model with generic parameter θ:

$$y_t = \theta u_t + n_t.$$

We construct a prediction by dropping the noise term n_t whose characteristics are unknown:

$$\hat{y}_t(\theta) = \theta u_t, \quad \epsilon_t(\theta) = y_t - \hat{y}_t(\theta) = y_t - \theta u_t.$$

Next, we compute the prediction errors $\epsilon_t(\theta)$ from the observed data for $t = 1, \ldots, 7$ and calculate

$$f_t(\theta) = u_t \epsilon_t(\theta), \quad t = 1, \ldots, 7.$$

The rest of the construction is the same as for the preliminary example of Section 3: We consider the same group of subsets as given in the bullet table in that example and construct the $g_i(\theta)$ functions as in (4). Then, we extract the interval where at least two functions are below zero and at least two are above zero. The reader can verify that the theoretical analysis for the example in Section 3 goes through here to conclude that the obtained interval has probability 0.5 to contain the true θ^0. Interestingly, the property of w_t to be independent and symmetrically distributed have been replaced here by analogous properties of the input signal; the advantage is that - if the experiment can be designed - these properties can be easily enforced and no restrictive conditions on the noise are required anymore.

A simulation example was run where $\theta^0 = 1$ and the noise was the sequence shown in Figure 15. This noise sequence was obtained as a realization of a biased independent Gaussian process with mean 0.5 and variance 0.1. The obtained $g_i(\theta)$ functions and the corresponding confidence region are given in Figure 16.

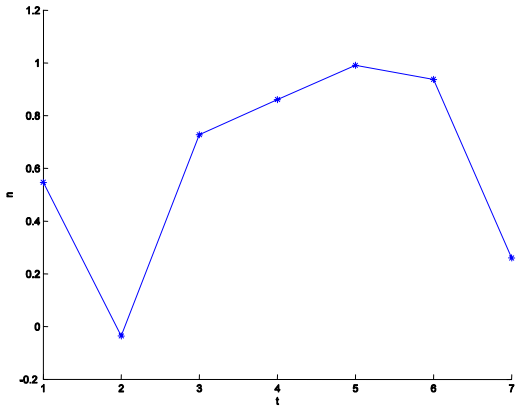

Fig. 15. Noise sequence.

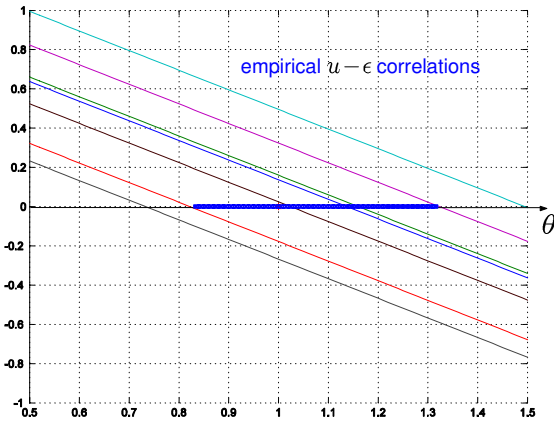

Fig. 16. The $g_i(\theta)$ functions.

9. UNMODELED DYNAMICS IN THE U TO Y TRANSFER FUNCTION: AN EXAMPLE

Suppose that a system has structure

$$y_t = \theta_0^0 u_t + \theta_1^0 u_{t-1} + n_t,$$

while - for estimation purposes - we use the reduced order model

$$y_t = \theta u_t + n_t.$$

The noise has been indicated with a generic n_t to signify that it can be whatever, and not just a white signal. In fact, a perspective similar to the previous section is taken and we regard n_t as a generic unknown deterministic signal.

After determining a region for the parameter θ, one sensible question to ask is: Does this region contain with a given probability the system parameter θ_0^0 linking u_t to y_t?

Reinterpreting the above question we are asking whether the projection of the true transfer function $\theta_0^0 + \theta_1^0 z^{-1}$ onto the 1-dimensional space spanned by constant transfer functions is contained in the estimated set with a certain probability.

