Non-Asymptotic Confidence Sets for the Parameters of Linear Transfer Functions

Marco C. Campi, *Senior Member, IEEE*, and Erik Weyer, *Member, IEEE*

***Abstract*—We consider the problem of constructing confidence sets for the parameters of input-output transfer functions based on observed data. The assumptions on the noise affecting the system are reduced to a minimum; the noise can virtually be anything, but in return the user must be able to select the input signal. In this paper a procedure for solving this problem is developed in the general framework of leave-out sign-dominant confidence regions. The procedure returns confidence regions that are guaranteed to contain the true transfer function with a user-chosen probability for any finite data set.**

***Index Terms*—Confidence regions, finite sample results, linear systems, system identification, transfer function estimation.**

I. Introduction

A model is never a perfect description of a real system. Consequently, it is important to determine if the uncertainty in the model is within limits that can be tolerated by the application at hand. Developing methods and procedures for assessing the model quality is therefore a central issue in system identification.

In this paper we present a procedure which gives rigorously guaranteed non-asymptotic confidence regions for the parameters of a linear dynamical system which is affected by arbitrary noise. The procedure consists of a simple input design step, followed by an algorithm named LSCR (Leave-out Sign-dominant Correlation Regions) which constructs the confidence set from the observed data points. The procedure can be applied and the results are valid for any finite number of data points.

Problem Addressed in This Paper: Consider a dynamical system

$$y_t = G^0(z^{-1})u_t + v_t$$

as in Fig. 1. The transfer function $G^0(z^{-1})$ belongs to a set of transfer functions $G(\theta, z^{-1})$ parameterized by θ, that is $G^0(z^{-1}) = G(\theta^0, z^{-1})$ for some θ^0. The structure of the model class $G(\theta, z^{-1})$ is known, but θ^0 itself is unknown. The noise v_t describes all other sources, apart from u_t, that cause variation in y_t, and v_t is independent of u_t.

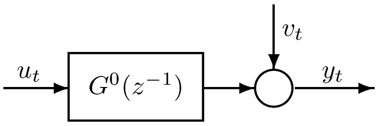

Fig. 1. Dynamical system.

We have access to the system for experimentation. The problem consists in using a finite number of input and output data collected at time $t = 1, 2, \ldots, N$ to determine a confidence region $\hat{\Theta}$ for θ^0. The region $\hat{\Theta}$ must contain θ^0 with a given probability chosen by the user, and moreover $\hat{\Theta}$ must be constructed without any a-priori knowledge of the strength, distribution or correlation pattern of the noise.

In order to put our results into perspective we next review some common methods for construction of confidence regions, before providing more details in the form of a preview example on the approach developed in this paper.

Asymptotic Theory: A standard approach to deriving confidence regions is to employ asymptotic system identification theory (see, e.g., [1] or [2]). Although used with success in many applications, asymptotic results are only guaranteed when the number of data points N tends to infinity. When the number of data points is finite, asymptotic theory may generate misleading results, even for large data sets, see, e.g., [3], [4]. Moreover, the asymptotic theory looses relevance for scarce data sets.

In order to make practical use of the asymptotic theory a full description of the system is required, that is both $G^0(z^{-1})$ as well as the noise transfer function $H^0(z^{-1})$ ($v_t = H^0(z^{-1})e_t$, e_t white) must belong to the specified model classes.

Set Membership Identification: In set membership identification, e.g., [5]–[10], the parameter confidence regions are guaranteed for any finite N. No noise model is necessary, and unmodeled dynamics in $G^0(z^{-1})$ is also allowed. All system components that are not described by $G^0(z^{-1})$ are assumed to be *bounded by known constants*, and the parameter region is determined as the set of parameter values that do not invalidate the *a priori* bounds. No probabilistic framework is assumed in this approach.

Bootstrap: The underlying idea in bootstrap methods used in system identification (e.g., [11] and [12]) is that the prediction errors evaluated at the estimated parameters are representative for the behaviour of the noise process. The effect of the noise on the estimated model is then assessed by generating new noise sequences by random resampling from the prediction errors. For this approach to apply, the noise must be modelled in full so that the prediction error becomes white. Moreover, while asymptotic properties of bootstrap have been studied, e.g., in [13], finite sample results are substantially lacking.

Manuscript received December 05, 2007; revised January 19, 2009, July 10, 2009, December 03, 2009, and April 14, 2010; accepted April 22, 2010. First published April 29, 2010; current version published December 02, 2010. This work was supported by the Australian Research Council, under Discovery Grant DP0558579 and the MIUR project "New methods for Identification and Adaptive Control for Industrial Systems." Recommended by Associate Editor T. Zhou.

M. C. Campi is with the Department of Information Engineering, University of Brescia, Brescia 25123, Italy (e-mail: marco.campi@ing.unibs.it).

E. Weyer is with the Department of Electrical and Electronic Engineering, The University of Melbourne, Parkville, VIC 3010, Australia (e-mail: ewey@unimelb.edu.au).

Digital Object Identifier 10.1109/TAC.2010.2049416

Leave-Out Sign-Dominant Confidence Regions: In [14], it was recognized that a methodology able to work out confidence regions *without assuming a-priori knowledge on the noise level* was highly desirable. A procedure called leave-out sign-dominant correlation regions (LSCR) was developed. LSCR as introduced in [14] is able to return guaranteed confidence regions for any finite number of data points under two assumptions: **i)** both $G^0(z^{-1})$ and $H^0(z^{-1})$ belong to specified classes of transfer functions; **ii)** the noise entering $H^0(z^{-1})$ is zero mean and symmetrically distributed.

The Method of the Present Paper: In this paper, it is assumed that we can select the input u_t.[1] However, it can be subject to constraints, such as bounds on the magnitude, $|u_t| \leq U$, and the experimentation time may be limited. Under these conditions a procedure within the LSCR framework is designed with the following properties:

i) the procedure works for any noise v_t and no a-priori knowledge on the noise characteristics is required. The noise can be white or correlated, zero mean or biased, the signal-to-noise ratio can be high or low;

ii) for any size N of the data sample, θ^0 belongs to the constructed confidence region $\hat{\Theta}$ with a guaranteed probability specified by the user;

iii) $\hat{\Theta}$ shrinks around θ^0 for increasing N.

Thus, the method presented in this paper is a significant step forward from the results in [14] in that it requires no assumptions on the noise or noise model. However, differently from [14], we do assume that the system input can be designed. In this paper we also derive expressions for the shape and size of the confidence sets which shed light on the role of user chosen quantities such as the input signal. A further aspect worth noting is that in this paper unlike [14] we introduce and employ stochastic strings in the LSCR algorithm, a choice that offers significant computational advantages.

Property (i) is important since noise characteristics are hardly known in practice. v_t can, e.g., describe external influences generated by other systems, measurement noise, etc.; often it is even difficult to figure out what all the external influences are. Note that all possible noise situations are encompassed by the system in Fig. 1 since, under linearity, noise can be recast as additive at the output without any loss of generality. Property (ii) is fundamental for the use of the results in robust design problems since the number of data points used for model building is always finite. Property (iii) is a desirable and natural property as it allows us to make stronger and stronger claims about the nature of the system as the number of data points increases.

Further discussions and earlier results on finite sample properties of system identification methods can be found in [15]–[19].

A. Preview Example

A simple example serves the purpose of giving a preview of the main ideas behind the method introduced in this paper. Consider the system

$$y_t = b^0 u_t + v_t. \tag{1}$$

We have no information about the noise process v_t, other than that v_t is independent of u_t.

Our goal is to generate $N = 15$ input data $u_1, \ldots, u_{15}$ and to construct a confidence interval $\hat{\Theta}$ from the observed output data such that $Pr\{b^0 \in \hat{\Theta}\} = 0.80$. Next we describe the input design and the procedure for constructing $\hat{\Theta}$ followed by simulation results. The proof of why the confidence region $\hat{\Theta}$ has the property $Pr\{b^0 \in \hat{\Theta}\} = 0.80$ is given later in Theorem 1.

Input Design: Let u_t, $t = 1, \ldots, 15$, be independent and identically distributed (iid) with distribution

$$u_t = \begin{cases} 1 & \text{with probability } 1/2 \\ -1 & \text{with probability } 1/2. \end{cases} \tag{2}$$

Procedure for Construction of the Confidence Interval $\hat{\Theta}$: Rewrite the system as a model with generic parameter b

$$y_t = bu_t + v_t.$$

By ignoring the noise term we obtain a predictor and a prediction error

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

Compute the prediction errors $\epsilon_t(b) = y_t - bu_t$ for $t = 1, \ldots, 15$ as a function of b using the observed data $u_1, \ldots, u_{15}, y_1, \ldots, y_{15}$ and calculate

$$f_t(b) = u_t \epsilon_t(b), \quad t = 1, \ldots, 15.$$

Using the $f_t(b)$'s, we want to compute empirical estimates of the correlation $E[u_t \epsilon_t(b)]$. Note that $E[u_t \epsilon_t(b)] = (b^0 - b)E[u_t^2] + E[u_t v_t] = (b^0 - b)$. Hence, the empirical estimates will be zero mean random variables for $b = b^0$. 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) values "too many" times.

We select 19 subsets of data at random and compute the empirical estimates

$$g_i(b) = \sum_{t=1}^{15} h_{i,t} \cdot f_t(b), \quad i = 1, \ldots, 19$$

where $h_{i,t}$ are iid with distribution

$$h_{i,t} = \begin{cases} 0 & \text{with probability } 1/2 \\ 1 & \text{with probability } 1/2 \end{cases} \tag{3}$$

i.e., $h_{i,t}$ determines if $f_t(b)$ is used when we compute the ith estimate of the correlation. (The estimates are actually scaled estimates, but this is of no consequence as only the sign is used in the procedure.)

Next we plot $g_i(b)$, $i = 1, \ldots, 19$, as functions of b. Since it is very unlikely that all the $g_i(b^0)$'s have the same sign, we discard the regions where none or only one function is less than zero or greater than zero. The resulting interval is the confidence region for b^0.

Simulation Results: **(A)** The value $b^0 = 1$ was used, and v_t was an independent sequence of normally distributed variables

[1] Extensions to the case when u_t is not user-chosen are briefly outlined in Section IV-A.

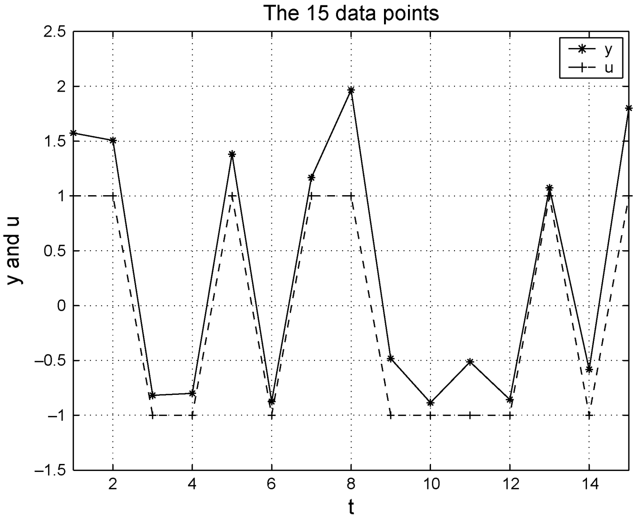

Fig. 2. Input-output data—case (A).

