Guaranteed Non-asymptotic Confidence Regions in System Identification

M.C. Campi [a], E. Weyer [b]

[a] *Department of Electrical Engineering and Automation, University of Brescia, Via Branze 38, 25123 Brescia, Italy*

[b] *Department of Electrical and Electronic Engineering, The University of Melbourne, Parkville, VIC 3010, Australia*

Abstract

In this paper we consider the problem of constructing confidence regions for the parameters of identified models of dynamical systems. Taking a major departure from the previous literature on the subject, we introduce a new approach called 'Leave-out Sign-dominant Correlation Regions' (LSCR) which delivers confidence regions with guaranteed probability. All results hold rigorously true for any finite number of data points and no asymptotic theory is involved. Moreover, prior knowledge on the noise affecting the data is reduced to a minimum. The approach is illustrated on several simulation examples, showing that it delivers practically useful confidence sets with guaranteed probabilities.

Key words: Confidence sets, Uncertainty evaluation, General linear models, Finite sample results, System identification

⋆ This paper was not presented at any IFAC meeting. Corresponding author E. Weyer. Tel. +613-83449726. Fax +613-83446678

Email addresses: campi@ing.unibs.it (M.C. Campi), e.weyer@ee.unimelb.edu.au (E. Weyer).

1 Introduction, preview example, and discussion

2 Introduction, preview example, and discussion

Models and uncertainty

Models of dynamical systems are used in many fields of science and engineering. It is however widely recognised that a model is of limited use if no quality tag is attached to it. The quality tag should give a description of the uncertainties associated with the model, and the accuracy of the model should be taken into account when it is used in practice.

A good technique or methodology for model uncertainty evaluation should meet the following two requirements:

(1) it is applicable under general conditions;
(2) it provides a non-conservative evaluation of the system uncertainties.

Regarding the first item we note that evaluation methods that can be used for a large class of systems and noise characteristics are desirable. For example, restrictive assumptions on the noise (e.g. that it is Gaussian or bounded), means that the theory is not applicable to many real life systems, and, even if it is, the verification of the assumptions may be difficult in a given application. The second point is important because loose uncertainty evaluations generate conservativeness in the belief that the model is less reliable than it actually is. For example, a robust controller looses in performance as the level of uncertainty increases. The reader is referred to our recent paper Campi et al. (2004) for a broader discussion on these points.

One additional point that needs to be kept in mind is that, in system identification (e.g. Ljung (1999), Söderström and Stoica (1988)), one always uses a *finite* number of data points. And, in fact, uncertainty in the model is due to such a finiteness, as no limit to the accuracy would exist if an infinite amount of information were available. Likewise, for the evaluation of model quality and construction of confidence sets one will only have a finite amount of data available. Thus, a sound uncertainty evaluation method must provide results valid when the number of data is finite, and, possibly, small.

The new theory presented in this paper

In this paper we introduce a novel approach for the construction of confidence regions in system identification called 'Leave-out Sign-dominant Correlation Regions', LSCR for short. LSCR exhibits the following features:

(i) *It provides regions to which the true system parameter belongs with an exact guaranteed probability.*

This means that if the user decides to determine a, say, 95% confidence region, the

method returns a region which has exactly probability 0.95 of containing the true system parameter; no overbounding is introduced and, therefore, no conservativeness in the found regions is present (yet, in order to obtain confidence regions of suitable shape one may be willing to intersect different regions in which case some overbounding is introduced).

(ii) *The prior assumptions on the noise are reduced to a minimum.*

The only assumption is that the noise is symmetrically distributed around zero. Apart from that, 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 σ^2 can be any (unknown) number: $\sigma^2 = 0.001, 0.1, 10$ or 10^3, and yet the confidence region has guaranteed exact probability. Of course, depending on the strength of the noise, the region will be wider or smaller; the important point is that knowledge of the noise characteristics is not a-priori required: the method *let the data speak* and automatically outputs regions that are correct relative to the existing level of the noise.

(iii) *LSCR is guaranteed for any data set size.*

Evaluating uncertainty is more important when the uncertainty is significant which is the case when the information conveyed by data is limited. LSCR holds rigorously for any information content and size of the data set.

The above features credit the LSCR method with a promise of making a great impact on the model quality evaluation theory (see Section 2.2 for further comments and comparison with existing results). Yet, the final word is not written as this paper leaves an important point substantially open for further discussion: working out procedures for construction of the confidence regions at low computational cost for high order systems.

To make things concrete from the beginning, a simple example that shows how LSCR works is given in the next section. Section 2.2 contains additional remarks and comparison with the literature.

2.1 A preview example

Consider the system

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

where $a^0 = 0.2$ and $\{w_t\}$ is an independent sequence of random variables uniformly distributed between -1 and 1 (the distribution of w_t is given for completeness of description, but it is not used in the algorithm). 9 data points were generated according to (1) and shown in Figure 1. Our goal is to form a confidence region for a^0 from the available data set. The small number of data points is used to facilitate comprehension of the example as well as a full description of the obtained results.

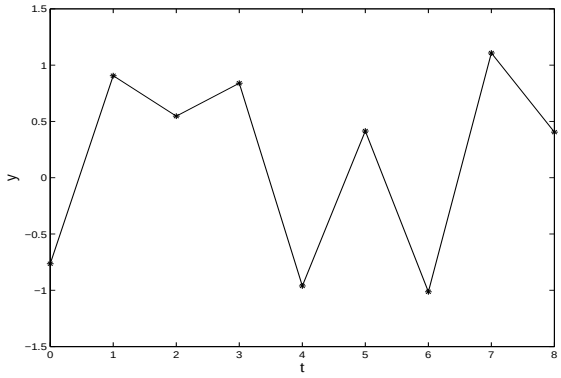

Fig. 1. Data for the preview example

Rewrite the system as a model with generic parameter a:

$$y_t + ay_{t-1} = w_t.$$

The predictor and prediction error associated with the model are

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

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

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

Using the $f_{t-1}(a)$'s, we want to form empirical estimates of the correlation $E[\epsilon_{t-1}(a)\epsilon_t(a)]$. For $a = a^0$ we have that $E[\epsilon_{t-1}(a^0)\epsilon_t(a^0)] = E[w_{t-1}w_t] = 0$. We therefore expect the empirical estimates to be zero mean random variables for $a = a^0$. Based on this observation, we compute a number of estimates of the correlation using different subsets of the data, and we discard those regions in parameter space where the empirical estimates take positive (or negative) value too many times (from which the name of the method derives: Leave-out Sign-dominant Correlation Regions). To eventually provide rigorous results, these empirical estimates, however, need to be constructed very carefully as illustrated in the following.

First, we generate a set G of subsets of $I = \{1, \ldots, 7\}$ which is a group with respect to the symmetric difference, i.e. $(I_i \cup I_j) - (I_i \cap I_j) \in G$, if $I_i, I_j \in G$. The set I is the index set for the seven functions $f_1(a), f_2(a), \ldots, f_7(a)$, and each set in the group G gives the indices of the functions $f_i(a)$ used for computing one particular empirical estimate. The group considered in this example is described by the incident matrix below where each row corresponds to a subset in the group. A 1 means that the element is in the set, while a 0 means that the element is not in the set.

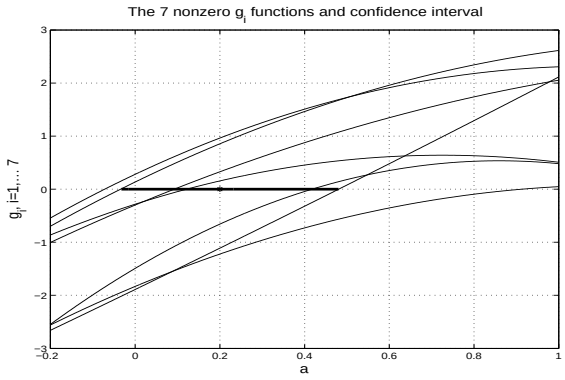

Fig. 2. The $g_i(a)$ functions for the preview example together with the confidence interval.

	1	2	3	4	5	6	7
I_1	1	1	0	1	1	0	0
I_2	1	0	1	1	0	1	0
I_3	0	1	1	0	1	1	0
I_4	1	1	0	0	0	1	1
I_5	1	0	1	0	1	0	1
I_6	0	1	1	1	0	0	1
I_7	0	0	0	1	1	1	1
I_8	0	0	0	0	0	0	0

The estimates of the correlation $E[\epsilon_{t-1}(a)\epsilon_t(a)]$ (in fact a re-scaled version as no normalization is present) are then given by

$$g_i(a) = \sum_{k \in I_i} f_k(a), \quad i = 1, \ldots, 8$$

($g_8(a) = 0$ since we let $g_i(a) = 0$ if $I_i = \emptyset$). The seven nonzero $g_i(a)$'s are plotted in Figure 2 as functions of a.

Following the LSCR idea, we recognise that it is very unlikely that all the $g_i(a^0)$'s have the same sign, and we therefore discard the rightmost and leftmost regions where at most one function out of the seven non-zero functions is less than zero or greater than zero. The resulting interval $[-0.04, 0.48]$, is the confidence region for a^0. It is a rigorous fact (stated in Theorem 3.1) that the confidence region constructed this way has probability $1 - 2 \cdot 2/8 = 0.5$ to contain the true parameter value a^0.