We generate an input signal u_t in the same way as in the previous section, this time over the time interval $t = 0, \ldots, 7$, and inject it into the system. Then, the predictor and prediction error are constructed the same way as in the previous section, while we add a *sign* function to $f_t(\theta)$:

$$f_t(\theta) = sign(u_t \epsilon_t(\theta)), \quad t = 1, \ldots, 7. \tag{9}$$

Corresponding to the true parameter value, i.e. $\theta = \theta_0^0$, an easy inspection reveals that $sign(u_t \epsilon_t(\theta_0^0)) = sign(u_t(\theta_1^0 u_{t-1} + n_t))$ is an independent and symmetrically distributed process (it is in fact a Bernoullian process taking on values ± 1 with probability 0.5 each). Thus, with $f_t(\theta)$ as in (9), the situation is similar to what we had in the previous section and again the theory goes through to prove that an interval for θ constructed as in the previous section has probability 0.5 to contain θ_0^0, despite the presence of unmodeled dynamics.

A simulation example was run with $\theta_0^0 = 1$, $\theta_1^0 = 0.5$ and where the noise was again the realization of a biased Gaussian process given in Figure 15. As $sign(u_t \epsilon_t(\theta))$ only can take on the values -1, 1 and 0, it is possible that two or more of the $g_i(\theta)$ functions will take on the same value on an interval (in technical terms, the assumption on the existence of densities in Theorem 1 does not hold). This tie can be broken by introducing a random ordering (e.g. by adding a random constant number between -0.1 and 0.1 to the $g_i(\theta)$ functions) and one can see that the theory remains valid. The obtained $g_i(\theta)$ functions and confidence region are in Figure 17.

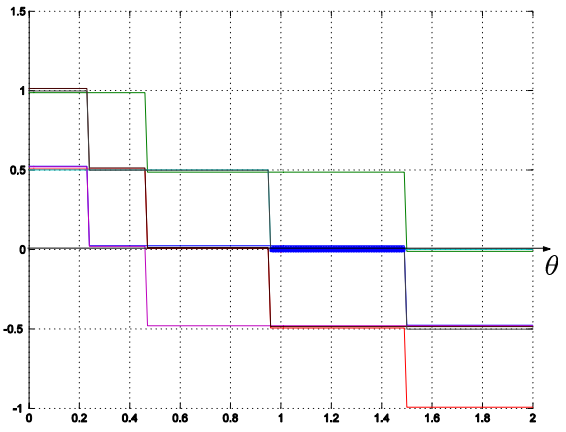

Fig. 17. The $g_i(\theta)$ functions.

Though presented on simple examples, the approaches illustrated in Sections 8 and 9 to deal with unmodeled dynamics bear a general breath of applicability.

PART III: Nonlinear systems

Interestingly, the identification set-up developed in previous sections in a linear setting generalizes naturally to nonlinear systems. Such an extension is presented in this part III with reference to a simple example.

10. IDENTIFICATION OF NONLINEAR SYSTEMS: AN EXAMPLE

Consider the following nonlinear system

$$y_t = \theta^0(y_{t-1}^2 - 1) + w_t, \tag{10}$$

where w_t is an independent and symmetrically distributed sequence and θ^0 is an unknown parameter.

This system can be made explicit with respect to w_t as follows:

$$w_t = y_t - \theta^0(y_{t-1}^2 - 1),$$

and - by substituting θ^0 with a generic θ and renaming the so-obtained right-hand-side as $w_t(\theta)$ - we have

$$w_t(\theta) = y_t - \theta(y_{t-1}^2 - 1).$$

Note that $w_t(\theta)$ coincides in this example with the prediction error for the model with parameter θ. It is not true, however, that the above construction generates the prediction error for any nonlinear system. For example, if we make explicit system $y_t = \theta^0 y_{t-1} w_t$ with respect to w_t we get $w_t = y_t/\theta^0 y_{t-1}$, and further substituting a generic θ we have $w_t(\theta) = y_t/\theta y_{t-1}$. This is not the prediction error since the predictor is in this case $\hat{y}_t(\theta) = 0$, so that the prediction error is here given by $y_t - \hat{y}_t = y_t$. Note also that for linear systems $w_t(\theta)$ is always equal to the prediction error $\epsilon_t(\theta)$ so that the construction suggested here for the generation of $w_t(\theta)$ generalizes the construction of $\epsilon_t(\theta)$ for linear systems.