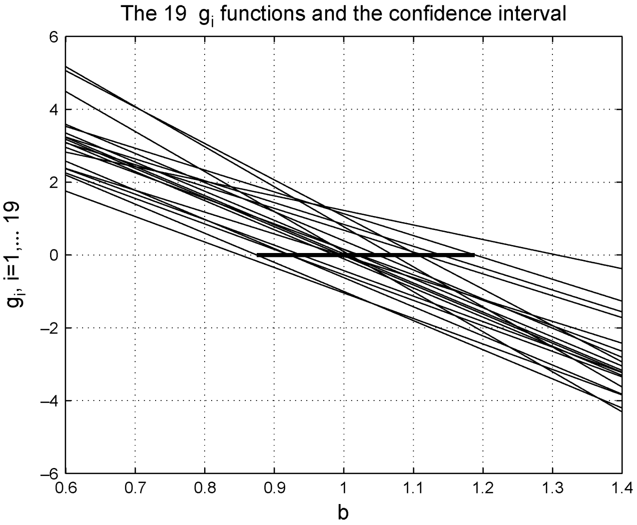

Fig. 3. $g_i(b)$ functions together with the confidence interval—case (A).

with mean 0.5 and variance 0.1, i.e., the noise was biased. The input-output data are shown in Fig. 2, and the functions $g_i(b)$, $i = 1, \dots, 19$, are plotted in Fig. 3. We then discarded the values of b for which none or only one function was greater/smaller than 0. As shown in Fig. 3, the resulting confidence interval for b^0 is [0.874, 1.119].

No knowledge about the noise v_t was used in the construction of the confidence interval, yet it is a rigorous fact (stated in Theorem 1) that the so-constructed confidence interval has exact probability $1 - 2 \cdot 2/20 = 0.8$ of containing the true parameter value b^0. Despite the fact that the noise is biased, the procedure provides a guaranteed confidence interval.

As expected, due to the small number of data points, this confidence interval is rather large and the associated probability is low. Next we repeated the experiment with 2000 data points, and computed 999 empirical estimates of the correlation using random subsets. We kept the values of b where at least 25 of the 999 functions are greater than 0 and at least 25 are smaller than zero. The resulting interval was $\hat{\Theta} = [0.983, 1.024]$. This procedure returns a $\hat{\Theta}$ that contains the true parameter value b^0 with exact probability $1 - 25 \cdot 2/1000 = 0.95$ (see Theorem 1).

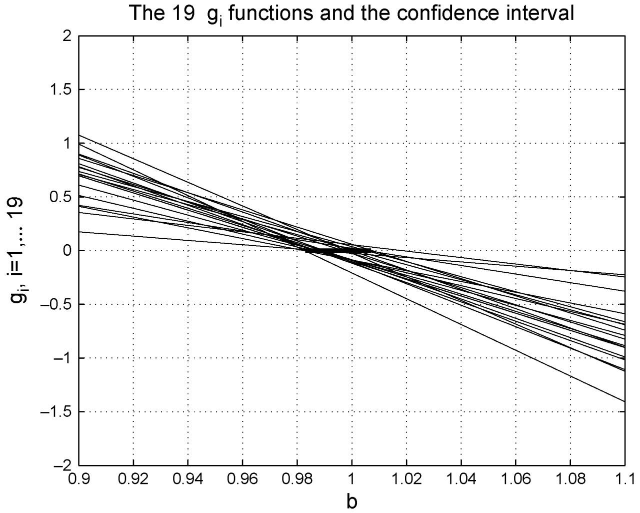

Fig. 4. $g_i(b)$ functions together with the confidence interval—case (B).

(B) As before $b^0 = 1$ and $N = 15$, but this time v_t was normally distributed with mean 0 and variance 0.001, that is very little noise affects the data. The $g_i(b)$ functions are shown in Fig. 4 and the 80% confidence interval narrows down to $\hat{\Theta} = [0.983, 1.007]$. We see that when the noise is reduced the interval becomes smaller and this is achieved automatically, without making any *a priori* assumption on the noise level.

Remark 1: What is crucial in the above example is that a certain sign property of the $g_i(b^0)$'s is valid irrespective of the characteristics of the noise v_t, and this is what makes the procedure valid regardless of what the noise is. Certainly, the level and the type of noise impact on the final result (i.e., $\hat{\Theta}$ will be larger or smaller depending on the noise characteristics.) What is important is that the procedure can be implemented without any a-priori knowledge on the noise. To state it simply, the procedure lets the data speak without assuming what they have to tell us. □

B. Organization of the Paper

For the sake of clarity we first present the basic version of our procedure which is applicable to systems where the transfer function between input and output $G^0(z^{-1}) = B^0(z^{-1})/A^0(z^{-1})$ is parameterized in terms of the numerator and denominator coefficients. All the main ideas and key technical points in the proofs are introduced in this part. Then, we extend the setting to a larger model class, and we introduce a more general version of the procedure using filtered signals.

In Sections II and III we introduce the data generating system and we present the basic version of our algorithm. Moreover, in Section III we also prove that the constructed confidence sets have guaranteed probability for any finite number of data points, and we show that the confidence sets shrink around the true parameter as the number of data points increases. In Section IV-A we present the extended version of the algorithm, and in Section IV-B we show how user choices in the algorithm affect the shape of the obtained confidence regions. Conclusions are given in Section V.

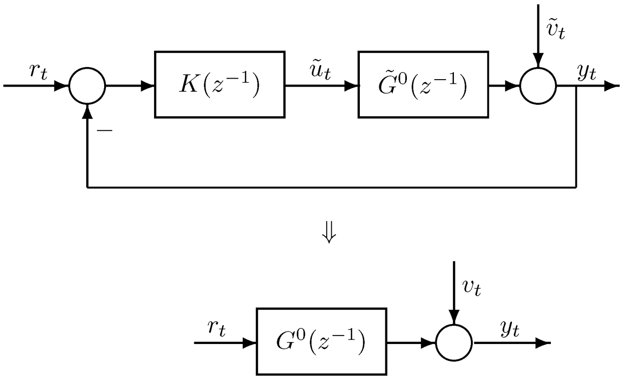

Fig. 5. Closed loop system recast as an open loop system.

II. Data Generating System

The data generating system is given by

$$y_t = G^0(z^{-1})u_t + v_t \tag{4}$$

where $G^0(z^{-1}) = B^0(z^{-1})/A^0(z^{-1})$ where z^{-1} is the backward shift operator $(z^{-1}u_t = u_{t-1})$ and

$$\begin{aligned} A^0(z^{-1}) &= 1 + a_1^0 z^{-1} + \cdots + a_{n_a^0}^0 z^{-n_a^0}, \\ B^0(z^{-1}) &= b_1^0 z^{-1} + \cdots + b_{n_b^0}^0 z^{-n_b^0}. \end{aligned}$$

Assumptions:

1) The user can choose the input signal u_t, and the choice of u_t does not affect v_t.
In mathematical terms, u_t and v_t are independent.
2) Upper bounds n_a and n_b ($n_a \geq n_a^0, n_b \geq n_b^0$) on the model orders are known.

There are no assumptions on v_t. No upper bound on its magnitude is assumed, and it is allowed to have non-zero mean and any autocorrelation properties.

Remark 2: There is no loss of generality in having v_t additive at the output. Disturbances not entering at the output can for example be represented by $v_t = H^0(z^{-1})e_t$ where e_t is the real disturbance. □

Remark 3: Assumption 1 entails an open loop configuration. Closed loop systems can be cast in the present setting as shown in Fig. 5 where r_t plays the role of u_t and

$$G^0(z^{-1}) = \frac{\tilde{G}^0(z^{-1})K(z^{-1})}{1 + \tilde{G}^0(z^{-1})K(z^{-1})} \tag{5}$$

and $v_t = (1/(1 + \tilde{G}^0(z^{-1})K(z^{-1})))\tilde{v}_t$.

Assuming that $K(z^{-1})$ is known, (5) provides an expression for $G^0(z^{-1})$ as a function of $\tilde{G}^0(z^{-1})$. If a parameterized model of the unknown $\tilde{G}^0(z^{-1})$ transfer function is substituted in this expression, a model of $G^0(z^{-1})$ parameterized in terms of the unknown parameters of $\tilde{G}^0(z^{-1})$ is obtained, and we can apply LSCR to this model. The functional dependence of $G^0(z^{-1})$ on the unknown parameters can possibly be complex, but this is not a problem since the LSCR algorithm does not require any particular parameterization of the transfer function. When applying the LSCR algorithm to the system in the lower part of Fig. 5, r_t is selected by the user in the same way as u_t was selected in the original system in Fig. 1. □

III. Procedure for Construction of Confidence Regions

Our goal is to construct confidence regions for the transfer function between the input u_t and the output y_t. The procedure developed below consists of an easy input design step, followed by an algorithm that constructs the region from the observed data. Let

$$\begin{aligned} A(z^{-1}, \theta) &= 1 + a_1 z^{-1} + \cdots + a_{n_a} z^{-n_a} \\ B(z^{-1}, \theta) &= b_1 z^{-1} + \cdots + b_{n_b} z^{-n_b} \end{aligned}$$

and let $n := n_a + n_b$.

Input design.

Let u_t, $t = 1 - n, \ldots, N$, be an iid sequence of random variables symmetrically distributed around 0. u_t is applied to the system, and the initial conditions of the system are arbitrary.

Construction of confidence sets.

1) Using the data, compute the predictions[2] as function of θ:

$$\hat{y}_t(\theta) = \left(1 - A(z^{-1}, \theta)\right) y_t + B(z^{-1}, \theta)u_t = \phi_t^T \theta, \qquad t = 1, \ldots, N$$

where

$$\begin{aligned} \phi_t &= [-y_{t-1}, \ldots, -y_{t-n_a}, u_{t-1}, \ldots, u_{t-n_b}]^T \\ \theta &= [a_1, \ldots, a_{n_a}, b_1, \ldots, b_{n_b}]^T . \end{aligned}$$

2) Compute the prediction errors

$$\begin{aligned} \epsilon_t(\theta) &= y_t - \hat{y}_t(\theta) = A(z^{-1}, \theta)y_t - B(z^{-1}, \theta)u_t \\ &= y_t - \phi_t^T \theta, \quad t = 1, \ldots, N. \end{aligned}$$

3) Form the vector

$$\xi_t = [u_{t-1}, \ldots, u_{t-n}]^T, \quad t = 1, \ldots, N$$

and compute the vector

$$f_t(\theta) = \xi_t \epsilon_t(\theta), \quad t = 1, \ldots, N.$$

4) Select an integer M and construct M binary ({0,1}-valued) stochastic strings of length N as follows: Let $h_{0,1}, \ldots, h_{0,N} = 0, \ldots, 0$ be the string of all zeros. Every element of the remaining strings takes the values 0 or 1 each with probability 0.5, and the elements are independent of each other. Moreover, each string is constructed independently of previous strings. However, if a string turns out to be equal to an already constructed string, this string is removed and another string to be used in its place is constructed according to the same rule. Name the constructed non-zero strings $h_{1,1}, \ldots, h_{1,N}$; $h_{2,1}, \ldots, h_{2,N}; \ldots; h_{M-1,1}, \ldots, h_{M-1,N}$. Compute

$$g_i(\theta) = \sum_{t=1}^{N} h_{i,t} \cdot f_t(\theta) = \sum_{t=1}^{N} h_{i,t} \cdot \xi_t \epsilon_t(\theta), \quad i = 0, \ldots, M-1.$$

Note that $g_0(\theta) \equiv 0$.