As expected, due to the small number of data points, this confidence interval is rather large and the associated probability is low. Next we increase the number of data points to 1025 and using the group with incidence matrix $R(1023)$ (see

Appendix A.5) we keep the region in parameter space where at least 25 of the 1023 nonzero $g_i(a)$ functions are greater than 0 and at least 25 are smaller than zero. The resulting interval [0.12, 0.2425], contains the true parameter value a^0 with exact probability $1 - 25 \cdot 2/1024 = 0.9512 > 95\%$ (see Theorem 3.1).

A verification of the theoretical confidence result was performed by running the last simulation with 1025 data points 5000 times. The empirical frequency of a^0 being in the confidence interval was 0.9490, in good agreement with the theoretical result.

2.2 Some discussion on the existing literature on model quality evaluation

Quite often, uncertainty evaluations and confidence ellipsoids are derived based on the asymptotic theory of system identification. It is common experience of theorists and practitioners that this theory - though applied heuristically with a finite number of data points - in many situations delivers sensible results. On the other hand, the correctness of the results is not guaranteed, and contributions (Bittanti et al. (2002), Garatti et al. (2003,2004)) have appeared that show that the asymptotic theory may fail to be reliable in certain situations. Moreover, when the available data is scarce, using asymptotic results makes no sense at all. Thus, there is a need for developing techniques that provide results guaranteed for finite data samples.

Our earlier finite sample results (e.g. Campi and Weyer (2002) and Weyer and Campi (2002)) were data independent, in the sense that they were uniform with respect to the considered class of data generating systems, and they could essentially be evaluated without any data. Because of the uniformity, it was realised that the results could be quite conservative for the particular system at hand. The approach presented here is data based as it uses data generated by the actual system at hand, and avoids the problems due to uniformity. Finite sample results using a data based approach have also been developed in Campi et al. (2002, 2004), and of course many popular techniques such as bootstrap are data based (Tjärnström and Ljung (2002), Bittanti and Lovera (2000)). However, few rigorous finite sample results exists for bootstrap methods.

Similarly to set membership identification, e.g. Milanese and Vicino (1991), Bai et al. (1995,1996), Vicino and Zappa (1996), Giarre' et al. (1997), Garulli et al. (2000, 2002), LSCR returns regions for the true system parameter. However, unlike the typical setting in set membership identification, LSCR does not assume that the disturbances are deterministic or bounded.

In this paper we develop a methodology for construction of confidence sets for a general linear system based on a finite number of data points. The confidence sets have guaranteed probability of containing the true parameter. The sets are constructed with no a-priori knowledge on the noise level, and they are concentrated around the true parameter. The developed theory is rigorously valid for any finite data sample and gives us a practically useful method for model quality evaluation which stands

on a solid theoretical footing.

The mathematical approach taken in this paper is inspired by the work of Hartigan (Hartigan (1969,1970)) in the statistical literature on estimating a constant in noise. In this paper we consider a different problem from Hartigan (1969,1970) as we are interested in dynamical systems and the theory developed herein departs significantly from the work of Hartigan. Still, some basic concepts used in this paper are common with those used by Hartigan and, in a sense, this paper can also be seen as a contribution in the direction of fertilizing the area of system identification with ideas inspired from a certain area of the statistical literature.

2.3 Organisation of the paper

This paper is organised in three main sections: Section 3 present the LSCR procedure for constructing confidence regions. LSCR is applied to ARMA, ARMAX, and general linear models in Section 4 and simulation examples are finally provided in Section 5.

3 Confidence regions for linear systems

3.1 Data generating system

The data are generated by a general linear system

$$y_t = G^0(z^{-1})u_t + H^0(z^{-1})w_t,$$

where $G^0(z^{-1})$ and $H^0(z^{-1})$ are stable rational transfer functions. z^{-1} is the backward shift operator ($z^{-1}y_t = y_{t-1}$). $H^0(z^{-1})$ is monic and has a stable inverse and $G^0(z^{-1})$ has a delay of 1 or more time units. $\{w_t\}$ is a zero-mean independent sequence (noise). No a-priori knowledge of the noise level is assumed. The system operates in open loop, that is $\{w_t\}$ and $\{u_t\}$ are independent. Closed loop systems are discussed in Section 4.4.

3.2 Model structure

The model class consists of full order models

$$y_t = G(z^{-1}, \theta)u_t + H(z^{-1}, \theta)w_t,$$

which are parameterised by θ. We assume that there exists a unique parameter θ^0 such that $G(z^{-1}, \theta^0) = G^0(z^{-1})$ and $H(z^{-1}, \theta^0) = H^0(z^{-1})$. Moreover, we assume

that θ is restricted to a set Θ such that $H(z^{-1},\theta)$ is monic, $G(z^{-1},\theta)$, $H(z^{-1},\theta)$ and $H^{-1}(z^{-1},\theta)$ are stable and $G(z^{-1},\theta)$ has a delay of 1 or more time units for all $\theta \in \Theta$.

Remark 3.1 *It should be noted that the goal of the present paper is to construct confidence regions for the system parameter (as opposed to identifying a nominal model). Similar to other existing techniques for model quality evaluation (e.g. standard asymptotic theory (Ljung (1999a)) and bootstrap techniques (Tjärnström and Ljung (2002))), a full description of the system is adopted. It is important to remark, however, that this in no way enforces a full order nominal model: one can use a reduced order nominal model and then verify its reliability through a full order model for quality evaluation along the approach of this paper. Moreover, the quality evaluation can as well be directly applied to the model error similarly to the model error modelling approach, Ljung (1999b,2001).*

3.3 Construction of confidence regions

We start by describing procedures for the determination of confidence sets Θ_r^ϵ and Θ_s^u based on correlation properties of ϵ (the prediction error) at different time instants and on cross-correlation properties of ϵ and u. Confidence sets $\hat{\Theta}$ for θ^0 can then be constructed by taking the intersection of a number of the Θ_r^ϵ and Θ_s^u sets as discussed at the end of this section.

Procedure for the construction of Θ_r^ϵ

(1) Compute the prediction errors

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

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

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

(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 $I_i, I_j \in G$). [1] Compute

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

[1] A group is a non-empty set G with a binary operation $\circ$ such that: $a \circ b \in G$, $\forall a, b \in G$ and $a \circ (b \circ c) = (a \circ b) \circ c$, $\forall a, b, c \in G$; there exists an identity element $e \in G$ such that $a \circ e = e \circ a = a$, $\forall a \in G$; for every $a \in G$, there exists an inverse a^{-1} such that $a \circ a^{-1} = a^{-1} \circ a = e$.

(4) Select an integer q in the interval $[1, (M+1)/2)$ and find the region Θ_r^ϵ such that at least q of the $g_{i,r}^\epsilon(\theta)$ 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 preview example in Section 2.1. In that example we had $H^{-1}(z^{-1},\theta) = 1 + az^{-1}$, $G(z^{-1},\theta) = 0$, and $K = 8, N = 7, r = 1, M = 8$ and $q = 2$.

Remark 3.2 *The group in the procedure has identity element $e = \emptyset$, and the inverse of an element is the element itself, i.e. $I_i^{-1} = I_i$. In the procedure, the group G can be freely selected. Thus, if $I = \{1,2,3,4\}$, 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_r^ϵ is guaranteed to be a confidence region in any case (see Theorem 3.1), the shape and size of Θ_r^ϵ is affected by the choice made. Moreover, 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. These sets are particularly well suited for use in point 3 above and we always use Gordon's construction. Gordon's procedure is summarised in appendix A.5 for completeness, and the reader is referred to Gordon (1974) for further discussion.*

The intuitive idea behind this construction is that, for $\theta = \theta^0$, the functions $g_{i,r}^\epsilon(\theta)$ assume positive or negative value at random ($\epsilon_t(\theta^0)$ is a zero mean independent sequence), 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_r^ϵ discards regions where all $g_{i,r}^\epsilon(\theta)$'s but a small fraction (q should be taken to be small compared to M) are of the same sign, we expect that $\theta^0 \in \Theta_r^\epsilon$ with high probability. This is put on solid mathematical grounds in Theorem 3.1 below, showing that the probability that $\theta^0 \in \Theta_r^\epsilon$ 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_r^ϵ shrinks and the probability that $\theta^0 \in \Theta_r^\epsilon$ decreases.

The procedure for construction of the sets Θ_s^u 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 Θ_s^u

(1) Compute the prediction errors

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

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

(2) Select an integer $s \geq 1$. 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.