Now, an inspection of the proof in Appendix A reveals that the only property that has a role in determining the result that the confidence region contains θ^0 with a given probability is that $\epsilon_t(\theta^0) = w_t$. Since this same property holds here for $w_t(\theta)$, i.e. $w_t(\theta^0) = w_t$, we can argue that proceeding in the same way as for the preliminary example of Section 3 where $\epsilon_t(\theta)$ is replaced by $w_t(\theta)$ still generates in the present context a guaranteed confidence region.

Is it all so simple? It is indeed, as far as the guarantee that θ^0 is in the region is concerned. The dark side of the medal is that second order

statistics are in general not enough to spot the real parameter value for nonlinear systems, so that results like those in Section 6 do not apply to conclude that the region shrinks around θ^0.

In actual effects, if e.g. $\theta^0 = 0$ and $E[w_t^2] = 1$, some easy computation reveals that $E[w_{t-r}(\theta)w_t(\theta)] = 0$ for any θ and for any $r > 0$ so that the second-order statistics are useless.

Now, the good news is that LSCR can be upgraded to higher-order statistics with little effort. A general presentation of the related results can be found in Dalai et al. (2005). Here, it suffices to say that we can e.g. consider the third-order statistic $E[w_t(\theta)^2 w_{t+1}(\theta)]$ and the theory goes through unaltered.

As an example, we generated 9 samples of y_t, $t = 0, \ldots, 8$ for system (10) where w_t is zero-mean Gaussian with variance 1. Then, we constructed

$$g_i(\theta) = \frac{1}{4} \sum_{k \in I_i} w_k(\theta)^2 w_{k+1}(\theta), \quad i = 1, \ldots, 8,$$

where the sets I_i are as in Section 3. These functions are displayed in Figure 18. The interval marked in blue where at least two functions are below zero and at least two are above zero has probability 0.5 to contain $\theta^0 = 0$.

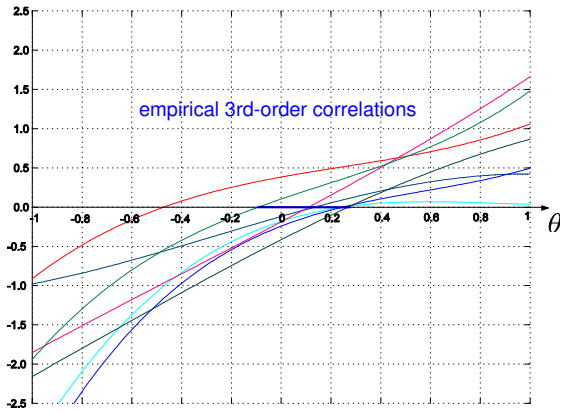

Fig. 18. The $g_i(\theta)$ functions.

11. CONCLUSIONS

In this paper we have provided an overview of the LSCR method for system identification. Most of the existing theory for system identification is asymptotic in the number of data points while in practice one will only have a finite number of samples available. Although the asymptotic theory often delivers sensible results when applied heuristically to a finite number of data points, the results are not guaranteed. The LSCR method delivers guaranteed finite sample results, and it produces a set of models to which the true system belongs with a given probability for any finite number of data points.

As illustrated by the simulation examples, the method is not only grounded on a solid theoretical basis, but it also delivers practically useful confidence sets.

The LSCR method takes a global approach and can, when the situation requires, produce a confidence set which consists of disjoint regions, and hence it has advantages over confidence ellipsoids based on the asymptotic theory.

By allowing the user to choose the input signal, the LSCR method can be extended to deal with unmodeled dynamics, and it can also be extended to non-linear systems by considering higher-order statistics.

REFERENCES

Bartlett, P.L (2003). Prediction algorithms: complexity, concentration and convexity. *Proc. of the 13th IFAC Symposium on System Identification*, Rotterdam, The Netherlands, pp. 1507-1517.

Bittanti, S. and M. Lovera (2000). Bootstrap-based estimates of uncertainty in subspace identification methods. *Automatica*, Vol. 36, pp. 1605-1615.

Bosq, D. (1998). *Nonparametric Statistics for Stochastic Processes - Estimation and Prediction*. Lecture Notes in Statistics 110. Springer Verlag.

Campi, M.C. and S. Garatti (2003). Correlation approach for ARMAX model identification: A counterexample to the uniqueness of the asymptotic solution. *Internal report of the university of Brescia*.