[2] The predictors are obtained from (4) by ignoring v_t. The predictors are not the one step ahead predictors commonly used in system identification as these predictors require more knowledge about the noise than is available in the present setting.

5) Let $g_i^k(\theta)$ denote the kth element of the vector $g_i(\theta)$, $k=1,\dots,n$. Select an integer q in the interval $[1, M/2n]$. Construct the regions $\hat{\Theta}_N^k$ such that at least q of the $g_i^k(\theta)$ functions are strictly larger than $g_0^k(\theta) \equiv 0$ and at least q are strictly smaller than $g_0^k(\theta) \equiv 0$. The confidence set is given by

$$\hat{\Theta}_N = \cap_{k=1}^N \hat{\Theta}_N^k. \tag{6}$$

Remark 4: The algorithm above has connections with Instrumental Variable (IV) methods for system identification. The main idea behind IV methods is that the prediction errors should be uncorrelated with past data. By taking ξ_t in point 3 above as the instrumental variable, the estimate would be given by

$$\hat{\theta}_N = \left\{ \theta \text{ such that } \sum_{t=1}^N \xi_t \epsilon_t(\theta) = 0 \right\}.$$

In our approach, the confidence set $\hat{\Theta}_N$ is constructed by excluding the regions in parameter space where the components of $\sum_{t=1}^N h_{i,t} \cdot \xi_t \epsilon_t(\theta)$ takes on positive or negative values too many times. □

Remark 5: One aspect that deserves some further comments is how the region $\hat{\Theta}_N^k$ should be constructed in practice. One can easily ascertain whether a given θ value belongs to $\hat{\Theta}_N^k$ by simply inspecting the sign of functions $g_i^k(\theta)$. This suggests that $\hat{\Theta}_N^k$ can be constructed by exploring a grid of θ values. This method is practical for problems where θ has few elements, but it becomes computationally intensive when θ has many elements. Finding suitable ways to construct regions $\hat{\Theta}_N^k$ at low computational effort is an important topic for future research. □

Let θ^0 be the vector of the coefficients of the system in (4), where suitable zeros have been added to match the size of θ^0 with that of θ. The intuitive idea behind the algorithm is that for $\theta = \theta^0$ the functions $g_i^k(\theta) = \sum_{t=1}^N h_{i,t} \cdot u_{t-k} \epsilon_t(\theta)$ take on positive and negative values at random since $g_i^k(\theta^0) = \sum_{t=1}^N h_{i,t} \cdot u_{t-k} A^0(z^{-1}) v_t$ where u_{t-k} are symmetrically distributed around 0. It is therefore unlikely that only a small fraction of them are positive or negative, and point 5 in the algorithm excludes the regions in parameter space where this happens. The probability that θ^0 belongs to each of the $\hat{\Theta}_N^k$ is given in the next theorem.

Theorem 1: Consider a $k \in \{1,\dots,n\}$ and assume that $Pr\{g_i^k(\theta^0) = 0\} = 0$, $i \neq 0$. Then

$$Pr\left\{\theta^0 \in \hat{\Theta}_N^k\right\} = 1 - 2q/M \tag{7}$$

where $\hat{\Theta}_N^k$ is constructed in point 5 of the algorithm above, and q and M are introduced in points 5 and 4 respectively. □

Proof: See Appendix A-A. ■

An immediate consequence of Theorem 1 is

Corollary 2: Under the assumptions in Theorem 1

$$Pr\{\theta^0 \in \hat{\Theta}_N\} \geq 1 - 2nq/M \tag{8}$$

where $\hat{\Theta}_N$ is given by (6). □

Note that the probability in (7) is the exact probability, not a lower bound. The inequality in (8) is due to that the events $\{\theta^0 \notin \hat{\Theta}_N^k\}$, $k = 1,\dots,n$, may overlap.

The only reason for the assumption $Pr\{g_i^k(\theta^0) = 0\} = 0$ is to prevent ties with the zero function $g_0(\theta) \equiv 0$ at $\theta = \theta^0$ from occurring in point 5 of the procedure. This assumption is mild. It is for example satisfied whenever u_t and v_t admit densities. One particular case of some interest where the assumption is not satisfied is when there is no noise, in which case all the $g_i^k(\theta)$ functions are 0 for $\theta = \theta^0$. Such ties with the zero function $g_0^k(\theta)$ can be broken by *a priori* assigning a random ordering to the $g_i^k(\theta)$, $i = 0, 1, \dots, M-1$, functions and by resorting to this ordering whenever one of the non-zero functions $g_i^k(\theta)$, $i = 1, 2, \dots, M-1$, takes on the value 0. Using this ordering permits one to drop the assumption $Pr\{g_i^k(\theta^0) = 0\} = 0$, and an inspection of the proof reveals that the above theorem holds true under this generalization. However, we have chosen to keep the assumption in order to simplify the presentation.

Theorem 1 and Corollary 2 show that the constructed confidence sets have guaranteed probability. The next Theorem 3 shows that the confidence sets concentrate around the true parameter θ^0 as the number of data points increases.

Theorem 3: In addition to the assumptions of Theorem 1, assume that

1) $z^{n_a} A^0(z^{-1})$ and $z^{n_b} B^0(z^{-1})$ have no common factors;
2) $A^0(z^{-1})$ is asymptotically stable;
3) $|u_t| \leq U$ for some U; moreover, $|v_t| \leq K t^\alpha$ with probability 1 for some K and $\alpha < 1/2$, where K and α are allowed to depend on the noise realization.

Suppose that an initial confidence set $\hat{\Theta}_{\bar{N}}$ is obtained from the algorithm and then this set is updated as new data are observed by extending the stochastic strings $h_{i,t}$ with new $\{0,1\}$ random variables for $t = \bar{N}+1, \bar{N}+2, \dots$ while M is kept fixed. Then, for all $\epsilon > 0$

$$Pr\left\{\exists N(\epsilon) \text{ such that } \hat{\Theta}_N \subseteq \{\theta : \|\theta - \theta^0\| \leq \epsilon\} \right. \\ \left. \text{for all } N > N(\epsilon)\right\} = 1$$

that is, there exists a realization dependent $N(\epsilon)$ such that the confidence set is included in an ϵ-neighborhood of θ_0 for all $N > N(\epsilon)$. □

Proof: See Appendix A–B. ■

The first assumption is a standard one ensuring that the system has a unique representation within the model class. The assumption on v_t is very mild since v_t is allowed to grow unbounded. It is for example satisfied if v_t is white Gaussian noise. The assumption that u_t is bounded can be relaxed but we have decided to maintain it since it appears natural and simplifies the proof.

A. Simulation Example

In this section, we show through an example how the obtained confidence sets depend on the noise characteristics and the number of data points. The system is given by

$$y_t = -a^0 y_{t-1} + b^0 u_{t-1} + \tilde{v}_t \tag{9}$$

where $a^0 = -0.7$ and $b^0 = 0.3$. This corresponds to $G^0(z^{-1}) = b^0 z^{-1}/(1 + a^0 z^{-1})$ $=0.3z^{-1}/(1 - 0.7z^{-1})$ and $v_t = (1/(1 + a^0 z^{-1}))\tilde{v}_t = (1/(1 - 0.7z^{-1}))\tilde{v}_t$ in (4). u_t is independent and uniformly distributed on $[-\sqrt{3}, \sqrt{3}]$, i.e., it is zero mean with variance 1.

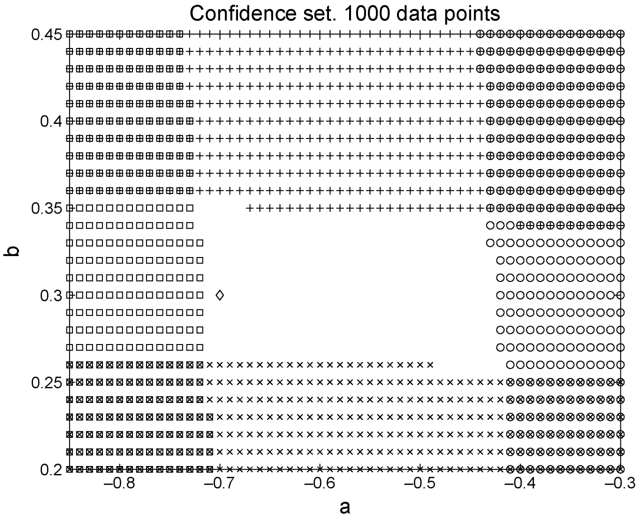

Fig. 6. Non-asymptotic confidence region for (a^0, b^0) (blank region). 1000 data points. $\tilde{v}_t$ is lowpass filtered white noise. $\diamond = \mathrm{true\ parameter}$.

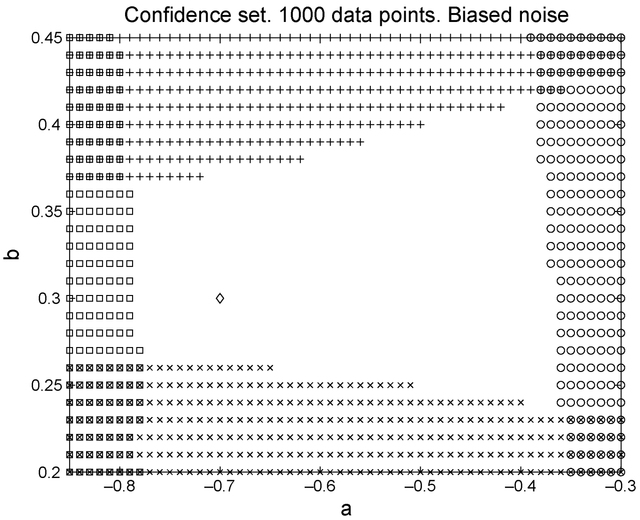

Fig. 7. Non-asymptotic confidence region for (a^0, b^0) (blank region). 1000 data points. $\tilde{v}_t$ is biased lowpass filtered white noise. $\diamond = \mathrm{true\ parameter}$.

For $\tilde{v}_t$ we consider two different processes:

A) $\tilde{v}_t^A$ is lowpass filtered white noise

$$\tilde{v}_t^A = c_0 \tilde{v}_{t-1}^A + \sqrt{1 - c_0^2} w_t$$

where $c_0 = 0.5$ and w_t is white Gaussian noise with variance 0.12;

B) $\tilde{v}_t^B$ is a biased version of $\tilde{v}_t^A$: $\tilde{v}_t^B = \tilde{v}_t^A + 0.5$.