Theorem 3.1 *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:*

$$Pr\{\theta^0 \in \Theta^\epsilon_r\} = 1 - 2q/M, \tag{2}$$
$$Pr\{\theta^0 \in \Theta^u_s\} = 1 - 2q/M. \tag{3}$$

■

Proof. See appendix A.1. ■

Remark 3.3 *The only reason for requiring that the variables w_t and $w_t u_\tau$ admit densities is to avoid that the functions $g^\epsilon_{i,r}(\theta)$ and $g^u_{i,s}(\theta)$ can take on the same value with nonzero probability. This condition prevents ties from occurring in point 4 of the procedures for constructing Θ^ϵ_r and Θ^u_s. It can be dropped by using a random ordering in case of ties, but we have preferred to maintain the condition to avoid unduly complications.*

When the $\{w_t\}$ process 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 $(y_t - y_{t-1}) = G(z^{-1}, \theta)(u_t - u_{t-1}) + H(z^{-1}, \theta)(w_t - w_{t-1})$; here, $w_t - w_{t-1}$, $t = 2, 4, 6, \ldots$ are independent and symmetrically distributed around 0.

The noise assumption is mild enough to accommodate a number of situations. In particular, one can describe possible outliers by allowing the noise to take on large values with small probability. Importantly, the procedures return regions of guaranteed probability despite that we do not assume any a-priori knowledge on the noise level: 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.

The evaluations (2) and (3) 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 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.

$$\hat{\Theta} = \cap_{r=1}^{n_\epsilon} \Theta_r^\epsilon \cap_{s=1}^{n_u} \Theta_s^u. \tag{4}$$

The next 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 the number of parameters in the model, and these issues will be discussed in detail in the next section.

We conclude this section with a fact which is immediate from Theorem 3.1.

Theorem 3.2 *Under the assumptions of Theorem 3.1,*

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

where $\hat{\Theta}$ *is given by (4).* ∎

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. This is illustrated later on in the simulation example in Section 5.1.

Remark 3.4 *Note that for each one of the sets* Θ_r^ϵ *and* Θ_s^u *we could have chosen a different group and a different value of* q *with the effect that the probabilities on the right hand side of (2) and (3) change to* $1 - 2q_r^\epsilon/M_r^\epsilon$ *and* $1 - 2q_s^u/M_s^u$*, where* q_r^ϵ *and* q_s^u *are the numbers chosen in point 4 of the constructions and* M_r^ϵ *and* M_s^u *are the number of elements in the chosen group. However, in order to keep the notation relatively simple we have presented our results with a fixed* q *and* M*. The generalisation is however straightforward.*

The confidence region $\hat{\Theta}$ can be used to validate/invalidate a given model. However, we do not want to focus solely on this specific application and we prefer to see $\hat{\Theta}$ as the output of an identification procedure returning a guaranteed set, independently of the use we make of this set. As an alternative example to validating a model, $\hat{\Theta}$ could be used directly in robust control design.

Finally, note that the probability in Theorem 3.2 is with respect to the observed data sequences, and it bounds the probability that we observe realisations of $\{w_t\}$ and $\{u_t\}$ such that $\theta^0 \in \hat{\Theta}$. Here, θ^0 is deterministic and the random element is $\hat{\Theta}$

because this set depends on observations and hence the event that $\hat{\Theta}$ contains θ^0 depends on the observed data. In theory we may even obtain $\hat{\Theta} = \emptyset$, but this just tells us that we have had the misfortune of observing one of the (rare) realisations where $\theta^0 \notin \hat{\Theta}$.

4 Confidence sets for different model classes

As we have seen in the previous section, Theorem 3.1 quantifies the probability that θ^0 belongs to the regions Θ_r^ϵ and Θ_s^u. It holds for any finite N and is non-conservative. On the other hand, Theorem 3.1 deals only with one side of the medal in the study of uncertainty evaluation techniques. A good evaluation method must have two properties: the provided region must have guaranteed probability (and this is what Theorems 3.1 and 3.2 deliver); and the region must be bounded, and, in particular, it should concentrate around θ^0 as the number of data points increases. In this section we show how this second requirement can be achieved by choosing n_ϵ and n_u in (4). The choices depend on the model class, and next we consider ARMA and ARMAX models, followed by general linear model classes.

4.1 ARMA models

4.1.1 Data generating system and model class

The data generating system is given by

$$y_t = \frac{C^0(z^{-1})}{A^0(z^{-1})} w_t,$$

where

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

In addition to the assumptions in Section 3.1 and in Theorem 3.1, we assume that $A^0(z^{-1})$ and $C^0(z^{-1})$ 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

$$y_t = \frac{C(z^{-1}, \theta)}{A(z^{-1}, \theta)} w_t,$$

where

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

$\theta = [a_1 \ \cdots \ a_n \ c_1 \ \cdots \ c_p]^T$, and the assumptions in Section 3.2 are in place.

4.1.2 Confidence regions for ARMA models

We next give a result 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 4.1 *Let* $\epsilon_t(\theta) = \frac{A(z^{-1},\theta)}{C(z^{-1},\theta)} y_t$ *be the prediction error associated with the ARMA model class. Then,* $\theta = \theta^0 = [a_1^0 \ \cdots \ a_n^0 \ c_1^0 \ \cdots \ c_p^0]^T$ *is the unique solution to the set of equations:*

$$E[\epsilon_{t-r}(\theta)\epsilon_t(\theta)] = 0, \quad r = 1, \ldots, n+p. \tag{5}$$

Proof. See Appendix A.2. ∎

Theorem 4.1 shows that if we simultaneously impose $n + p$ correlation conditions, where n and p are the orders of the $A(z^{-1}, \theta)$ and $C(z^{-1}, \theta)$ 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 (4). As $N \to \infty$, the functions $\frac{1}{N_i} g_{i,r}^\epsilon(\theta) \to E[\epsilon_{t-r}(\theta)\epsilon_t(\theta)]$, provided that N_i, the number of elements in the set I_i, also tends to infinity (this is the case for the groups in Gordon (1974)). This means that each region Θ_r^ϵ gets smaller and the intersection of them gives an uncertainty region shrinking around the true parameter θ^0. This leads to the following construction of confidence regions for ARMA models.

Confidence region for ARMA models

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

Theorem 3.2 guarantees that this set contains θ^0 with probability at least $1 - (n + p)q/M$, and Theorem 4.1 shows that the confidence set concentrates around θ^0.

4.2 ARMAX models

4.2.1 Data generating system and model class

The data generating system is given by

$$y_t = \frac{B^0(z^{-1})}{A^0(z^{-1})} u_t + \frac{C^0(z^{-1})}{A^0(z^{-1})} w_t,$$

where

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

In addition to the assumptions in Section 3.1 and in Theorem 3.1, we assume that $A^0(z^{-1})$ and $B^0(z^{-1})$ 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 $\{w_t\}$ (open loop configuration, closed loop systems are considered in Section 4.4).

The model class is

$$y_t = \frac{B(z^{-1},\theta)}{A(z^{-1},\theta)} u_t + \frac{C(z^{-1},\theta)}{A(z^{-1},\theta)} w_t,$$

where

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

$\theta = [a_1 \ \cdots \ a_n \ b_1 \ \cdots \ b_m \ \ c_1 \ \cdots \ c_p]^T$ and the assumptions in Section 3.2 are in place.

4.2.2 Confidence regions for ARMAX models

The next theorem shows that we can choose correlation equations such that the solution is unique and equal to θ^0, provided the input signal $\{u_t\}$ is white.

Theorem 4.2 *Let $\epsilon_t(\theta) = \frac{A(z^{-1},\theta)}{C(z^{-1},\theta)} y_t - \frac{B(z^{-1},\theta)}{C(z^{-1},\theta)} u_t$ 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 $\theta = \theta^0 = [a_1^0 \ \cdots \ a_n^0 \ b_1^0 \ \cdots \ b_m^0 \ c_1^0 \ \cdots \ c_p^0]^T$ is the unique solution to the set of equations:*

$$E[u_{t-s}\epsilon_t(\theta)] = 0, \quad s = 1, \ldots, n+m, \tag{6}$$
$$E[\epsilon_{t-r}(\theta)\epsilon_t(\theta)] = 0, \quad r = 1, \ldots, p. \tag{7}$$

Proof. See Appendix A.3. ∎

Guided by this result, we choose $n_\epsilon = p$ and $n_u = n + m$ in (4) to arrive at

Confidence region for ARMAX models

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

Remark 4.1 *Interestingly enough, the conclusion of Theorem 4.2 does not hold true for coloured input sequences, see the simulation example in Section* **??**.

Thus, with $\{u_t\}$ white, no problems arise and the situation is similar to the ARMA case. On the other hand, assuming that $\{u_t\}$ is white is often unrealistic and hence we are well advised to discuss how to remove such an assumption.

Suppose that $\{u_t\}$ is prefiltered by a filter $L(z^{-1})$ before it used in point 2 in the construction of Θ_s^u, that is point 2 is substituted by

2'. Select an integer $s \geq 1$. For $t = 1 + s, \ldots, N + s = K$, compute

$$f_{t-s,s}^{u}(\theta) = (L(z^{-1})u_{t-s})\epsilon_t(\theta).$$

Then, Theorem 3.1 (and 3.2) remains valid, as can be verified by inspecting its proof. In fact, we can also allow the filter $L(z^{-1})$ to be dependent on the input signal, that is, it can be constructed from the input signal without affecting the validity of Theorem 3.1. Moreover, if the filter $L(z^{-1})$ is appropriately chosen, θ^0 is the unique solution to the correlation equations, as stated in the next theorem.