Campi, M.C. and P.R. Kumar (1998). Learning Dynamical Systems in a Stationary Environment. *Systems and Control Letters*, Vol. 34, pp. 125-132.

Campi, M.C. and E. Weyer (2002). Finite sample properties of system identification methods. *IEEE Trans. on Automatic Control*, Vol. 47, pp. 1329-1334.

Campi, M.C. and E. Weyer (2005). Guaranteed non-asymptotic confidence regions in system identification. *Automatica*, Vol. 41, pp. 1751-1764.

Campi, M.C., S.K. Ooi and E. Weyer (2004). Non-asymptotic quality assessment of generalised FIR models with periodic inputs. *Automatica*, Vol. 40, pp. 2029-2041.

Cherkassky, V. and F. Mulier (1998). *Learning from data*. John Wiley.

Dahleh, M.A., T.V. Theodosopoulos and J.N. Tsitsiklis (1993). The sample complexity of worst-case identification of FIR linear sys-

tems. *System and Control Letters*, Vol. 20, pp. 157-166.

Dahleh, M.A., E.D. Sontag, D.N.C. Tse and J.N. Tsitsiklis (1995). Worst-case identification of nonlinear fading memory systems. *Automatica*, Vol. 31, pp. 503-508.

Dalai, M., E. Weyer and M.C. Campi (2005). Parametric identification of nonlinear systems: Guaranteed confidence regions. In *Proc. of the 44th IEEE CDC*, Seville, Spain, pp. 6418-6423.

Ding, F. and T. Chen (2005). Performance bounds on forgetting factor least-squares algorithms for time-varying systems with finite measurement data. *IEEE Trans on Circuits and Systems-I*, Vol. 52, pp. 555-566.

Dunstan, W.J. and R.R. Bitmead (2003). Empirical estimation of parameter distributions in system identification, In *Proc. of the 13th IFAC Symposium on System Identification*, Rotterdam, The Netherlands.

Efron, B., and R.J. Tibshirani (1993). *An introduction to the bootstrap*, Chapman and Hall.

Garatti, S., M.C. Campi and S. Bittanti (2004). Assessing the quality of identified models through the asymptotic theory - When is the result reliable? *Automatica*, Vol. 40, pp. 1319-1332.

Garatti, S., M.C. Campi and S. Bittanti (2006). The asymptotic model quality assessment for instrumental variable identification revisited. *System and Control Letters*, to appear.

Garulli, A., L. Giarre' and G. Zappa (2002). Identification of approximated Hammerstein models in a worst-case setting. *IEEE Trans. on Automatic Control*, Vol. 47, pp. 2046-2050.

Garulli, A., A. Vicino and G. Zappa (2000). Conditional central algorithms for worst-case set membership identification and filtering. *IEEE Trans. on Automatic Control*, Vol. 45, pp. 14-23.

Giarre', L., B.Z. Kacewicz and M. Milanese (1997a). Model quality evaluation in set membership identification. *Automatica*, Vol. 33, pp. 1133-1139.

Giarre', L., M. Milanese and M. Taragna (1997b). H_∞ identification and model quality evaluation. *IEEE Trans. on Automatic Control*, Vol. 4, pp. 88-199.

Goldenshluger, A. (1998). Nonparametric estimation of transfer functions: rates of convergence and adaptation. *IEEE Trans. on Information Theory*, Vol. 44, pp. 644-658.

Gordon, L. (1974). Completely separating groups in subsampling. *Annals of Statistics*, Vol. 2, pp. 572-578.

Harrison, K.J., J.A. Ward and D.K. Gable (1996). Sample complexity of worst-case H^∞-identification. *Systems and Control Letters*, Vol. 27, pp. 255-260.

Hartigan, J.A. (1969). Using subsample values as typical values. *Journal of American Statistical Association*, Vol. 64, pp. 1303-1317.

Hartigan, J.A. (1970). Exact confidence intervals in regression problems with independent symmetric errors. *Annals of Mathematical Statistics*, Vol. 41, pp. 1992-1998.

Heath, W.P. (2001). Bias of indirect non-parametric transfer function estimates for plants in closed loop. *Automatica*, Vol. 37, pp. 1529-1540.