The predictor is given by

$$\hat{y}_t(a, b) = -a y_{t-1} + b u_{t-1}.$$

For each noise scenario, the system (9) was simulated from $t = -1$ to $t = 1000$. The prediction errors were given by

$$\epsilon_t(a, b) = y_t + a y_{t-1} - b u_{t-1}, \quad t = 1, \ldots, 1000$$

and $M = 960$ binary strings were constructed according to the procedure in point 4 of the algorithm. Next, we computed the empirical correlations

$$g_i^1(a, b) = \sum_{t=1}^{1000} h_{i,t} \cdot u_{t-1} \epsilon_t(a, b), \quad i = 0, \ldots, 959,$$
$$g_i^2(a, b) = \sum_{t=1}^{1000} h_{i,t} \cdot u_{t-2} \epsilon_t(a, b), \quad i = 0, \ldots, 959.$$

In order to obtain a 95% confidence set, we discarded those values of a and b for which zero was among the 12 largest or smallest values of $g_i^1(a, b)$ or zero was among the 12 largest or smallest values of $g_i^2(a, b)$. Then, according to Corollary 2, (a^0, b^0) belongs to the constructed region with probability at least $1 - 2 \cdot 2 \cdot 12/960 = 0.95$.

The obtained confidence sets are the blank areas in Fig. 6 (unbiased noise) and Fig. 7 (biased noise). The areas marked with x is where 0 is among the 12 smallest values of $g_i^1(a, b)$ and the areas marked with + is where 0 is among the 12 largest values of $g_i^1(a, b)$. Likewise for $g_i^2(a, b)$ where $\Box$ represents where 0 are among the 12 smallest values and o among the 12 largest. The true value (a^0, b^0) is marked with a diamond. As we can see, each step in the construction of the confidence set excludes a particular region.

In this case the final confidence set was obtained as the intersection of two sets, and the probability in Corollary 2 is a lower bound. To verify the tightness of this bound, we ran 1 000 000 Monte Carlo simulations and found that the empirical probability with which the true parameter belonged to the confidence set was 0.951, both with unbiased and biased noise, showing that the theoretical lower bound 0.95 is not conservative.

Using the algorithm for the construction of $\hat{\Theta}_{1000}$ 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. Moreover, the results were obtained without using any a priori knowledge about the noise.

Obviously, the size of the confidence set depends on the noise and the set is larger when the noise is biased. Also, the boundaries of the confidence set are nearly parallel to the axes when the noise is unbiased, while they are forming an angle with the "a" axis when the noise is biased. The shape of the confidence sets is studied in Section IV-B.

Next we increased the number of data points to $N = 4000$, and generated $M = 960$ binary strings of length 4000. As before, the regions in parameter space where zero were among the 12 largest or smallest values of $g_i^1(a, b)$ and $g_i^2(a, b)$ were excluded in order to obtain a 95% confidence set. The results are shown in Figs. 8 and 9. The size of the sets are smaller than with 1000 data points illustrating that the confidence set concentrates around the true parameter as the number of data points increases.

IV. Variations of the Algorithm

In this section we introduce a number of generalizations and extensions of the algorithm that provide the user with more flexibility and freedom. The main extensions are:

- generalized predictors can be used;
- the input signal does not have to be white;
- the instrumental variable ξ_t can be an arbitrary iid sequence which is uncorrelated with the noise but correlated with the input signal;

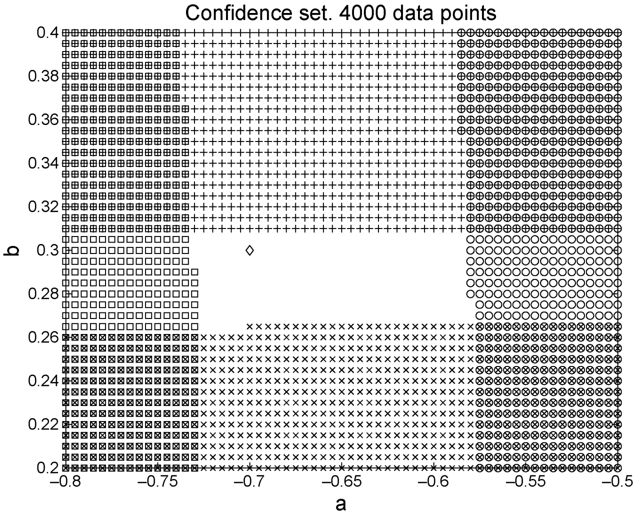

Fig. 8. Non-asymptotic confidence region for (a^0, b^0) (blank region). 4000 data points. $\tilde{v}_t$ is lowpass filtered white noise. $\diamond = \mathrm{true\ parameter}$.

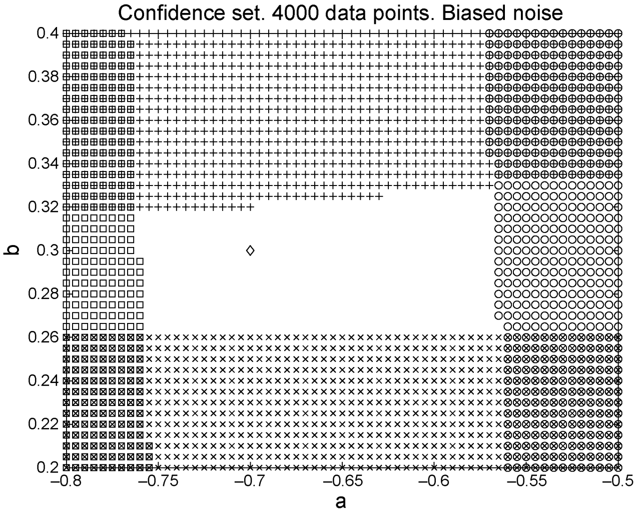

Fig. 9. Non-asymptotic confidence region for (a^0, b^0) (blank region). 4000 data points. $\tilde{v}_t$ is biased lowpass filtered white noise. $\diamond = \mathrm{true\ parameter}$.

- filtered prediction errors can be used.

The theory from Section III still remains valid with these extensions, and in Section IV-B an analysis is presented which sheds light on the role of the user choices, showing in particular how they affect the shape of the confidence set.

A. *Generalized Procedure for Construction of Confidence Regions*

1) Data Generating System: The data generating system is as before

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

and we assume that the input signal u_t is user-chosen and independent of v_t. However, we do not require u_t to be an independent sequence.

2) Generalized Predictors: Here, generalized predictors, still in the regression form from Section III

$$\hat{y}_t(\theta) = \phi_t^T\theta \tag{11}$$

are considered with

$$\phi_t = \left[\phi_t^1, \ldots, \phi_t^N\right]^T$$

where ϕ_t^j, $j = 1, \ldots, n$, are linear in the observations: $\phi_t^j = L_u^j(z^{-1})u_t + L_y^j(z^{-1})y_t$ with $L_u^j(z^{-1})$ and $L_y^j(z^{-1})$ asymptotically stable filters. Typical examples of predictors in the form (11) are Laguerre predictors (see, e.g., [20], [21]) and predictors using general orthonormal basis functions (see, e.g., [22]).

As in Section III, we assume that the true system can be represented in the model class, i.e., there exists a θ^0 such that

$$\hat{y}_t(\theta^0) = \phi_t^T\theta^0 = y_t + \text{term depending on } v_t. \tag{12}$$

3) Procedure:

Input design.

Let $u_t = L(z^{-1})\tilde{u}_t$, where $\tilde{u}_t$ is an iid sequence of random variables symmetrically distributed around 0 and $L(z^{-1})$ is an asymptotically stable filter.

Construction of confidence sets.

1) Compute the predictions

$$\hat{y}_t(\theta) = \phi_t^T\theta, \quad t = 1, \ldots, N.$$

2) Compute the prediction errors $\epsilon_t(\theta) = y_t - \phi_t^T\theta$, $t = 1, \ldots, N$, and the filtered prediction errors

$$\begin{aligned}\epsilon_t^{F_k}(\theta) &= F_k(z^{-1})\epsilon_t(\theta)\\ &= F_k(z^{-1})\left(y_t - \phi_t^T\theta\right), \quad t{=}1, \ldots, N, k{=}1, \ldots, n\end{aligned}$$

where $F_k(z^{-1})$ are asymptotically stable filters.

3) Choose a vector sequence of instrumental variables

$$\xi_t = \left[\xi_t^1, \ldots, \xi_t^N\right]^T, \quad t = 1, \ldots, N$$

such that, for each $k = 1, \ldots, n$, ξ_t^k, $t = 1, \ldots, N$, is a sequence of independent random variables symmetrically distributed around 0 and independent of v_t, and compute

$$f_t(\theta) = \begin{bmatrix} \xi_t^1\epsilon_t^{F_1}(\theta) \\ \xi_t^2\epsilon_t^{F_2}(\theta) \\ \vdots \\ \xi_t^N\epsilon_t^{F_N}(\theta) \end{bmatrix}, \quad t = 1, \ldots, N. \tag{13}$$

4) Select an integer M and construct M binary stochastic strings $h_{i,t}$ of length N as in point 4 of the algorithm in Section III and compute

$$g_i(\theta) = \sum_{t=1}^{N} h_{i,t} \cdot f_t(\theta), \quad i = 0, \ldots, M-1.$$

5) Let $g_i^k(\theta)$ denote the kth element of the vector $g_i(\theta)$, $k = 1, \ldots, n$. Select an integer q in the interval $[1, M/2n]$. Construct the regions $\hat{\Theta}_N^k$ such that at least q of the $g_i^k(\theta)$ functions are larger than $g_0^k(\theta) = 0$ and at least q are smaller than $g_0^k(\theta) = 0$. The confidence set is given by

$$\hat{\Theta}_N = \cap_{k=1}^{N}\hat{\Theta}_N^k. \tag{14}$$

By a comparison, it is easily seen that the procedure in Section III is a particular case of the above procedure. The only changes from Section III are in point 1, 2 and 3. Different filters can be used for each scalar correlation equation, and this allows us to control the shape of the confidence set as discussed in Section IV-B. A typical choice of instrumental variables are delayed versions of the signal $\tilde{u}_t$ used in the input design, and this choice satisfies the condition stated in point 3 that ξ_t^k should be independent of v_t. Theorem 1 and Corollary 2 hold unchanged in the present generalized setting and they are restated below.

Theorem 4: Consider a $k \in \{1, \ldots, n\}$ and assume that $Pr\{g_i^k(\theta^0) = 0\} = 0, i \neq 0$. Then

$$Pr\left\{\theta^0 \in \hat{\Theta}_N^k\right\} = 1 - 2q/M. \tag{15}$$

□

Proof: See Appendix A-C. ■

Corollary 5: Under the assumptions in Theorem 4

$$Pr\{\theta^0 \in \hat{\Theta}_N\} \geq 1 - 2nq/M. \tag{16}$$

□

Under some mild stability and regularity conditions, it holds that the confidence set concentrates around θ^0. We thus have the following equivalent to Theorem 3.

Theorem 6: In addition to the assumptions of Theorem 4, assume that

1) $\xi_t = [\tilde{u}_{t-1}, \ldots, \tilde{u}_{t-n}]^T$.

2) $G^0(z^{-1})$ is asymptotically stable.

3) $|\tilde{u}_t| \leq U$ for some U; moreover, $|v_t| \leq Kt^\alpha$ with probability 1 for some K and $\alpha < 1/2$, where K and α are allowed to depend on the noise realization.

4)
$$\lim_{t\to\infty} E \begin{bmatrix} \xi_t^1 F_1(z^{-1})\phi_t^T \\ \xi_t^2 F_2(z^{-1})\phi_t^T \\ \vdots \\ \xi_t^N F_N(z^{-1})\phi_t^T \end{bmatrix} = C \tag{17}$$

with C non-singular, and $F_k(z^{-1})\phi_t^T$ is a row vector whose elements are the elements of ϕ_t^T filtered through $F_k(z^{-1})$.