Theorem 4.3 *Assume $u_t = Q(z^{-1})\nu_t$ with $\{\nu_t\}$ a white wide-sense stationary sequence of random variables with spectral density $\Phi_\nu(\omega) = \lambda_\nu^2 > 0$ and $Q(z^{-1})$ is a rational and stable transfer function. Let $L(z^{-1}) = Q(z^{-1})^{-1}Q(z)^{-1}$, then θ^0 is the unique solution to the set of equations*

$$E[(L(z^{-1})u_{t-s})\epsilon_t(\theta)] = 0, \quad s = 1, \ldots, n+m \tag{8}$$
$$E[\epsilon_{t-r}(\theta)\epsilon_t(\theta)] = 0, \quad r = 1, \ldots, p. \tag{9}$$

∎

Proof. See Appendix A.4. ∎

Remark 4.2 *The fact that the filter $L(z^{-1})$ is unstable is not much of a concern since all operations are performed in batch so that $L(z^{-1})u_{t-s}$ can be computed as a solution having a causal as well as an anti-causal component.*

Also note that an imprecise estimation of $Q(z^{-1})$ does not affect the validity of Theorem 3.1, so that the obtained region does have the guaranteed probability of containing the true θ^0. The issue here is the shape of the region, which, if uniqueness is missing, may comprise spurious portions around the solutions of equations (8) and (9) that do not correspond to θ^0, see Campi and Weyer (2004) for an example. In

that example it is also shown that the spurious regions disappear even if the applied filter is only an approximation of $Q(z^{-1})^{-1}Q(z)^{-1}$.

4.3 General linear models

We now turn to the case of a general linear system

$$y_t = G^0(z^{-1})u_t + H^0(z^{-1})w_t, \tag{10}$$

where $\{u_t\}$ is wide-sense stationary with spectral density $\Phi_u(\omega)$ and $\{w_t\}$ is a white noise sequence with spectral density $\Phi_w(\omega) = \lambda_w^2 > 0$. Moreover, we assume that $\{u_t\}$ and $\{w_t\}$ are independent. As usual, we consider a full order model class $y_t = G(z^{-1}, \theta)u_t + H(z^{-1}, \theta)w_t$. Moreover, the assumptions in Sections 3.1 and 3.2 and in Theorem 3.1 are in place.

For ARMA and ARMAX models we saw in the previous sections that by computing scaled empirical values of certain correlations for a finite number of lags we obtained well shaped confidence regions concentrated around θ^0. In the present general setting it can be shown that imposing $E[\epsilon_{t-r}(\theta)\epsilon_t(\theta)] = 0,\ \forall r \geq 1$, and $E[u_{t-s}\epsilon_t(\theta)] = 0,\ \forall s \geq 1$, returns θ^0 as the unique solution. Computing an infinite number of correlations is however not possible, but it is common experience in practice that imposing as many correlations as there are parameters is sufficient in order to obtain a well shaped region around θ^0. In this section a heuristic guideline for how to choose appropriate correlations is given. The guideline is based on the relative strength of the external signals. It is perhaps worth mentioning once more that it is only the shape of the obtained regions which is under discussion whereas Theorem 3.1 (and 3.2) are always valid so that the regions do have a guaranteed confidence.

Guideline 1 *Compute as many empirical correlations as there are parameters in the model. Let* n *be the number of parameters which* $G(z^{-1}, \theta)$ *and* $H(z^{-1}, \theta)$ *have in common,* m *the number of parameters which appear exclusively in* $G(z^{-1}, \theta)$ *and* p *the number of parameters which appear exclusively in* $H(z^{-1}, \theta)$*. Choose at least* m *correlations* $u_{t-s}\epsilon_t(\theta),\ s = 1, \ldots, m$*, and at least* p *correlations* $\epsilon_{t-r}(\theta)\epsilon_t(\theta)$*,* $r = 1, \ldots, p$*. When choosing the last* n *correlations take into account the a priori information about the energy in the signals* u_t *and* w_t *and how exciting they are. Favour correlations of the type* $u_{t-s}\epsilon_t(\theta)$ *if* u_t *is the stronger signal and correlations of the type* $\epsilon_{t-r}(\theta)\epsilon_t(\theta)$ *if* w_t *is the stronger signal.*

This guideline is put to practice in Section 5.2.

4.4 Closed loop systems

Consider a general linear system (10), and assume now that the input is generated by a feedback controller

$$u_t = K(z^{-1})(\tilde{r}_t - y_t),$$

where $\tilde{r}_t$ is a reference signal, so that the closed loop system is stable. In this context we regard the closed loop system as a whole with inputs $\tilde{r}_t$ and w_t, i.e. (the argument z^{-1} is suppressed)

$$\begin{aligned} y_t &= (1 + KG^0)^{-1} KG^0 \tilde{r}_t + (1 + KG^0)^{-1} H^0 w_t \\ &= \bar{G}^0 \tilde{r}_t + \bar{H}^0 w_t. \end{aligned}$$

It is now clear that we are in exactly the same situation as in Section 4.3 with $\tilde{r}_t$ replacing u_t, and therefore the standard approach can be applied in the present context by imposing correlation conditions between $\epsilon_{t-r}(\theta)$ and $\epsilon_t(\theta)$ and between $\tilde{r}_{t-s}$ and $\epsilon_t(\theta)$.

The closed loop system is parameterised in terms of the parameters of the open loop model. The LSCR approach determines directly whether a parameter value is in the confidence set or not, and hence we do not need to perform a deconvolution to find the confidence regions for the parameters of G and H. A closed loop simulation example is given in Section 5.2.

Remark 4.3 *It may be of interest to note that imposing correlation relations $u_{t-s}\epsilon_t(\theta)$ does not preserve the result in Theorem 3.1 since $\{u_t\}$ and $\{\epsilon_t(\theta^0)\}$ are not independent processes and the proof for the open loop configuration no longer applies.*

5 Simulation examples

In this section, we present simulation examples illustrating the developed methodology.

5.1 First order ARMA model

Consider the ARMA system

$$y_t + a^0 y_{t-1} = w_t + c^0 w_{t-1}, \tag{11}$$

where $a^0 = -0.5, c^0 = 0.2$ and $\{w_t\}$ is an independent sequence of zero mean normally distributed random variables with variance 1. 1025 data points were generated according to (11). As a model class we used $y_t + ay_{t-1} = w_t + cw_{t-1},\ |a| < 1,\ |c| < 1$, with associate 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 (a^0, c^0) we calculated

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

and then computed

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

using the group in Appendix A.5. Next we discarded those values of a and c for which zero was among the 12 largest and smallest values of $g^{\epsilon}_{i,1}(a,c)$ and $g^{\epsilon}_{i,2}(a,c)$. Then according to Theorem 3.2 (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 3. The area marked with `x` is where 0 is among the 12 smallest values of $g^{\epsilon}_{i,1}$, the area marked with `+` is where 0 is among the 12 largest values of $g^{\epsilon}_{i,1}$. Likewise for $g^{\epsilon}_{i,2}$ 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 3 also plotted the confidence ellipsoid obtained using the asymptotic theory (Ljung (1999a), 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.

The reader is also referred to Campi and Weyer (2004) for a simulation example of an ARMAX system, omitted here due to space limitations. In that ARMAX example, the input signal is non-white and there exist parameter values other than the true ones which make the expected value of the correlations zero in Theorem 4.2 (see discussion in Section 4.2). However, it is demonstrated that the simple filtering procedure discussed after Remark 4.1 removes the parameter values which do not correspond to the true ones from the confidence region.

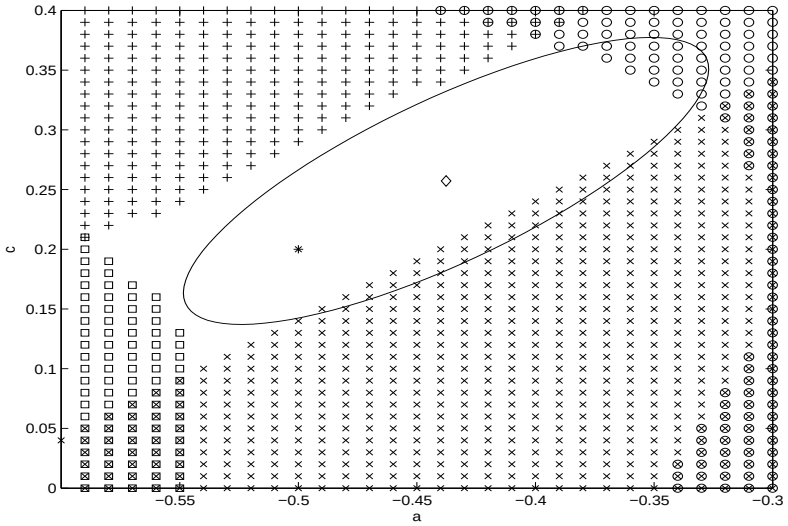

Fig. 3. Non-asymptotic confidence region for (a^0, c^0) (blank region) and asymptotic confidence ellipsoid. $\star$ = true parameter, $\diamond$ = estimated parameter using a prediction error method.