Hjalmarsson, H. and B. Ninness (2004). An exact finite sample variance expression for a class of frequency function estimates. In *Proc. of the 43rd IEEE CDC*, Bahamas.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, Vol. 58, pp. 13-30.

Karandikar R. and M. Vidyasagar (2002). Rates of uniform convergence of empirical means with mixing processes. *Statistics and Probability Letters*, Vol. 58, pp. 297-307.

Meir, R. (2000). Nonparametric Time Series Prediction Through Adaptive Model Selection. *Machine Learning*, Vol. 39, pp. 5-34.

Ljung, L. (1999). *System Identification - Theory for the User*. 2nd Ed., Prentice Hall.

Milanese, M. and M. Taragna (2005). H_∞ set membership identification. A survey. *Automatica*, Vol. 41, pp. 2019-2032.

Milanese, M. and A. Vicino (1991). Optimal estimation theory for dynamic systems with set membership uncertainty: an overview. *Automatica*, Vol. 27, pp. 997-1009.

Modha, D.S. and E. Masry (1996). Minimum complexity regression estimation with weakly dependent observations. *IEEE Trans. Information Theory*, Vol. 42, pp. 2133-2145.

Modha, D.S. and E. Masry (1998). Memory-Universal Prediction of Stationary Random Processes. *IEEE Trans. Information Theory*, Vol. 44, pp. 117-133.

Ninness, B. and H. Hjalmarson (2004). Variance Error quantifications that are exact for finite model order. *IEEE Trans. on Automatic Control*, Vol. 49, pp. 1275-1291.

Politis, D.N. (1998). Computer-intensive methods in statistical analysis. *IEEE Signal Processing Magazine*, Vol. 15, pp. 39-55.

Politis, D.N., J.P. Romano and M. Wolf (1999). *Subsampling*. Springer.

Poolla, K. and A. Tikku (1994). On the Time Complexity of Worst-Case System Identification. *IEEE Trans. on Automatic Control*, Vol. 39, pp. 944-950.

Richmond, S. (1964). *Statistical Analysis.* 2nd Ed., The Ronald Press Company, N.Y.

Shao, J., and D. Tu (1995). *The Jackknife and Bootstrap*. Springer.

Smith, R. and G.E. Dullerud (1996). Continuous-time control model validation using finite experimental data. *IEEE Trans. on Auotmatic Control*, Vol. 41, pp. 1094-1105.

Söderström, T. and P. Stoica (1988). *System Identification*. Prentice Hall.

Spall, J.C. (1995). Uncertainty Bounds for Parameter Identification with Small Sample Sizes. *Proc. of the 34th IEEE CDC*, New Orleans, Louisiana, USA, pp. 3504-3515.

Tjärnström, F. and L. Ljung (2002). Using the Bootstrap to estimate the variance in the case of undermodeling. *IEEE Trans. on Automatic Control*, Vol. 47, pp. 395-398.

Vapnik, V. (1998). *The nature of statistical learning theory*. Springer.

Vapnik, V.N. and A.Ya. Chervonenkis (1968). Uniform convergence of the frequencies of occurence of events to their probabilities. *Soviet Math. Doklady*, Vol. 9, pp. 915-968.

Vapnik, V.N. and A.Ya. Chervonenkis (1971). On the uniform convergence of relative frequencies to their probabiliies. *Theory of Prob. and its Appl.*, Vol. 16, pp. 264-280.

Venkatesh, S.R. and M. A. Dahleh (2001). On system identification of complex systems from finite data. *IEEE Trans. on Automatic Control*, Vol. 46, pp. 235-257.

Vicino, A. and G. Zappa (1996). Sequential approximation of feasible parameter sets for identification with set membership uncertainty. *IEEE Trans. on Automatic Control*, Vol. 41, pp. 774-785.

Vidyasagar, M. (2002). *A theory of Learning and Generalization*. 2nd Ed., Springer Verlag.

Vidyasagar, M. and R. Karandikar (2002). A learning theory approach to system identification and stochastic adaptive control. In *IFAC Symp. on Adaptation and Learning*.

Vidyasagar, M. and R. Karandikar (2006). A learning theory approach to system identification and stochastic adaptive control. In G. Calafiore and F. Dabbene (Eds.) *Probabilistic and randomized methods for design under uncertainty*. Springer.