Suppose that an initial confidence set $\hat{\Theta}_{\bar{N}}$ is obtained from the algorithm and then this set is updated as new data are observed by extending the stochastic strings $h_{i,t}$ with new {0,1} random variables for $t = \bar{N}+1, \bar{N}+2, \ldots$ while M is kept fixed. Then, for all $\epsilon > 0$

$$Pr\left\{\exists N(\epsilon)\text{such that } \hat{\Theta}_N \subseteq \{\theta : \|\theta - \theta^0\| \leq \epsilon\}\right.$$
$$\text{for all } N > N(\epsilon)\} = 1.$$

□

The assumptions on u_t now apply to $\tilde{u}_t$. The assumption on v_t is unchanged, and the assumption that C is non-singular replaces the assumption that $z^{n_a}A^0(z^{-1})$ and $z^{n_b}B^0(z^{-1})$ have no common factors. The proof of Theorem 6 follows mutatis-mutandis that of Theorem 3 and is therefore omitted.

Remark 6: A fundamental requirement in order for the constructed region to contain θ^0 with given probability is that the vector ξ_t is component-wise an independent sequence which is also independent of the noise affecting the system. Constructing ξ_t with this property is easy, but securing that the region shrinks around θ^0 also demands that ξ_t exhibits suitable correlation properties with the system input. This property has been achieved in this paper by assuming that the system input can be chosen. However, the ideas developed here carry over to when u_t is not user-chosen, in which case ξ_t can, e.g., be constructed by suitable whitening filters applied to u_t. □

B. Size and Shape of the Confidence Sets

In Theorem 6 we have shown that the confidence set will eventually be included in an ϵ-neighborhood of θ^0. In this section we examine the shape of the confidence set and the rate at which it shrinks around θ^0. As the confidence set is constructed by excluding the regions where the functions $g_i^k(\theta)$ take on positive or negative values too many times, the boundary of the confidence set is pieced together by patches of surfaces described by $g_i^k(\theta) = 0$ for some values of k and i.

The next theorem shows that $g_i^k(\theta) = 0$ can be written as a linear equation in $\tilde{\theta} = \theta^0 - \theta$ with stochastic perturbations which are asymptotically normally distributed with zero mean and covariances which tend to zero as $1/N$.

Theorem 7: In addition to the assumptions in Theorem 4 and 6, assume that

1) v_t is given by

$$v_t = L_e(z^{-1})e_t + \bar{v}_t$$

where $\bar{v}_t$ is a bounded deterministic sequence, $L_e(z^{-1})$ is an asymptotically stable filter, and e_t is a sequence of independent random variables with zero mean and bounded moments of order $4 + \delta$ for some $\delta > 0$.

2) The limit in (29) in Lemma 9 in Appendix A-D exists.

Let $\tilde{\theta} = \theta^0 - \theta$, then

$$g_i(\theta) = \sum_{t=1}^{N} h_{i,t} \begin{bmatrix} \xi_t^1 \epsilon_t^{F_1}(\theta) \\ \xi_t^2 \epsilon_t^{F_2}(\theta) \\ \vdots \\ \xi_t^N \epsilon_t^{F_N}(\theta) \end{bmatrix} = 0 \tag{18}$$

can be written as

$$C\tilde{\theta} + \frac{1}{\sqrt{N}}\bar{C}_i\tilde{\theta} = \frac{1}{\sqrt{N}}\bar{w}_i \tag{19}$$

where C is given by (17), and $\bar{C}_i$ and $\bar{w}_i$ are a random matrix and a random vector whose elements are asymptotically jointly Gaussian with zero mean. □

Proof: See Appendix A-D. ■

The main implication of this theorem is that it provides us with information about the shape and size of the boundary of the confidence set. Let C^k and $\bar{C}_i^k$ denote the kth row of C and $\bar{C}_i$ respectively, and let $\bar{w}_i^k$ be the kth element of $\bar{w}$. The equation

$$C^k\tilde{\theta} + \frac{1}{\sqrt{N}}\bar{C}_i^k\tilde{\theta} = \frac{1}{\sqrt{N}}\bar{w}_i^k$$

describes a hyperplane in the parameter space which, for some values of i, will form the boundary of the confidence set. Asymptotically, this hyperplane is orthogonal to vector C^k, hence the matrix C determines the asymptotic shape of the confidence set. The stochastic fluctuations in the normal vector is given by $(1/\sqrt{N})\bar{C}_i^k$ which tends to zero as $1/\sqrt{N}$. Moreover, the stochastic translation of the hyperplane is due to the term $(1/\sqrt{N})\bar{w}_i^k$, and this term essentially determines the size of the confidence set.

The importance of the terms in (19) are summarized as follows:

C	shape of confidence set;
$1/\sqrt{N}\bar{C}_i$	fluctuation in shape of confidence set;
$1/\sqrt{N}\bar{w}_i$	size of the confidence set.

When C has only one non-zero element in each row, each correlation equation will asymptotically determine the confidence in one parameter. The above considerations are now made more concrete by re-examining the example in Section III-A in the light of Theorem 7.

C. Simulation Example (Continued)

In the example, $\xi_t = [u_{t-1}, u_{t-2}]^T$ and $\phi_t = [-y_{t-1}, u_{t-1}]^T$ and hence

$$C = E\begin{bmatrix} u_{t-1} \\ u_{t-2} \end{bmatrix} [-y_{t-1}, u_{t-1}] = \begin{bmatrix} 0 & 1 \\ -b^0 & 0 \end{bmatrix}. \tag{20}$$

The boundaries of the confidence set are therefore asymptotically parallel to the axes, in agreement with Figs. 8 and 9. The first equation asymptotically determines the confidence in the b parameter while the second equation determines the confidence in the a parameter. We also note that when the number of data points is increased with a factor 4 (Figs. 8 and 9) the size of the confidence set is reduced with a factor two relative to Figs. 6 and 7 in both the a and b directions and the boundaries become more parallel to the axes in accordance with the theory.

The term responsible for the size of the confidence set is

$$\frac{2}{\sqrt{N}} \sum_{t=1}^{N} h_{i,t} \begin{bmatrix} u_{t-1} \\ u_{t-2} \end{bmatrix} \tilde{v}_t \tag{21}$$

[to verify this formula one has to inspect the proof of Theorem 7, see (28)]. The two elements in the vector in (21) are about equal, and since $b^0 = 0.3$ we expect from the value of C in (20) that the confidence set will be about 1/0.3 times wider in the a direction than in the b direction which is in agreement with Figs. 6–9. Having boundaries which are asymptotically parallel to the axes is beneficial in that the construction of the confidence region becomes easier from an algorithmic point of view. In fact, in order to find the approximate range for a parameter say θ^k, one can set all other parameters to arbitrary values and use the appropriate correlation equation to determine the approximate range of θ^k. Besides the obvious theoretical interest, Theorem 7 provides insight on how to obtain sets with boundaries parallel to the axes.

V. Conclusion

We have presented an algorithm for construction of confidence sets for the parameters of linear transfer functions. The constructed confidence sets contain the true parameters with a guaranteed probability for any finite number of data points and the results are non-conservative. As the confidence sets give a description of the model uncertainty, the results in this paper are relevant for robust control systems design.

Theory-wise, the most remarkable feature of the introduced algorithm is that it works with basically no assumptions on the noise. This is also of practical importance since the noise characteristics are hardly known in most real applications. In addition, we have shown that under natural conditions the confidence set shrinks around the true parameter values as the number of data points increases, and the asymptotic size and shape of the confidence set have been derived.

Appendix A
Proofs

A. Proof of Theorem 1

The following preliminary proposition is instrumental to the proof of Theorem 1.

Proposition 1: Let H be the stochastic $M \times N$ matrix with elements $h_{i,t}, i = 0, \ldots, M-1, t = 1, \ldots, N$, constructed according to point 4 of the algorithm for the construction of confidence sets in Section III, and further let $\eta := [\eta_1, \ldots, \eta_N]^T$ be a vector independent of H of mutually independent random variables symmetrically distributed around 0. Given a $\bar{i} \in [0, M-1]$, let $H_{\bar{i}}$ be the $M \times N$ matrix whose rows are all equal to the $\bar{i}$th row of H. Then, $H\eta$ and $(H - H_{\bar{i}})\eta$ have the same M-dimensional distribution provided that the $\bar{i}$th element of $(H - H_{\bar{i}})\eta$ (which is 0) is repositioned as first element of the vector. □

Proof: Let $\mathcal{H}$ be the set of all deterministic {0,1}-valued $M \times N$ matrices whose first row is all zeros and where the rows are all different from each other. An inspection of point 4 of the algorithm in Section III reveals that the stochastic matrix H constructed there takes on a value in $\mathcal{H}$, and each matrix in $\mathcal{H}$ carries the same probability to be obtained.

Given a specific matrix $H \in \mathcal{H}$, introduce the notation $|H - H_{\bar{i}}|$ to indicate the matrix where each element of $H - H_{\bar{i}}$ has been substituted by its absolute value. Consider the following map:

$$\mathrm{map} : H \to |H - H_{\bar{i}}\,| \text{ and further reposition the } \bar{i}\text{−th row as the first row.}$$

It is easy to verify that this map transforms elements $H \in \mathcal{H}$ into elements of $\mathcal{H}$ and, moreover, if $H_1 \neq H_2$, then $\mathrm{map}(H_1) \neq \mathrm{map}(H_2)$. That is, the map is one-to-one on $\mathcal{H}$.

Now, due to that the map is one-to-one and that the stochastic matrix H constructed in point 4 takes all possible matrices in $\mathcal{H}$ with the same probability, it turns out that $\mathrm{map}(H)$ has the same probability distribution as H.

Introduce next the new variables

$$\tilde{\eta}_t := \begin{cases} \eta_t, & if\ h_{\bar{i},t} = 0 \\ -\eta_t, & if\ h_{\bar{i},t} = 1 \end{cases}$$

and let $\tilde{\eta}$ be the vector with elements $\tilde{\eta}_t$. We show below that vector $\tilde{\eta}$: **(i)** is independent of H (so that $\tilde{\eta}$ is also independent of $\mathrm{map}(H)$); and **(ii)** it has the same distribution as η.