5.2 Closed loop system, general linear model structure

The following example is taken from Garatti et al. (2004). Consider the system

$$y_t = \frac{b^0 z^{-1}}{1 + a^0 z^{-1}} u_t + (1 + h^0 z^{-1}) w_t, \tag{12}$$

with $\theta^0 = [a^0 \ b^0 \ h^0]^T = [-0.7 \ 0.3 \ 0.5]^T$. $\{w_t\}$ is white Gaussian noise with variance 1, and the input is generated by

$$u_t = \tilde{r}_t - y_t \tag{13}$$

where $\tilde{r}_t$ is white Gaussian noise with variance 10^{-6}.

In Garatti et al. (2004) the asymptotic variance of the parameter estimation error was computed using asymptotic system identification theory, and it was shown that this theory could give misleading results for the above system. The reason for this can be explained as follows.

Using a standard quadratic criterion $V_N(\theta) = 1/N \sum_{t=1}^{N} \epsilon^2(t, \theta)$, the estimate is given by $\hat{\theta}_N = \arg\min_\theta V_N(\theta)$. Asymptotically the estimate $\hat{\theta}_N$ converges to a value which minimises $V(\theta) = E[\epsilon^2(t, \theta)]$. For $\tilde{r}_t \equiv 0$, there are two isolated parameters which minimise $V(\theta)$. These values are the true parameter θ^0 and $\bar{\theta} = [h^0 \quad a^0 - h^0 + b^0 \quad a^0]^T$. When the input signal is different from zero, but poorly exciting, the only minimum is θ^0, but $V(\theta^0)$ and $V(\bar{\theta})$ are close, and since the estimate is found by minimising $V_N(\theta)$, it will often end up being close to $\bar{\theta}$ which is now only a local minimum of $V(\theta)$. The asymptotic theory for evaluation of the variance of $\hat{\theta}_N - \theta^0$ is based on a Taylor series expansion of $\sqrt{N} V_N'(\theta)$ around the true parameter θ^0 ($'$ and $''$ denote

first and second derivative w.r.t. θ), i.e.

$$0 = \sqrt{N}V_N'(\hat{\theta}_N) = \sqrt{N}V_N'(\theta^0) + V_N''(\xi_N)\sqrt{N}(\hat{\theta}_N - \theta^0),$$

where ξ_N is a point between $\hat{\theta}_N$ and θ^0. When the asymptotic expressions are used in the finite sample case, $V''(\xi_N)$ is replaced by $V_N''(\hat{\theta}_N)$. This can give rise to a large error when $\hat{\theta}_N$ is far from θ^0 and this is what happens in this example. The net result is that the obtained confidence region is deceivingly small.

As is clear from above, one reason why the asymptotic theory gives unreliable results is that it is local in nature in the sense that it is based on a Taylor expansion evaluated in $\hat{\theta}_N$. It will only deliver a confidence set around the estimated parameter. This is in contrast to the non-asymptotic theory developed here which is global in nature as no local approximations are involved.

Returning to our approach for generating a confidence region we consider a full order model $y_t = \frac{bz^{-1}}{1+az^{-1}}u_t + (1+hz^{-1})w_t$. The prediction errors are given by

$$\begin{aligned}\epsilon_t(\theta) &= \frac{1}{1+hz^{-1}}y_t - \frac{bz^{-1}}{(1+az^{-1})(1+hz^{-1})}u_t \\ &= \frac{1+(a+b)z^{-1}}{(1+az^{-1})(1+hz^{-1})}y_t - \frac{bz^{-1}}{(1+az^{-1})(1+hz^{-1})}\tilde{r}_t.\end{aligned}$$

As the system operates in closed loop, we consider $\tilde{r}_t$ as the input signal, and the model structure is $y_t = \bar{G}(z^{-1},\theta)\tilde{r}_t + \bar{H}(z^{-1},\theta)w_t$ with

$$\bar{G}(z^{-1},\theta) = \frac{bz^{-1}}{1+(a+b)z^{-1}} \quad \text{and} \quad \bar{H}(z^{-1},\theta) = \frac{(1+hz^{-1})(1+az^{-1})}{1+(a+b)z^{-1}}.$$

We have three parameters, one which belongs to $\bar{H}(z^{-1})$ only, and two which belong to both $\bar{G}(z^{-1})$ and $\bar{H}(z^{-1})$. Using the Guideline in Section 4.3, we compute three correlations, one of them being $\epsilon_{t-1}(\theta)\epsilon_t(\theta)$. As $\tilde{r}_t$ is a poorly exciting signal compared to w_t, we choose the other two correlations to be $\epsilon_{t-2}(\theta)\epsilon_t(\theta)$ and $\epsilon_{t-3}(\theta)\epsilon_t(\theta)$.

We generated 2047+3 data points ($N = 2047$) according to (12) and (13). The group was constructed as in the appendix A.5 ($M = 2048$), and we computed

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

in the parameter space, making the standard assumptions that $\bar{G}(z^{-1},\theta)$ and $\bar{H}(z^{-1},\theta)$ were stable ($|a+b|<1$) and that $\bar{H}(z^{-1},\theta)$ 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 $1-3\cdot 2\cdot 34/2048 = 0.9004$ confidence set. Note that, despite the heuristics in choosing the correlation functions, the probability of the obtained set is still guaranteed. The confidence set is shown in Figure 4. The set consists of two separate regions, one around the true

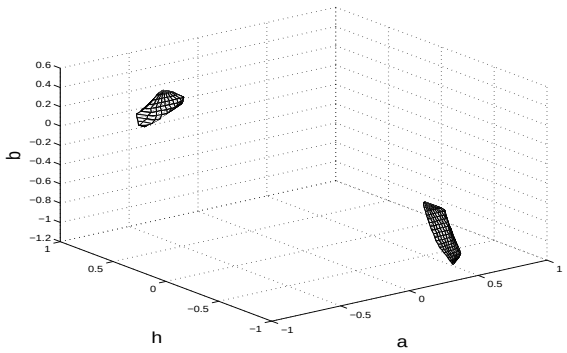

Fig. 4. 90% confidence region.

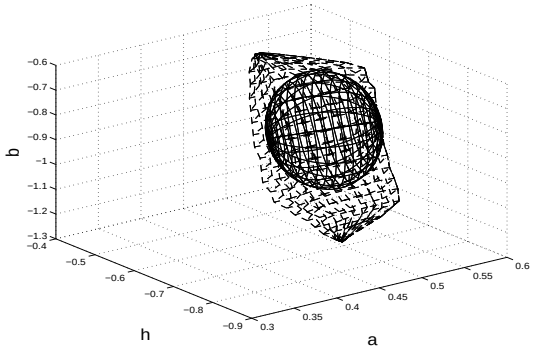

Fig. 5. Asymptotic confidence 90% ellipsoid (-), and the part of the non-asymptotic confidence set around $\bar{\theta}$ (- -).

parameter θ^0 and one around $\bar{\theta}$, the other minimum of $V(\theta)$ when $\tilde{r}_t \equiv 0$. This illustrates the global features of the approach, producing two separate regions far apart in the parameter space as the confidence set.

We also computed the parameter estimate as in Garatti et al. (2004). The parameter estimate turned out to be close to $\bar{\theta}$, and the asymptotic 90% confidence ellipsoid is shown in Figure 5 together with the part of our non-asymptotic confidence set which is concentrated around $\bar{\theta}$. As we can see, the asymptotic theory, due to its local nature, produces a misleading result, since the confidence region is situated around a parameter value corresponding to a local minimum and it does not include the true parameter θ^0. A close up of the part of the non-asymptotic confidence region around the true parameter θ^0 is shown in Figure 6.

6 Conclusions

In this paper, we have derived a new method, Leave-out Sign-dominated Correlation Regions (LSCR), for the construction of confidence sets for general linear models. The method is based on computing empirical correlation functions using subsamples

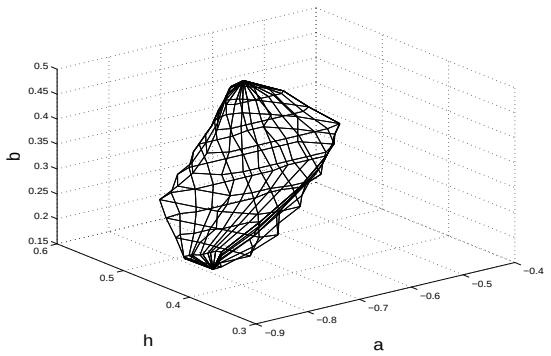

Fig. 6. Close up of the non-asymptotic confidence region around θ^0.

and discarding regions in the parameter space where only a small fraction of the empirical functions are greater or smaller than zero. The developed methodology is grounded on a solid theoretical basis, giving guaranteed probabilities for the true parameter to belong to the constructed set for any finite number of data points, and, as illustrated by the simulation examples, it produces practically useful confidence sets.