Welsh, J.S. and G.C. Goodwin (2002). Finite sample properties of indirect nonparametric closed-loop identification. *IEEE Trans. on Automatic Control*, Vol. 47, pp. 1277-1292.

Weyer, E. (2000). Finite sample properties of system identification of ARX models under mixing conditions. *Automatica* Vol. 36, pp. 1291-1299.

Weyer, E. and M.C. Campi (2002). Non-asymptotic confidence ellipsoids for the least squares estimate. *Automatica*, Vol. 38, pp. 1539-1547.

Weyer, E., R. C. Williamson and I. M. Y. Mareels (1999). Finite sample properties of linear model identification. *IEEE Trans. on Automatic Control*, Vol. 44, pp. 1370-1383.

Yu, B. (1994). Rates of convergence for empirical processes of stationary mixing sequences. *Annals of Probability*, Vol. 22, pp. 94-116.

Appendix A. PROOF OF THE RESULT

In the confidence region construction, we eliminate the regions in parameter space where all functions $g_i(\theta)$, $i = 1, \ldots, 7$, are above zero, or at most one of them is below zero and where all functions are below zero, or at most one of them is above zero. Therefore, the true parameter value θ^0 falls outside the confidence region when - corresponding to $\theta = \theta^0$ - all functions $g_i(\theta)$, $i = 1, \ldots, 7$, are bigger than $g_8(\theta) = 0$, or at most one of them is smaller than $g_8(\theta)$ and where all functions are smaller than $g_8(\theta)$, or at most one of them is bigger than $g_8(\theta)$. We claim that each one of these events has probability $1/8$ to happen, so that the total probability that θ^0 falls outside the confidence region is $4 \cdot 1/8 = 0.5$, as claimed in the RESULT.

We next concentrate on one condition only and compute the probability that 'all functions $g_i(\theta^0)$, $i = 1, \ldots, 7$, are bigger than $g_8(\theta^0) = 0$'. The probability of the other conditions can be derived similarly.

The considered condition writes

$$\begin{cases} w_1w_2 + w_2w_3 + w_4w_5 + w_5w_6 > 0 \\ w_1w_2 + w_3w_4 + w_4w_5 + w_6w_7 > 0 \\ w_2w_3 + w_3w_4 + w_5w_6 + w_6w_7 > 0 \\ w_1w_2 + w_2w_3 + w_6w_7 + w_7w_8 > 0 \\ w_1w_2 + w_3w_4 + w_5w_6 + w_7w_8 > 0 \\ w_2w_3 + w_3w_4 + w_4w_5 + w_7w_8 > 0 \\ w_4w_5 + w_5w_6 + w_6w_7 + w_7w_8 > 0. \end{cases} \tag{A.1}$$

To compute the probability that all these 7 inequalities are simultaneously true let us ask the following question: What would we have written if instead of comparing $g_i(\theta^0)$, $i = 1, \ldots, 7$, with $g_8(\theta^0)$ we would have compared $g_i(\theta^0)$, $i = 2, \ldots, 8$, with $g_1(\theta^0)$? The conditions would have been:

$$\begin{cases} w_1w_2 + w_3w_4 + w_4w_5 + w_6w_7 \\ \qquad > w_1w_2 + w_2w_3 + w_4w_5 + w_5w_6 \\ w_2w_3 + w_3w_4 + w_5w_6 + w_6w_7 \\ \qquad > w_1w_2 + w_2w_3 + w_4w_5 + w_5w_6 \\ w_1w_2 + w_2w_3 + w_6w_7 + w_7w_8 \\ \qquad > w_1w_2 + w_2w_3 + w_4w_5 + w_5w_6 \\ w_1w_2 + w_3w_4 + w_5w_6 + w_7w_8 \\ \qquad > w_1w_2 + w_2w_3 + w_4w_5 + w_5w_6 \\ w_2w_3 + w_3w_4 + w_4w_5 + w_7w_8 \\ \qquad > w_1w_2 + w_2w_3 + w_4w_5 + w_5w_6 \\ w_4w_5 + w_5w_6 + w_6w_7 + w_7w_8 \\ \qquad > w_1w_2 + w_2w_3 + w_4w_5 + w_5w_6 \\ 0 > w_1w_2 + w_2w_3 + w_4w_5 + w_5w_6, \end{cases}$$