To verify these two properties without too much notational clutter, suppose that $N = 2$. Fix a specific matrix $\bar{H} \in \mathcal{H}$ and concentrate on the event where $H = \bar{H}$. The entries of the $\bar{i}$ row of $\bar{H}$ will take on fixed numerical values, for the sake of concreteness say $(h_{\bar{i},1}, h_{\bar{i},2}) = (0, 1)$. Then, over the event where $H = \bar{H}$, for given sets E_1 and E_2, we have: $Pr\{H = \bar{H}, \tilde{\eta}_1 \in E_1, \tilde{\eta}_2 \in E_2\} = Pr\{H = \bar{H}, \eta_1 \in E_1, -\eta_2 \in E_2\} =$ [since H and η are independent] $= Pr\{H = \bar{H}\} \cdot Pr\{\eta_1 \in E_1, -\eta_2 \in E_2\}$ =[since η_1 and η_2 are independent] $=$

$Pr\{H = \bar{H}\} \cdot Pr\{\eta_1 \in E_1\} \cdot Pr\{-\eta_2 \in E_2\} =$ [since η_2 is symmetrically distributed] $= Pr\{H = \bar{H}\} \cdot Pr\{\eta_1 \in E_1\} \cdot Pr\{\eta_2 \in E_2\} = Pr\{H = \bar{H}, \eta_1 \in E_1, \eta_2 \in E_2\}$, showing that $(\tilde{\eta}_1, \tilde{\eta}_2)$ and (η_1, η_2) have the same distribution, conditionally to that $H = \bar{H}$. Since the same holds for any other choice of $\bar{H}$, the conclusion is drawn that $(H, \tilde{\eta}_1, \tilde{\eta}_2)$ has the same joint distribution as (H, η_1, η_2). Generalizing to any N, we similarly get that $(H, \tilde{\eta})$ has the same joint distribution as (H, η). Now, property **(i)** that $\tilde{\eta}$ is independent of H follows from that η is independent of H, since the joint distribution of H and $\tilde{\eta}$ is the same as that of H and η. Moreover, the marginals of $\tilde{\eta}$ and η are obviously the same, and this is property **(ii)**.

To conclude the proof, observe now that $(H - H_{\bar{i}})\eta = |H - H_{\bar{i}}|\tilde{\eta}$, so that the vector $(H - H_{\bar{i}})\eta$ where the $\bar{i}$th element is repositioned as the first entry is the same as $\mathrm{map}(H) \cdot \tilde{\eta}$. Since $\mathrm{map}(H)$ is distributed as H, $\tilde{\eta}$ is distributed as η, and $\mathrm{map}(H)$ and $\tilde{\eta}$ are independent, $\mathrm{map}(H) \cdot \tilde{\eta}$ has the same distribution as $H\eta$ and the proposition is established. ■

Remark 7: Interestingly enough, the above proposition admits an easy generalization that can prove useful in the construction of modified identification algorithms. Suppose that, in the construction of strings $h_{i,1}, \ldots, h_{i,N}$ in point 4 of the algorithm, a string is discarded if it turns out to be equal to an already constructed string or if it differs from an already constructed string by 1 element only. Then, the result stated in Proposition 1 keeps true. Moreover, we can push this process further and ask that a string be discarded whenever it differs from an already constructed string by less than ν elements. By applying the so modified proposition, the proof of Theorem 1 given below goes through. The advantage of having strings that are well apart from each other is that they carry more diversified information, so that the resulting confidence regions are in general tighter. □

We turn now to the proof of Theorem 1.

We first prove that each variable $g_i^k(\theta^0)$, $i = 0, \ldots, M-1$, has the same probability $1/M$ to be in the generic rth position (i.e., there are exactly $r-1$ other variables smaller than the variable under consideration).

Pick a variable $g_{\bar{i}}^k(\theta^0)$. $g_{\bar{i}}^k(\theta^0)$ is in the rth position if the inequality

$$g_{\bar{i}}^k(\theta^0) > g_i^k(\theta^0)$$

is satisfied for exactly $r-1$ choices of $i \in \{0, \ldots, M-1\}$, or, equivalently, if

$$g_i^k(\theta^0) - g_{\bar{i}}^k(\theta^0) < 0 \tag{22}$$

holds for $r-1$ selections of i. Let $\eta_t := u_{t-k} A^0(z^{-1}) v_t$, and note that (22) can be rewritten as

$$\begin{aligned} g_i^k(\theta^0) - g_{\bar{i}}^k(\theta^0) &= \sum_{t=1}^{N} h_{i,t} \cdot \eta_t - \sum_{t=1}^{N} h_{\bar{i},t} \cdot \eta_t \\ &= \sum_{t=1}^{N} (h_{i,t} - h_{\bar{i},t}) \eta_t < 0 \end{aligned}$$

so that the requirement that "$g_i^k(\theta^0) - g_{\bar{i}}^k(\theta^0) < 0$ holds for $r-1$ selections of i" is the same as asking that $r-1$ entries of $(H - H_{\bar{i}})\eta$ are negative. Fix now a realization of the noise v_t, that is v_t is regarded from now on as deterministic; in mathematical terms, this corresponds to say that derivations are carried out conditionally to v_t. We next appeal to Proposition 1 and observe that, by an easy inspection, the assumptions there introduced are all satisfied (particularly, the fact that $\eta_1, \ldots \eta_N$ are mutually independent holds here conditionally to the picked realization of v_t, a fact that holds true due to that u_t is an independent sequence and that u_t and v_t are independent of each other, see Assumption 1[3]). From Proposition 1 we have that $(H - H_{\bar{i}})\eta$ has the same distribution as $H\eta$ and therefore the probability of the event where "$r-1$ entries of $(H - H_{\bar{i}})\eta$ are negative" is the same as the probability of the event where $r-1$ entries of $H\eta$ are negative. But this latter event does not depend on $\bar{i}$, showing that the probability of being in the rth position is the same for any $g_{\bar{i}}^k(\theta^0)$. Since this holds for any realization of the noise, it also holds unconditionally, and the probability is therefore $1/M$ since $\bar{i}$ can take on M possible values.[4]

To conclude the proof, let

$$\begin{aligned} A = \{&g_0^k(\theta^0) \text{ is in the 1−st or 2−nd or} \\ &\ldots \text{ or } q\text{−th position}\} \\ \cup \{&g_0^k(\theta^0) \text{ is in the } M\text{−th or } (M-1)\text{−th or} \\ &\ldots \text{ or } (M-q+1)\text{−th position}\}. \end{aligned}$$

Since $g_0^k(\theta^0) := 0$ has the same probability $1/M$ to be in the generic rth position, $Pr\{A\} = 2q/M$. Next, suppose the probabilistic event A has occurred. Then, either $g_i^k(\theta^0) > 0$ holds for at most $q-1$ selections of i or $g_i^k(\theta^0) < 0$ holds for at most $q-1$ selections of i, so that $\theta^0 \notin \hat{\Theta}_N^k$ (recall the construction of $\hat{\Theta}_N^k$ in point 5 of the algorithm). Viceversa, if the probabilistic event A has not occurred, then $g_i^k(\theta^0) > 0$ holds for at least q selections of i and $g_i^k(\theta^0) < 0$ holds for at least q selections of i, yielding $\theta^0 \in \hat{\Theta}_N^k$. Thus, $Pr\{\theta^0 \in \hat{\Theta}_N^k\} = 1 - Pr\{A\} = 1 - 2q/M$ and (7) is proven. ■

B. Proof of Theorem 3

The proof is technical. We start with a preliminary convergence result given in the following lemma. In the lemma we have used the argument ω to show the explicit dependence of some empirical matrices and vectors on the observed realization of the stochastic processes.

Lemma 8: Let $\tilde{v}_t = A^0(z^{-1}) v_t$, $\xi_t = [u_{t-1}, \ldots, u_{t-n}]^T$, $\phi_t = [-y_{t-1}, \ldots, -y_{t-n_a}, u_{t-1}, \ldots, u_{t-n_b}]^T$ and furthermore let

$$C_{i,N}(\omega) = \frac{1}{N} \sum_{t=1}^{N} h_{i,t} \cdot \xi_t \phi_t^T \tag{23}$$

$$D_{i,N}(\omega) = \frac{1}{N} \sum_{t=1}^{N} h_{i,t} \cdot \xi_t \tilde{v}_t. \tag{24}$$

[3] Note that this is the very point where the open loop assumption plays a crucial role.

[4] Note that the probability that any two random variables $g_i^k(\theta^0)$ and $g_{\bar{i}}^k(\theta^0)$ coincide is zero. In fact, $g_i^k(\theta^0) - g_{\bar{i}}^k(\theta^0) = 0$ with non-zero probability implies that an element of $(H - H_{\bar{i}})\eta$ is zero with non-zero probability and this in turn means that an element of $H\eta$ other than the first one is zero with non-zero probability, a possibility which is excluded by the assumptions of the theorem.

Under the assumptions in Theorem 3

$$\max_{i\in\{0,\dots,M-1\}} \|C_{i,N}(\omega) - C\| \to 0 \text{ w.p. } 1 \text{ as } N \to \infty,$$
$$\max_{i\in\{0,\dots,M-1\}} \|D_{i,N}(\omega)\| \to 0 \text{ w.p. } 1 \text{ as } N \to \infty$$

where C is a non-singular matrix given by

$$C = \frac{\sigma_u^2}{2}\left[-A_\alpha \left| \begin{matrix} I_{n_b} \\ 0_{n_a\times n_b}\end{matrix}\right.\right]$$

with $\sigma_u^2 = E[u_t^2]$, I_{n_b} = identity matrix of size n_b, $0_{n_a\times n_b}$ = matrix of size $n_a \times n_b$ whose elements are all zero, and

$$A_\alpha = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ \alpha_1 & 0 & 0 & \cdots & 0 \\ \alpha_2 & \alpha_1 & 0 & \cdots & 0 \\ \alpha_3 & \alpha_2 & \alpha_1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \alpha_{n_a+1} & \alpha_{n_a} & \alpha_{n_a-1} & \cdots & \alpha_1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \alpha_{n-1} & \alpha_{n-2} & \alpha_{n-3} & \cdots & \alpha_{n-n_a} \end{bmatrix}$$

where α_j are the Markov coefficients of the system $B^0(z^{-1})/A^0(z^{-1})$, i.e.

$$\frac{B^0(z^{-1})}{A^0(z^{-1})} = \sum_{j=1}^{\infty} \alpha_j z^{-j}.$$

□

Proof: Since M is a fixed number, "$\max_{i\in\{0,\dots,M-1\}}$" does not influence the convergence results and the max operator is ignored in the sequel. The sequence $h_{i,t}$ is independent after $t = \bar{N}$ (it is not for $t \le \bar{N}$ due to the string removal described in point 4 of the algorithm). However, as the first $\bar{N}$ elements have no importance for asymptotic results, we treat $h_{i,t}$ as an independent sequence.

The proof is carried out by showing convergence with probability 1 for each element in the matrices $C_{i,N}$ and vectors $D_{i,N}$. We consider the different types of terms one by one.

A straightforward application of Hoeffding's inequality (see Appendix B) gives for $l = 1,\dots,n_b$

$$Pr\left\{\left|\frac{1}{N}\sum_{t=1}^{N} h_{i,t}u_{t-l}^2 - \sigma_u^2/2\right| \ge \epsilon\right\} \le 2e^{\frac{-2N\epsilon^2}{U^4}}$$

which establishes uniform convergence in probability of $(1/N)\sum_{t=1}^{N} h_{i,t}u_{t-l}^2$ to $\sigma_u^2/2$. Convergence with probability 1 follows by an application of the Borel-Cantelli Lemma (see, e.g., [23]).