The proposed method has a really broad applicability. Even though this has not been discussed in the present paper, the developed method applies unaltered to nonlinear systems as well. Particularly, our Theorem 3.1 is valid without any change for nonlinear systems, provided of course that at point 1 of the procedures $\hat{y}_t(\theta)$ is substituted by the predictor for the nonlinear system structure.

Acknowledgements. This work is partly supported by MIUR under the project "New methods for Identification and Adaptive Control for Industrial Systems". This research has also been supported by the Cooperative Research Centre for Sensor Signal and Information Processing (CSSIP) under the Cooperative Research Centre scheme funded by The Commonwealth Government.

References

[1] Åström K.J. and T. Söderström (1974). "Uniqueness of the maximum likelihood estimates of the parameters of an ARMA model". *IEEE Trans. on Automatic Control*, Vol. 29, no.6, pp. 769-773.

[2] Bai E.W., K.M. Nagpal and R. Tempo (1996). "Bounded-error parameter estimation: noise models and recursive algorithms." *Automatica*, Vol. 32., pp. 985-999.

[3] Bai E.W., R. Tempo and H. Cho (1995). "Membership set estimators: size, optimal inputs, complexity and relations with least squares." *IEEE Trans. on Circuits and Sytems*, Vol. 42., no. 5, pp. 266-277.

[4] Bittanti S., M.C. Campi and S. Garatti (2002). "New results on the asymptotic theory of system identification for the assessment of the quality of estimated models". In *Proc.*

41st Conf. on Decision and Control, Las Vegas, USA.

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

[6] Campi M.C. and S. Garatti (2003). "Correlation Approach for ARMAX Model Identification: a counterexample on the uniqueness of the asymptotic solution". Internal Report of the University of Brescia.

[7] Campi M.C., S.K. Ooi and E. Weyer (2002). "Non-asymptotic quality assessment of generalised FIR models". *Proceedings of IEEE Conference on Decision and Control*, pp. 3416-3421, Las Vegas, USA.

[8] Campi, M.C., S.K. Ooi and E. Weyer (2004). "Non-asymptotic quality assessment of generalised FIR models with periodic inputs". *Automatica*, Vol. 40, no. 12, pp. 2029-2041.

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

[10] Campi M.C. and E. Weyer (2004). "Non-asymptotic confidence sets for the parameters of ARMAX models". *IFAC Workshop on Adaptation and Learning in Control and Signal Processing (ALCOSP 2004)*, pp. 841-846, Yokohama, Japan.

[11] Garatti S., M.C. Campi and S. Bittanti(2003). "Model quality assessment for Instrumental Variable methods: use of the asymptotic theory in practice". In *Proc. 42nd Conf. on Decision and Control*, Maui, USA.

[12] 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., no. 8, pp. 1319-1332.

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

[14] 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., no. 1, pp. 14-23.

[15] Giarre' L., B.Z. Kacewicz and M. Milanese (1997). "Model quality evaluation in set membership identification." *Automatica*, Vol. 33., no. 6, pp. 1133-1139.

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

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

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

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

[20] Ljung, L. (1997). "Identification, model validation and control". In *Proc. 36th Conf. on Decision and Control*, plenary lecture, San Diego, USA.

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

[22] Ljung, L. (1999b). "Model validation and model error models". In Wittenmark B, Rantzer A (Eds) *The Åström Symposium on Control*, Studentlitteratur, Lund, Sweden, pp. 15-42.

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

[24] Milanese M. and A. Vicino (1991). "Optimal estimation theory for dynamic systems with set membership uncertainty: an overview." *Automatica*, Vol. 27., no. 6, pp. 997-1009.

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

[26] Tjärnström F., and L. Ljung (2002). "Using the bootstrap to estimate the variance in the case of undermodelling" *IEEE Transactions on Automatic Control.* Vol. 47, no. 2, pp. 395-398.

[27] 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.

[28] Weyer E. and M.C. Campi (2002). "Non-asymptotic confidence ellipsoids for the least squares estimate." *Automatica*, Vol. 38., no. 9, pp. 1539-1547.

A Proofs

A.1 Proof of Theorem 3.1

A.1.1 Proof of (2)

The proof is divided into a few steps in the form of propositions.

Proposition A.1 *Let $\{w_t\}$ be a sequence of independent random variables with symmetric distribution around zero. Let $I = \{1, \dots, N\}$, and let G be a collection of subsets $I_i \subseteq I$, $i = 1, \dots, M$, forming a group under the symmetric difference operation (i.e. $(I_i \cup I_j) - (I_i \cap I_j) \in G$, if $I_i, I_j \in G$). Pick any $\bar{I} \in G$ and an integer r. Then, the set of variables*

$$\left\{ \sum_{k \in I_i} w_k w_{k+r}, \quad i = 1, \dots, M \right\} \tag{A.1}$$

has the same M-dimensional joint distribution as the set of variables

$$\left\{ \sum_{k \in I_i} w_k w_{k+r} - \sum_{k \in \bar{I}} w_k w_{k+r}, \quad i = 1, \ldots, M \right\}, \tag{A.2}$$

provided that the order of the variables is suitably rearranged. ∎

Before providing the proof, we give a simple example illustrating the idea.

Example. Suppose that $I = \{1, 2, 3, 4\}$, $r = 2$, and $G = \{\{1, 2\}, \{3, 4\}, \emptyset, \{1, 2, 3, 4\}\}$. Take $\bar{I} = \{1, 2\}$. Proposition A.1 says that (by convention, $\sum_{k \in \emptyset} w_k w_{k+r} = 0$):

$$\{w_1 w_3 + w_2 w_4,\ w_3 w_5 + w_4 w_6,\ 0, w_1 w_3 + w_2 w_4 + w_3 w_5 + w_4 w_6\} \tag{A.3}$$

has the same distribution as

$$\{0,\ w_3 w_5 + w_4 w_6 - w_1 w_3 - w_2 w_4, -w_1 w_3 - w_2 w_4,\ w_3 w_5 + w_4 w_6\}. \tag{A.4}$$

Proof. The idea of the proof is to introduce new variables $\tilde{w}_k = -w_k$ for some of the w_k and to rewrite these w_k as $-\tilde{w}_k$ in (A.2) in such a way that the set (A.2) is written as (A.1) with some of the w_k replaced by $\tilde{w}_k$. As w_k is symmetrically distributed around 0, w_k and $\tilde{w}_k$ have the same distribution, and (A.2) and (A.1) will therefore have the same joint M-dimensional distribution. The sign changing procedure introduced below is illustrated in an example after the proof.

Consider the variables

$$w_1 w_{1+r} \quad w_2 w_{2+r} \quad w_3 w_{3+r} \quad \cdots \quad w_N w_{N+r}$$

and organise them in the following chains

$$\begin{aligned}
&w_1 w_{1+r}\ w_{1+r} w_{1+2r}\ w_{1+2r} w_{1+3r}\ \cdots \quad \text{(chain 1)} \\
&w_2 w_{2+r}\ w_{2+r} w_{2+2r}\ w_{2+2r} w_{2+3r}\ \cdots \quad \text{(chain 2)} \\
&\vdots \\
&w_r w_{2r} \quad w_{2r} w_{3r} \qquad w_{3r} w_{4r} \qquad \cdots \text{(chain } r\text{)}.
\end{aligned}$$

We consider one chain at a time, starting with the first one. We scan its elements from left to right. When an element belonging to the set $\{w_k w_{k+r}, k \in \bar{I}\}$ - say $w_{\bar{k}} w_{\bar{k}+r}$ - is encountered, the new variable $\tilde{w}_{\bar{k}+r} = -w_{\bar{k}+r}$ is introduced, and the element is rewritten as $-w_{\bar{k}} \tilde{w}_{\bar{k}+r}$. The next element is then rewritten as $\tilde{w}_{\bar{k}+r} \tilde{w}_{\bar{k}+2r}$ with $\tilde{w}_{\bar{k}+2r} = -w_{\bar{k}+2r}$. So is the next one: $\tilde{w}_{\bar{k}+2r} \tilde{w}_{\bar{k}+3r}$, and we proceed this way until another element in $\{w_k w_{k+r}, k \in \bar{I}\}$ - say $w_{\bar{\bar{k}}} w_{\bar{\bar{k}}+r}$ - is encountered. This element is rewritten as $-\tilde{w}_{\bar{\bar{k}}} w_{\bar{\bar{k}}+r}$, where $\tilde{w}_{\bar{\bar{k}}} = -w_{\bar{\bar{k}}}$, stopping the sequence of sign change. We then proceed scanning the first chain and we start changing the sign again when we

encounter the next element in $\{w_k w_{k+r}, k \in \bar{I}\}$. The procedure terminates when all elements in the first chain have been scanned. Then, we do the same for the other chains.