or, moving everything to the left-hand-side,

$$\begin{cases} -w_2w_3 + w_3w_4 - w_5w_6 + w_6w_7 > 0 \\ -w_1w_2 + w_3w_4 - w_4w_5 + w_6w_7 > 0 \\ -w_4w_5 - w_5w_6 + w_6w_7 + w_7w_8 > 0 \\ -w_2w_3 + w_3w_4 - w_4w_5 + w_7w_8 > 0 \\ -w_1w_2 + w_3w_4 - w_5w_6 + w_7w_8 > 0 \\ -w_1w_2 - w_2w_3 + w_6w_7 + w_7w_8 > 0 \\ -w_1w_2 - w_2w_3 - w_4w_5 - w_5w_6 > 0. \end{cases}$$

If we now let $\tilde{w}_2 = -w_2$, $\tilde{w}_5 = -w_5$, the latter condition re-writes as

$$\begin{cases} \tilde{w}_2w_3 + w_3w_4 + \tilde{w}_5w_6 + w_6w_7 > 0 \\ w_1\tilde{w}_2 + w_3w_4 + w_4\tilde{w}_5 + w_6w_7 > 0 \\ w_4\tilde{w}_5 + \tilde{w}_5w_6 + w_6w_7 + w_7w_8 > 0 \\ \tilde{w}_2w_3 + w_3w_4 + w_4\tilde{w}_5 + w_7w_8 > 0 \\ w_1\tilde{w}_2 + w_3w_4 + \tilde{w}_5w_6 + w_7w_8 > 0 \\ w_1\tilde{w}_2 + \tilde{w}_2w_3 + w_6w_7 + w_7w_8 > 0 \\ w_1\tilde{w}_2 + \tilde{w}_2w_3 + w_4\tilde{w}_5 + \tilde{w}_5w_6 > 0. \end{cases} \quad \text{(A.2)}$$

Except for the ' ~ ' showing up here and there, this latter set of inequalities is the same as the original set of inequalities (A.1) (the order in which the inequalities are listed is changed but the inequalities altogether are the same - this is a consequence of the group property of the sets I_i). Moreover, since the w_t variables are symmetrically distributed, the change of sign implied by the '~' is immaterial as far as the probability of satisfaction of the inequalities in (A.2) is concerned, so that we can conclude that (A.1) and (A.2) are satisfied with the same probability.

Now, instead of comparing the $g_i(\theta^0)$'s with $g_1(\theta^0)$, we could have compared with $g_2(\theta^0)$, or with $g_3(\theta^0)$ or ... or with $g_7(\theta^0)$ arriving all the time to a similar conclusion that the probability does not change. Since these 8 events ($g_1(\theta^0)$ is the smallest, $g_2(\theta^0)$ is the smallest, etc.) are disjoint and cover all possibilities and all of them have the same probability, we finally draw the conclusion that each and every event has exactly probability 1/8 to happen. It remains therefore proven that the initial condition (A.1) is satisfied with probability 1/8 and this concludes the proof.

Appendix B. GORDON'S CONSTRUCTION OF THE INCIDENT MATRIX OF A GROUP

Given $I = \{1, \ldots, N\}$, the incident matrix for a group $\{I_i\}$ of subsets of I is a matrix whose (i, j) element is 1 if $j \in I_i$ and zero otherwise. In Gordon (1974), the following construction procedure for an incident matrix $\bar{R}$ is proposed where $I = \{1, \ldots, 2^l - 1\}$ and the group has 2^l elements.

Let $R(1) = [1]$, and recursively compute ($k = 2, 3, \ldots, l$)

$$R(2^k - 1) = \begin{bmatrix} R(2^{k-1} - 1) & R(2^{k-1} - 1) & 0 \\ R(2^{k-1} - 1) & J - R(2^{k-1} - 1) & e \\ 0^T & e^T & 1 \end{bmatrix},$$

where J and e are, respectively, a matrix and a vector of all ones, and 0 is a vector of all zeros. Then, let

$$\bar{R} = \begin{bmatrix} R(2^l - 1) \\ 0^T \end{bmatrix}.$$

Gordon (1974) also gives construction of groups when the number of data points is different from $2^l - 1$.