For the terms $(1/N)\sum_{t=1}^{N} h_{i,t}u_{t-k}u_{t-l}$, $k \ne l$, we observe that $\{u_{t-k}u_{t-l}, t \in \{1,\dots,N\}\}$ can be written as union of two disjoint sets

$$\{u_{t-k}u_{t-l}, t\in\{1,\dots,N\}\} = \{u_{t-k}u_{t-l}, t \in I_1\} \cup \{u_{t-k}u_{t-l}, t \in I_2\}$$

such that each set on the right hand side consists of mutually independent variables. Then

$$Pr\left\{\left|\frac{1}{N}\sum_{t=1}^{N} h_{i,t}u_{t-k}u_{t-l}\right| \ge \epsilon\right\} \le Pr\left\{\left|\frac{1}{N}\sum_{t\in I_1} h_{i,t}u_{t-k}u_{t-l}\right| \ge \frac{\epsilon}{2}\right\} + Pr\left\{\left|\frac{1}{N}\sum_{t\in I_2} h_{i,t}u_{t-k}u_{t-l}\right| \ge \frac{\epsilon}{2}\right\}$$

and convergence with probability 1 of $(1/N)\sum_{t=1}^{N} h_{i,t}u_{t-k}u_{t-l}$ to 0 follows by applying Hoeffding's inequality and the Borel-Cantelli Lemma as above to each of these two terms.

Next we consider terms of the type $(1/N)\sum_{t=1}^{N} h_{i,t}u_{t-k}\tilde{v}_t$. Consider first $(1/N)\sum_{t=1}^{N} h_{i,t}u_{t-k}v_t$, that is the same expression as before with v_t replacing $\tilde{v}_t$. As in the proof of Theorem 1 we can treat v_t as a deterministic sequence, that is derivations are carried out conditionally to the realization of v_t. From the assumption that $|v_t| \le Kt^\alpha$ for some K and $\alpha < 1/2$, using again Hoeffding's inequality we have

$$Pr\left\{\left|\frac{1}{N}\sum_{t=1}^{N} h_{i,t}u_{t-k}v_t\right| \ge \epsilon\right\} \le 2e^{-\frac{2N\epsilon^2}{4U^2\max_{t\in\{1,\dots,N\}}|v_t|^2}} \le 2e^{-\frac{2N\epsilon^2}{4U^2K^2N^{2\alpha}}} = 2e^{-\frac{2N^{1-2\alpha}\epsilon^2}{4U^2K^2}}$$

and convergence follows as before. Since $\tilde{v}_t = A^0(z^{-1})v_t$ is simply a linear combination of v_t terms, uniform convergence of $(1/N)\sum_{t=1}^{N} h_{i,t}u_{t-k}\tilde{v}_t$ follows in the same way.

Finally, we consider the terms $(1/N)\sum_{t=1}^{N} h_{i,t}u_{t-k}y_{t-l}$. The system output y_t can be written as

$$y_t = \sum_{j=1}^{t-1+n} \alpha_j u_{t-j} + v_t + i.c._t$$

where $i.c._t$ accounts for the effect of the initial conditions. Hence

$$\frac{1}{N}\sum_{t=1}^{N} h_{i,t}u_{t-k}y_{t-l} = \frac{1}{N}\sum_{t=1}^{N} h_{i,t}u_{t-k}\sum_{j=1}^{t-l-1+n}\alpha_j u_{t-l-j} + \frac{1}{N}\sum_{t=1}^{N} h_{i,t}u_{t-k}v_{t-l} + \frac{1}{N}\sum_{t=1}^{N} h_{i,t}u_{t-k}i.c._{t-l}.$$

Similar to previous results, the second term on the right hand side converges to 0 with probability 1. Moreover, the last term is also vanishing due to the asymptotic stability of the system. When $l \le k-1$, the first term contains (take $j = k-l$) $(1/N)\sum_{t=1}^{N} h_{i,t}\alpha_{k-l}u_{t-k}^2$, which, similarly to previous results, converges to $\alpha_{k-l}\sigma_u^2/2$ with probability 1. If we remove $(1/N)\sum_{t=1}^{N} h_{i,t}\alpha_{k-l}u_{t-k}^2$ from the first term, the absolute value of the remaining expression can be bounded as follows (to simplify notations, in the derivation below we let $u_t = 0$ for $t \leq -n$):

$$\begin{aligned}\left|\frac{1}{N}\sum_{t=1}^{N} h_{i,t}u_{t-k}\sum_{j=1,j\neq k-l}^{t-l-1+n}\alpha_j u_{t-l-j}\right| &= \left|\frac{1}{N}\sum_{t=1}^{N} h_{i,t}u_{t-k}\sum_{j=1,j\neq k-l}^{\infty}\alpha_j u_{t-l-j}\right| \\ &\leq \sum_{j=1,j\neq k-l}^{\infty}|\alpha_j|\left|\frac{1}{N}\sum_{t=1}^{N} h_{i,t}u_{t-k}u_{t-l-j}\right|. \qquad (25)\end{aligned}$$

Since $A^0(z^{-1})$ is asymptotically stable, there exist H and λ such that $|\alpha_j| \leq H\lambda^j$. Thus, recalling that $\sum_{j=1,j\neq k-l}^{\infty} H\lambda^j j\epsilon \leq \epsilon H/(1-\lambda)^2$, it follows from (25) that the condition $|(1/N)\sum_{t=1}^{N} h_{i,t}u_{t-k}\sum_{j=1,j\neq k-l}^{t-l-1+n}\alpha_j u_{t-l-j}| \geq \epsilon$ implies that

$$\left|\frac{1}{N}\sum_{t=1}^{N} h_{i,t}u_{t-k}u_{t-l-j}\right| \geq \frac{j\epsilon}{H}(1-\lambda)^2 \qquad (26)$$

for at least some $j \neq k-l$. But, again using Hoeffding's inequality, (26) holds on an event whose probability is bounded by $\gamma e^{-\delta N j^2 \epsilon^2}$, for suitable constants γ and δ, so that (26) holds for some $j \neq k-l$ with probability no more than $\sum_{j=1}^{\infty}\gamma e^{-\delta N j^2\epsilon^2} \leq \sum_{j=1}^{\infty}\gamma e^{-\delta N j\epsilon^2} = \gamma(e^{-\delta N\epsilon^2}/(1-e^{\delta N\epsilon^2}))$, which proves that the left hand side of (25) goes to zero in probability. Convergence with probability 1 then follows by the Borel-Cantelli Lemma.

Combining the above results we get that $(1/N)\sum_{t=1}^{N} h_{i,t}u_{t-k}y_{t-l}$ converges to $\alpha_{k-l}\sigma_u^2/2$ with probability 1.

Putting everything together it follows that $C_{i,N}$ converges with probability 1 to C and that $D_{i,N}$ converges with probability 1 to 0.

The fact that C is nonsingular follows from the fact that $z^{n_a}A^0(z^{-1})$ and $z^{n_b}B^0(z^{-1})$ have no common factors, see the proof of Theorem 3.2 in [14]. ■

We now return to the proof of Theorem 3. The idea is to show that at least one entry of the vector $(1/N)\sum_{t=1}^{N} h_{i,t}\cdot\xi_t\epsilon_t(\theta)$ converges with probability 1 to a nonzero value uniformly in i as N tends to infinity for all $\theta \neq \theta^0$. This means that for N sufficiently large at least one entry of vector $\sum_{t=1}^{N} h_{i,t}\cdot\xi_t\epsilon_t(\theta)$ is larger than 0 or smaller than 0 for all $i \in \{0,\ldots,M-1\}$ and hence this parameter value is excluded from the confidence set in point 5 of the algorithm in Section III.

Proof: First we rewrite $(1/N)\sum_{t=1}^{N} h_{i,t}\cdot\xi_t\epsilon_t(\theta)$ in a form more suitable to draw our conclusions. Note that

$$\epsilon_t(\theta) = y_t - \hat{y}_t(\theta) = \phi_t^T\tilde{\theta} + A^0(z^{-1})v_t$$

where $\tilde{\theta} = \theta^0 - \theta = [\tilde{a}_1,\ldots,\tilde{a}_{n_a},\tilde{b}_1,\ldots,\tilde{b}_{n_b}]^T$. Then

$$\begin{aligned}\frac{1}{N}\sum_{t=1}^{N} h_{i,t}\cdot\xi_t\epsilon_t(\theta) &= \frac{1}{N}\sum_{t=1}^{N} h_{i,t}\cdot\xi_t\left(\phi_t^T\tilde{\theta} + \tilde{v}_t\right) \\ &= C_{i,N}(\omega)\tilde{\theta} + D_{i,N}(\omega)\end{aligned}$$

where $C_{i,N}(\omega)$ and $D_{i,N}(\omega)$ are given in Lemma 8.

Now, pick a $\tilde{\theta}$ such that $\|\tilde{\theta}\| > \epsilon$ and observe that $\|C_{i,N}(\omega)\tilde{\theta} + D_{i,N}(\omega)\| \geq \|C_{i,N}(\omega)\tilde{\theta}\| - \|D_{i,N}(\omega)\| \geq \|C\tilde{\theta}\| - \|(C_{i,N}(\omega) - C)\tilde{\theta}\| - \|D_{i,N}(\omega)\| \geq \|C\tilde{\theta}\| - \|(C_{i,N}(\omega) - C)\|\|\tilde{\theta}\| - \|D_{i,N}(\omega)\|$, where C is given in Lemma 8. From Lemma 8, $\|C_{i,N}(\omega) - C\|$ and $\|D_{i,N}(\omega)\|$ tend to zero with probability 1 uniformly in i. On the other hand, C being nonsingular, $\|C\tilde{\theta}\| \geq \underline{\sigma}(C)\|\tilde{\theta}\|$ where $\underline{\sigma}(C) > 0$ is the minimum singular value of C, so that at least one element of $C\tilde{\theta}$—say the jth element—is greater than or equal to $\underline{\sigma}(C)\|\tilde{\theta}\|/\sqrt{n}$ in magnitude. Then, it is easy to see that with probability 1 there exits an $N(\epsilon)$ such that for $N \geq N(\epsilon)$ the sign of the jth element of $C_{i,N}(\omega)\tilde{\theta} + D_{i,N}(\omega)$ is the same as that of the jth element of $C\tilde{\theta}$ for all i, that is it is either positive or negative for all i. Moreover, $N(\epsilon)$ does not depend on the specific $\tilde{\theta}$. Thus, after $N(\epsilon)$, all $\tilde{\theta}$ with $\|\tilde{\theta}\| > \epsilon$ will be excluded from $\hat{\Theta}_N$ in point 5 of the algorithm in Section III and $\hat{\Theta}_N$ concentrates in an ϵ-neighborhood of θ^0. ■

C. *Proof of Theorem 4*

From the assumptions it follows that $y_t - \hat{y}_t(\theta^0)$ is independent of ξ_t. Hence, $\xi_t^k\epsilon_t^{F_k}(\theta^0) = \xi_t^k F_k(z^{-1})(y_t - \hat{y}_t(\theta^0))$ is a sequence of variables that are, conditionally on v_t, independent and symmetrically distributed around 0. The remainder of the proof is identical to the proof of Theorem 1.