Next, set $v_k = w_k$ if w_k has not been substituted, and $v_k = \tilde{w}_k$ if w_k has been substituted and consider the variables

$$v_1 v_{1+r} \quad v_2 v_{2+r} \quad v_3 v_{3+r} \quad \cdots \quad v_N v_{N+r}.$$

If $k \in \bar{I}$, we have $v_k v_{k+r} = -w_k w_{k+r}$, while if $k \notin \bar{I}$, $v_k v_{k+r} = w_k w_{k+r}$. Thus, the ith element of (A.2) is given by

$$\begin{aligned} \sum_{k \in I_i} w_k w_{k+r} - \sum_{k \in \bar{I}} w_k w_{k+r} &= \sum_{k \in I_i - \bar{I}} w_k w_{k+r} - \sum_{k \in \bar{I} - I_i} w_k w_{k+r} \\ &= \sum_{k \in I_i - \bar{I}} v_k v_{k+r} + \sum_{k \in \bar{I} - I_i} v_k v_{k+r} = \sum_{k \in I_i \Delta \bar{I}} v_k v_{k+r}. \end{aligned} \tag{A.5}$$

As G is a group under the symmetric difference, $I_i \Delta \bar{I} \in G$, $\forall i$, and hence $\{I_i \Delta \bar{I}, i = 1, \ldots, M\} = \{I_i, i = 1, \ldots, M\}$. This means that (A.2) can be rewritten, by reordering the elements and resorting to (A.5), as

$$\left\{ \sum_{k \in I_i} v_k v_{k+r}, \quad i = 1, \ldots, M \right\}. \tag{A.6}$$

But, for every k, v_k and w_k have the same distribution and thus the set of variables in (A.6) has the same M-dimensional joint distribution as the set of variables in (A.1). ∎

Example. Consider the situation in the example before the proof. We arrange the variables in two chains $w_1 w_3,\ w_3 w_5$ and $w_2 w_4,\ w_4 w_6$. For the first chain, since $1 \in \bar{I}$ we rewrite $w_1 w_3$ as $-w_1 \tilde{w}_3$, $(\tilde{w}_3 = -w_3)$, and $w_3 w_5$ as $\tilde{w}_3 \tilde{w}_5$ $(\tilde{w}_5 = -w_5)$. For the second chain, since $2 \in \bar{I}$ we rewrite the elements as $-w_2 \tilde{w}_4$ and $\tilde{w}_4 \tilde{w}_6$, where we have introduced the new variables $\tilde{w}_4 = -w_4$ and $\tilde{w}_6 = -w_6$. Rewriting the set (A.4) in the new notation we obtain $\{0,\ \tilde{w}_3 \tilde{w}_5 + \tilde{w}_4 \tilde{w}_6 + w_1 \tilde{w}_3 + w_2 \tilde{w}_4,\ w_1 \tilde{w}_3 + w_2 \tilde{w}_4,\ \tilde{w}_3 \tilde{w}_5 + \tilde{w}_4 \tilde{w}_6\}$, which we see is a reordering of (A.3) with $\tilde{w}_i$ substituting w_i for $i = 3, \ldots, 6$.

The next proposition proves that the variables in the set (A.1) exhibit a precise ordering property.

Proposition A.2 *Let $\{w_t\}$ be a sequence of independent random variables with symmetric distribution around zero and such that $Pr\{w_t = c\} = 0$, for any t and any real c. Let $I = \{1, \ldots, N\}$, and let G be a collection of subsets $I_i \subseteq I$, $i = 1, \ldots, M$, forming a group under the symmetric difference operation, and pick an integer r. Then, the set of variables in (A.1) has the following property: each variable in the set has the same probability $1/M$ to be in the j-th position (i.e. there are exactly $j - 1$ other variables in the set (A.1) smaller than the variable under consideration) and this holds for any choice of j between 1 and M.* ∎

The condition $Pr\{w_t = c\} = 0$ in the proposition statement serves the purpose of avoiding ties: it prevents variables from having the same value on sets of nonzero probability. This condition follows e.g. by assuming that the variables w_t admit a density.

If we consider the situation described in the example before the proof of Proposition A.1, Proposition A.2 says that the variables in (A.3) have the same probability of being in a generic j-th position. In other words, if we were asked to bet on one of the variables to be e.g. smaller than all others, our probability of success would not be affected by the choice we make and each of the four variables have exactly probability 1/4 to be the smallest.

Proof. Pick a variable in the set (A.1), say $\sum_{k\in\bar{I}} w_k w_{k+r}$, $\bar{I} \in G$. This variable is in the j-th position if the inequality

$$\sum_{k\in\bar{I}} w_k w_{k+r} > \sum_{k\in I_i} w_k w_{k+r}$$

is satisfied for exactly $j-1$ choices of $I_i \in G$. This is equivalent to say that

$$\sum_{k\in I_i} w_k w_{k+r} - \sum_{k\in\bar{I}} w_k w_{k+r} < 0$$

holds for $j-1$ selections of I_i. Now, using Proposition A.1 we have:

$$\begin{aligned} &Pr\left\{\sum_{k\in I_i} w_k w_{k+r} - \sum_{k\in\bar{I}} w_k w_{k+r} < 0 \quad \text{for } j-1 \text{ selections of } I_i\right\} \\ &= Pr\left\{\sum_{k\in I_i} w_k w_{k+r} < 0 \quad \text{for } j-1 \text{ selections of } I_i\right\}, \end{aligned}$$

showing that the probability of the event on the left hand side does not depend on the chosen $\bar{I}$, since the right hand side does not contain $\bar{I}$. So, any $\bar{I}$ has the same probability that $\sum_{k\in\bar{I}} w_k w_{k+r}$ is in the j-th position and, there being M possible choices of $\bar{I}$, the probability is $1/M$. ∎

We now come to the proof of (2) in Theorem 3.1. Consider the event

$$
\begin{aligned}
A = & \left\{ \sum_{k \in I_i} w_k w_{k+r} < 0 \text{ for at most } q-1 \text{ selections of } I_i \right\} \bigcup \\
& \left\{ \sum_{k \in I_i} w_k w_{k+r} > 0 \quad \text{for at most } q-1 \text{ selections of } I_i \right\} \\
= & \{0 \text{ is in the 1-st or 2-nd or } \cdots \text{ or } q\text{-th position}\} \cup \\
& \{0 \text{ is in the } M\text{-th or } (M-1)\text{-th or } \cdots \text{or } (M-q+1)\text{-th position}\}
\end{aligned}
$$

In view of Proposition A.2 (note that 0 is one variable in set (A.1)),

$$Pr(A) = 2q/M. \tag{A.7}$$

Note now that $w_t = \epsilon_t(\theta^0)$, so that $\sum_{k \in I_i} w_k w_{k+r} = g^{\epsilon}_{i,r}(\theta^0)$ (recall the definition of $g^{\epsilon}_{i,r}(\theta)$ in the "Procedure for the construction of Θ^{ϵ}_r"). Suppose that we have extracted a probabilistic outcome s in A. Then, either $g^{\epsilon}_{i,r}(\theta^0) > 0$ for at most $q-1$ selection of I_i or it is less than 0 for at most $q-1$ selection of I_i, so that $\theta^0 \notin \Theta^{\epsilon}_r$ (recall the construction of Θ^{ϵ}_r). Vice versa, if $s \notin A$, then $g^{\epsilon}_{i,r}(\theta^0) > 0$ for at least q selection of I_i and it is less than 0 again for at least q selection of I_i, yielding $\theta^0 \in \Theta^{\epsilon}_r$. Using (A.7), the conclusion is drawn that $Pr\{\theta^0 \in \Theta^{\epsilon}_r\} = 1 - \frac{2q}{M}$ and the proof is completed.

A.1.2 Proof of (3)

The proof is similar to that of (2), but simpler, and we only sketch the differences.

For Θ^u_s, one observes that a result similar to Proposition A.2 holds showing that each variable in the set $\left\{\sum_{k \in I_i} w_{k+s}, i = 1, \ldots, M\right\}$ has the same probability $1/M$ to be in the generic j-th position. This result can be proven similarly to Proposition A.2, even though the proof is more straightforward. Keeping in mind that $\{u_t\}$ and $\{w_t\}$ are independent, we can treat u_t as if it were deterministic (technically, all derivations can be carried out conditionally on the σ-algebra generated by the process $\{u_t\}$) and hence each variable in the set $\left\{\sum_{k \in I_i} u_k w_{k+s},\ i = 1, \ldots, M\right\}$ has the same probability $1/M$ to be in the generic j-th position as $u_k w_{k+s}$ is also symmetrically distributed around 0. Then, one follows the same reasoning as in the proof of (2) to conclude that $Pr\{\theta^0 \in \Theta^u_s\} = 1 - \frac{2q}{M}$.

A.2 Proof of Theorem 4.1

In the proof, we use the following lemma, taken from Åström and Söderström (1974).

Lemma A.3 *Consider the function*

$$f(z) = \frac{g(z)}{\Pi_{i=1}^{\ell}(z - u_i)^{t_i}}$$

where g is analytic inside and on the unit circle, the numbers u_i are distinct and $t_i \geq 1$. Assume that

$$\oint f(z) z^{k-1} dz = 0, \quad k = 1, \ldots, q,$$

where the integration path is the unit circle and $q = \sum_{i=1}^{\ell} t_i$. Then, f is analytic inside and on the unit circle. ∎

We now turn to the proof of Theorem 4.1. Condition (5) can be re-written as

$$\begin{aligned}
0 &= \int_{-\pi}^{\pi} \Phi_\epsilon(\omega) e^{i\omega r} d\omega \\
&= \int_{-\pi}^{\pi} \left| \frac{A(e^{-i\omega}, \theta)}{C(e^{-i\omega}, \theta)} \frac{C^0(e^{-i\omega})}{A^0(e^{-i\omega})} \right|^2 \lambda_w^2 e^{i\omega r} d\omega \\
&= \oint \frac{z^n A(z^{-1}, \theta) z^p C^0(z^{-1})}{z^p C(z^{-1}, \theta) z^n A^0(z^{-1})} \cdot \frac{A(z, \theta) C^0(z)}{C(z, \theta) A^0(z)} \frac{\lambda_w^2}{i} z^{r-1} dz \\
&= \oint \frac{g(z)}{\Pi_{i=1}^{\ell} (z - u_i)^{t_i}} z^{r-1} dz = 0, \quad r = 1, \ldots, n + p,
\end{aligned}$$

where $g(z) = \frac{z^n A(z^{-1},\theta) z^p C^0(z^{-1}) A(z,\theta) C^0(z)}{C(z,\theta) A^0(z)} \frac{\lambda_w^2}{i}$ is analytic inside and on the unit circle, the numbers u_i are the distinct zeros of $z^p C(z^{-1}, \theta) z^n A^0(z^{-1})$ and t_i is their multiplicity. Then, by applying Lemma A.3 with $q = n + p$, we conclude that $\frac{g(z)}{\Pi_{i=1}^{\ell} (z-u_i)^{t_i}}$ is analytic inside and on the unit circle. In turn, this implies that the zeros of $z^p C(z^{-1}, \theta) z^n A^0(z^{-1})$ - which are all inside the unit circle - are cancelled by those of $z^n A(z^{-1}, \theta) z^p C^0(z^{-1})$. Since $z^n A^0(z^{-1})$ and $z^p C^0(z^{-1})$ have no common zeros, this gives $C(z^{-1}, \theta) = C^0(z^{-1})$ and $A(z^{-1}, \theta) = A^0(z^{-1})$, concluding the proof.

A.3 Proof of Theorem 4.2

In order to avoid notational cluttering, in this proof we write A for $A(z^{-1}, \theta)$ or $A(e^{-i\omega}, \theta)$ (where the argument can be deduced from the context), and the same convention applies to all other polynomials. Moreover, '$^-$' denotes complex conjugation, so that $\bar{A} = A(e^{i\omega}, \theta)$.

Note first that

$$\epsilon_t(\theta) = \frac{A}{C} y_t - \frac{B}{C} u_t = \frac{AB^0}{CA^0} u_t + \frac{AC^0}{CA^0} w_t - \frac{B}{C} u_t = \frac{AB^0 - BA^0}{CA^0} u_t + \frac{AC^0}{CA^0} w_t.$$

Since $\{u_t\}$ and $\{w_t\}$ are independent, condition (6) can be written as

$$
\begin{aligned}
0 &= \int_{-\pi}^{\pi} \frac{AB^0 - BA^0}{CA^0} \lambda_u^2 e^{i\omega s} d\omega \\
&= \oint \frac{z^n A z^m B^0 - z^m B z^n A^0}{z^p C z^n A^0} \frac{\lambda_u^2}{i} z^{s+p-m-1} dz \quad s = 1, \ldots, n+m \\
&= \oint \frac{z^n A z^m B^0 - z^m B z^n A^0}{z^p C z^n A^0} \frac{\lambda_u^2}{i} z^{\tilde{s}-1} dz, \quad \tilde{s} = 1+p-m, \ldots, n+p. \qquad \text{(A.8)}
\end{aligned}
$$

We distinguish two cases: $p < m$ and $p \geq m$.

Case 1: $\mathbf{p < m}$. From (A.8), we have

$$
\oint \frac{g(z)}{\Pi_{i=1}^{\ell}(z-u_i)^{t_i}} z^{\tilde{s}-1} dz = 0, \quad \tilde{s} = 1, \ldots, n+p,
$$

where $g(z) = (z^n A z^m B^0 - z^m B z^n A^0)\frac{\lambda_u^2}{i}$, the numbers u_i are the distinct zeros of $z^p C z^n A^0$ and t_i is their multiplicity. Taking $\tilde{s} = 1, \ldots, n+p$, we have voluntarily disregarded the equations in (A.8) valid for $\tilde{s} = 1+p-m, \ldots, 0$. Then, by applying Lemma A.3 with $q = n+p$, we conclude that $\frac{g(z)}{\Pi_{i=1}^{\ell}(z-u_i)^{t_i}}$ is analytic inside and on the unit circle. In turn, this implies that $g(z)$ has the following structure:

$$
g(z) = z^p C z^n A^0 \cdot \sum_{j=0}^{m-p-1} \alpha_j z^j. \qquad \text{(A.9)}
$$

Indeed, the first factor $z^p C z^n A^0$ has to be present since the zeros of the denominator $z^p C z^n A^0$ - that are all inside the unit circle - must be cancelled by zeros in the numerator, while the second factor with unknown coefficients α_j's is introduced to equalize the degree of $g(z) = (z^n A z^m B^0 - z^m B z^n A^0)\frac{\lambda_u^2}{i}$ (which is $n+m-1$ - remember that $z^m B = b_1 z^{m-1} + \cdots + b_m$ has degree $m-1$) with that of (A.9).

We now go to consider the so far neglected equations in (A.8) valid for $\tilde{s} = 1+p-m, \ldots, 0$. Taking into account the expression of $g(z)$ in (A.9), we have:

$$
\oint \sum_{j=0}^{m-p-1} \alpha_j z^{j+\tilde{s}-1} dz = (2\pi i)\alpha_{-\tilde{s}} = 0, \quad \tilde{s} = 1+p-m, \ldots, 0,
$$

which gives $\alpha_j = 0, j = 0, \ldots, m-p-1$, so concluding that $g(z) = 0$. Finally, recalling that $g(z) = (z^n A z^m B^0 - z^m B z^n A^0)\frac{\lambda_u^2}{i}$ and that A^0 and B^0 have no common factors, this yields $A = A^0$ and $B = B^0$.

Case 2: $\mathbf{p \geq m}$. Consider the equations in (A.8), valid for $\tilde{s} = 1+p-m, \ldots, n+p$. If $p = m$, $\tilde{s}$ starts from 1. If $p > m$, $\tilde{s}$ starts from $1+p-m > 1$. Yet, the range of validity of (A.8) can be extended to $\tilde{s} = 1, \ldots, n+p$ in the latter case too. The reason is that for $\tilde{s} = 1, \ldots, p-m$, the degree of the numerator of the integrand in (A.8) is no larger than $n+p-2$ while the degree of the denominator is $n+p$. Since the denominator has all its zeros inside the unit circle, we can then express

the integrand as a Taylor expansion $\sum_{j=2}^{\infty} \beta_j z^{-j}$, where convergence takes place in a co-disk containing the unit circle. Integrating such a series gives the desired result that the integral is zero.

Now, from the equations

$$0 = \oint \frac{z^n A z^m B^0 - z^m B z^n A^0}{z^p C z^n A^0} \frac{\lambda_u^2}{i} z^{\tilde{s}-1} dz, \quad \tilde{s} = 1, \ldots, n+p,$$

by virtue of Lemma A.3 we again conclude that $\frac{z^n A z^m B^0 - z^m B z^n A^0}{z^p C z^n A^0}$ is analytic inside and on the unit circle. But, the numerator has degree $n+m-1 \leq n+p-1$ which is smaller than the degree $n+p$ of the denominator, so the only possibility for this function to be analytic inside and on the unit circle is that it is zero. The conclusion that $A = A^0$ and $B = B^0$ then follows as for the case where $p < m$.

We now turn to conditions (7). Since, as we have proved, $A = A^0$, they write:

$$0 = \oint \frac{z^p C^0}{z^p C} \cdot \frac{\bar{C}^0}{\bar{C}} \frac{\lambda_w^2}{i} z^{r-1} dz, \quad r = 1, \ldots, p.$$

The fact that $C = C^0$ follows along the same line as in the proof of Theorem 4.1. This concludes the proof.

A.4 Proof of Theorem 4.3

u_t has spectrum $\Phi_u(\omega) = Q(e^{-i\omega})Q(e^{i\omega})\lambda_\nu^2$ and $L(z^{-1}) = Q(z^{-1})^{-1}Q(z)^{-1}$ so that condition (8) gives

$$\begin{aligned} 0 &= \int_{-\pi}^{\pi} \frac{AB^0 - BA^0}{CA^0} \bar{L}\Phi_u e^{i\omega s} d\omega = \int_{-\pi}^{\pi} \frac{AB^0 - BA^0}{CA^0} \bar{L}Q\bar{Q}\lambda_\nu^2 e^{i\omega s} d\omega \\ &= \int_{-\pi}^{\pi} \frac{AB^0 - BA^0}{CA^0} \lambda_\nu^2 e^{i\omega s} d\omega, \end{aligned}$$

and the rest follows as in the proof of Theorem 4.2.

A.5 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

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

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