D. *Proof of Theorem 7*

From (12) it follows that (18) can be written as:

$$\sum_{t=1}^{N} h_{i,t}\begin{bmatrix}\xi_t^1 F_1(z^{-1})\phi_t^T \\ \xi_t^2 F_2(z^{-1})\phi_t^T \\ \vdots \\ \xi_t^N F_N(z^{-1})\phi_t^T\end{bmatrix}\tilde{\theta} + \sum_{t=1}^{N} h_{i,t}\begin{bmatrix}\xi_t^1 F_1(z^{-1})\tilde{v}_t \\ \xi_t^2 F_2(z^{-1})\tilde{v}_t \\ \vdots \\ \xi_t^N F_N(z^{-1})\tilde{v}_t\end{bmatrix} = 0 \qquad (27)$$

where $\tilde{v}_t$ is the "term depending on v_t" in (12), with sign changed. Multiplying by $1/\sqrt{N}$ and adding and subtracting $\sqrt{N}C\tilde{\theta}/2$ where C is given in (17) gives

$$\begin{aligned}&\frac{1}{2}\sqrt{N}C\tilde{\theta} + \frac{1}{\sqrt{N}}\sum_{t=1}^{N}\left(h_{i,t}\begin{bmatrix}\xi_t^1 F_1(z^{-1})\phi_t^T \\ \xi_t^2 F_2(z^{-1})\phi_t^T \\ \vdots \\ \xi_t^N F_N(z^{-1})\phi_t^T\end{bmatrix} - \frac{1}{2}C\right)\tilde{\theta} \\ &\qquad + \frac{1}{\sqrt{N}}\sum_{t=1}^{N} h_{i,t}\begin{bmatrix}\xi_t^1 F_1(z^{-1})\tilde{v}_t \\ \xi_t^2 F_2(z^{-1})\tilde{v}_t \\ \vdots \\ \xi_t^N F_N(z^{-1})\tilde{v}_t\end{bmatrix} = 0. \qquad (28)\end{aligned}$$

Let $X_{i,t}$ be an n^2+n vector which contains all the elements of the matrix and vector occurring in the two last terms of (28)

$$X_t = \begin{bmatrix} h_{i,t}\xi_t^1 F_1(z^{-1})\phi_t^1 - \frac{1}{2}c_{11} \\ h_{i,t}xi_t^1 F_1(z^{-1})\phi_t^2 - \frac{1}{2}c_{12} \\ \vdots \\ h_{i,t}\xi_t^1 F_1(z^{-1})\phi_t^N - \frac{1}{2}c_{1n} \\ h_{i,t}\xi_t^2 F_2(z^{-1})\phi_t^1 - \frac{1}{2}c_{21} \\ \vdots \\ h_{i,t}\xi_t^N F_N(z^{-1})\phi_t^N - \frac{1}{2}c_{nn} \\ h_{i,t}\xi_t^1 F_1(z^{-1})\tilde{v}_t \\ \vdots \\ h_{i,t}\xi_t^N F_N(z^{-1})\tilde{v}_t \end{bmatrix}.$$

Theorem 7 now follows from the next lemma.

Lemma 9: Assume that the limit

$$\Gamma = \lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} EX_t X_t^T \tag{29}$$

exists. Then $(1/\sqrt{N})\sum_{t=1}^{N} X_t$ is asymptotically normally distributed with zero mean and covariance matrix Γ. □

Proof: $F_l(z^{-1})\phi_t^T$ is given by ($L_y(z^{-1})$ and $L_u(z^{-1})$ are vectors containing the filters $L_y^j(z^{-1})$ and $L_u^j(z^{-1})$)

$$\begin{aligned} F_l(z^{-1})\phi_t^T &= F_l(z^{-1})\left(L_y(z^{-1})y_t + L_u(z^{-1})u_t\right) \\ &= F_l(z^{-1})\left(L_y(z^{-1})\left(G^0(z^{-1})u_t + v_t\right)\right. \\ &\quad \left. + L_u(z^{-1})u_t\right) \\ &= F_l(z^{-1})\left(L_y(z^{-1})\left(G^0(z^{-1})L(z^{-1})\tilde{u}_t\right.\right. \\ &\quad \left.\left. + L_e(z^{-1})e_t + \bar{v}_t\right) + L_u(z^{-1})L(z^{-1})\tilde{u}_t\right). \end{aligned}$$

Similarly, $F_l(z^{-1})\tilde{v}_t$ is a linear filtering of e_t and $\bar{v}_t$. This means that a generic element of the X_t vector can be written as

$$\begin{aligned} h_{i,t}\left(\tilde{u}_{t-k}\left(M_{\tilde{u}}(z^{-1})\tilde{u}_t + M_e(z^{-1})e_t + M_{\bar{v}}(z^{-1})\bar{v}_t\right)\right) \\ -\frac{1}{2}E\tilde{u}_{t-k}M_{\tilde{u}}(z^{-1})\tilde{u}_t \end{aligned}$$

where the filters $M_{\tilde{u}}(z^{-1})$, $M_e(z^{-1})$ and $M_{\bar{v}}(z^{-1})$ are asymptotically stable. The proof for asymptotic normality now follows from the proof of the asymptotic normality of the vector S_N in equation (9.A.13) in Appendix 9A of [1]. ■

Appendix B
Hoeffding's Inequality

Suppose $Y_1, \ldots Y_m$ are independent random variables with $E[Y_i] = 0$ for each i, and such that $a_i \le Y_i \le b_i$ for each i. Then $Pr\{|\sum_{i=1}^m Y_i| \ge \alpha\} \le 2e^{-2\alpha^2/\sum_{i=1}^m (b_i-a_i)^2}$.

References

[1] L. Ljung, *System Identification—Theory for the User*, 2nd ed. Englewood Cliffs, NJ: Prentice Hall, 1999.

[2] T. Söderström and P. Stoica, *System Identification*. Englewood Cliffs, NJ: Prentice Hall, 1998.

[3] S. Garatti, M. C. Campi, and S. Bittanti, "Assessing the quality of identified models through the asymptotic theory—When is the result reliable?," *Automatica*, vol. 40, pp. 1319–1332, 2004.

[4] S. Garatti, M. C. Campi, and S. Bittanti, "The asymptotic model quality assessment for instrumental variable identification revisited," *Syst. Control Lett.*, vol. 55, pp. 494–500, 2006.

[5] E. W. Bai, K. M. Nagpal, and R. Tempo, "Bounded-error parameter estimation: Noise models and recursive algorithms," *Automatica*, vol. 32, pp. 985–999, 1996.

[6] E. W. Bai, R. Tempo, and H. Cho, "Membership set estimators: Size, optimal inputs, complexity and relations with least squares," *IEEE Trans. Circuits Syst.*, vol. 42, no. 5, pp. 266–277, May 1995.

[7] A. Vicino and G. Zappa, "Sequential approximation of feasible parameter sets for identification with set membership uncertainty," *IEEE Trans. Autom. Control*, vol. 41, no. 6, pp. 774–785, Jun. 1996.

[8] A. Garulli, L. Giarre', and G. Zappa, "Identification of approximated Hammerstein models in a worst-case setting," *IEEE Trans. Autom. Control*, vol. 47, no. 7, pp. 2046–2050, Dec. 2002.

[9] A. Garulli, A. Vicino, and G. Zappa, "Conditional central algorithms for worst-case set membership identification and filtering," *IEEE Trans. Autom. Control*, vol. 45, no. 1, pp. 14–23, Jan. 2000.

[10] L. Giarre', B. Z. Kacewicz, and M. Milanese, "Model quality evaluation in set membership identification," *Automatica*, vol. 33, pp. 1133–1139, 1997.

[11] S. Bittanti and M. Lovera, "Bootstrap-based estimates of uncertainty in subspace identification methods," *Automatica*, vol. 36, pp. 1605–1615, 2000.

[12] F. Tjärnström and L. Ljung, "Using the bootstrap to estimate the variance in the case of undermodelling," *IEEE Trans. Autom. Control*, vol. 47, no. 2, pp. 395–398, Feb. 2002.

[13] J. Shao and D. Tu, *The Jackknife and Bootstrap*. New York: Springer Verlag, 1995.

[14] M. C. Campi and E. Weyer, "Guaranteed non-asymptotic confidence regions in system identification," *Automatica*, vol. 41, pp. 1751–1764, 2005.

[15] M. C. Campi and E. Weyer, "Finite sample properties of system identification methods," *IEEE Trans. Autom. Control*, vol. 47, no. 8, pp. 1329–1334, Aug. 2002.

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

[17] M. C. Campi, S. K. Ooi, and E. Weyer, "Non-asymptotic quality assessment of generalized FIR models with periodic inputs," *Automatica*, vol. 40, pp. 2029–2041, 2004.

[18] M. C. Campi and E. Weyer, "Non-asymptotic confidence sets for input-output transfer functions," in *Proc. 45th Conf. Decision Control*, San Diego, CA, 2006, pp. 157–162.

[19] M. C. Campi and E. Weyer, "Identification with finitely many data points: The LSCR approach," in *Proc. Symp. Syst. Ident. (SYSID'06)*, Newcastle, Australia, 2006, [CD ROM].

[20] B. Wahlberg, "System identification using Laguerre models," *IEEE Trans. Autom. Control*, vol. 36, no. 5, pp. 551–562, May 1991.

[21] B. Wahlberg and E. Hannan, "Parameteric signal modelling using Laguerre filters," *Annals Appl. Probabililty*, vol. 3, pp. 467–496, 1993.

[22] P. S. C. Heuberger, T. J. de Hoog, P. M. J. van den Hof, and B. Wahlberg, "Orthonormal basis functions in time and frequency domains: Hambo transform theory," *SIAM J. Control Optim.*, vol. 42, no. 4, pp. 1347–1373, 2003.

[23] A. N. Shiryaev, *Probability*, 2nd ed. New York: Springer, 1996.

Marco C. Campi (SM'08) received the Doctor degree in electronic engineering from the Politecnico di Milano, Milano, Italy, in 1988.

He is Professor of automatic control at the University of Brescia, Brescia, Italy. From 1988 to 1989, he was a Research Assistant at the Department of Electrical Engineering, Politecnico di Milano. From 1989 to 1992, he was a Researcher at the Centro di Teoria dei Sistemi, National Research Council (CNR), Milano, and in 1992, he joined the University of Brescia. He has held visiting and teaching positions at many universities and institutions including the Australian National University, Canberra, Australia; the University of Illinois at Urbana-Champaign; the Centre for Artificial Intelligence and Robotics, Bangalore, India; the University of Melbourne, Australia; and the Kyoto University, Japan. He is an Associate Editor of *Systems and Control Letters*, and a past Associate Editor of *Automatica* and the *European Journal of Control*. His research interests include system identification, stochastic systems, adaptive and data-based control, robust convex optimization, robust control and estimation, and learning theory.

Dr. Campi received the "Giorgio Quazza" Prize as author of the best original thesis for year 1988, and the IEEE CSS George S. Axelby Outstanding Paper Award for the article "The Scenario Approach to Robust Control Design" in 2008. From 2002 to 2008, he served as Chair of the Technical Committee IFAC on Stochastic Systems (SS) and he is currently vice-chair for the Technical Committee IFAC on Modeling, Identification and Signal Processing (MISP). Moreover, he has been a distinguished lecturer of the Control Systems Society.

Erik Weyer (M'92) received the Siv. Ing. and Ph.D. degrees from the Norwegian Institute of Technology, Trondheim, in 1988 and 1993, respectively.

From 1994 to 1996, he was a Research Fellow at the University of Queensland, Brisbane, Australia, and since 1997 he has been with the Department of Electrical and Electronic Engineering, the University of Melbourne, Melbourne, Australia, where he is currently an Associate Professor. His research interests are in the area of system identification and control.