Automatica 45 (2009) 2175–2186

Contents lists available at ScienceDirect

Automatica

journal homepage: www.elsevier.com/locate/automatica

Non-asymptotic confidence regions for model parameters in the presence of unmodelled dynamics☆

Marco C. Campi [a], Sangho Ko [b], Erik Weyer [c,*]

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

[b] School of Aerospace and Mechanical Engineering, Korea Aerospace University. 100, Hanggongdae-gil, Hwajeon-dong, Deogyang-gu, Goyang, Gyeonggi-do, 412-791, South Korea

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

ARTICLE INFO

Article history:
Received 11 October 2008
Accepted 26 May 2009
Available online 3 August 2009

Keywords:
System identification
Undermodelling
Confidence regions
Orthonormal basis functions
Finite sample results

ABSTRACT

This paper deals with the problem of constructing confidence regions for the parameters of truncated series expansion models. The models are represented using orthonormal basis functions, and we extend the 'Leave-out Sign-dominant Correlation Regions' (LSCR) algorithm such that *non-asymptotic* confidence regions for the parameters can be constructed in the presence of unmodelled dynamics. The constructed regions have guaranteed probability of containing the true parameters for any finite number of data points. The algorithm is first developed for FIR models and then extended to models with generalized orthonormal basis functions. The usefulness of the developed approach is demonstrated for FIR and Laguerre models in simulation examples.

1. Introduction

One of the intrinsic tasks in system identification is to evaluate how close the model is to the true system. This depends heavily on the quality and the size of the observed input–output data set and the specific rule used to construct a (set of) model(s) from the observed data. This work focuses on truncated series expansion models represented by orthonormal basis functions, and it develops a method for constructing confidence regions for the coefficients of the series expansion using only finitely many input–output data points. This is of importance since in any practical situation the observed number of data points will always be finite, and models of dynamical systems are of limited value if there are no quality tags attached which describe the accuracy of the models.

System identification using orthonormal basis functions flourished during the 1990s. This was motivated by efficient parameterizations enabling the use of simple linear regression techniques which produced models with enhanced qualities (Ninness, Hjalmarsson & Gustafsson, 1999). Typical examples of orthonormal basis functions are the pulse basis functions corresponding to the FIR (Finite Impulse Response) models, the Laguerre functions (Wahlberg, 1991), the Kautz functions (Wahlberg, 1994) and more generally orthonormal basis functions as in Heuberger, Van den Hof and Bosgra (1995) and Van den Hof, Heuberger and Bokor (1995).

In this setting the model class is given by

$$G(z^{-1}, \boldsymbol{\theta}) = \sum_{k=1}^{L} \theta_k \mathcal{B}_k(z^{-1}),$$

and the true system is given by

$$G^0(z^{-1}) = \sum_{k=1}^{\infty} \theta_k^0 \mathcal{B}_k(z^{-1}),$$

where $\mathcal{B}_k(z^{-1})$ are the orthonormal basis functions, θ_k, $k = 1, 2, \ldots, L$, are the model parameters and θ_k^0, $k = 1, 2, \ldots$, are the true system parameters. The problem considered in this paper is how to construct a guaranteed confidence set for the true system parameters $\theta_1^0, \theta_2^0, \ldots, \theta_L^0$ based on a finite number of observed data u_t, y_t from the system

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

where n_t is an arbitrary noise sequence.

For this purpose, we extend the LSCR (Leave-out Sign-dominant Correlation Regions) algorithm introduced in Campi and Weyer (2005). See Campi and Weyer (2006a) for an overview. The LSCR algorithm in Campi and Weyer (2005), inspired by the work

☆ The research of M.C. Campi was supported by MIUR under the project "Identification and Adaptive Control of Industrial Systems". The research of S. Ko and E. Weyer was supported by the Australian Research Council under the Discovery Grant Scheme, Project DP0558579. The material in this paper was partially presented at the 46th IEEE Conference on Decision and Control, New Orleans, LA, USA, Dec 12–14, 2007. This paper was recommended for publication in revised form by Associate Editor Wolfgang Scherrer under the direction of Editor Torsten Söderström.

* Corresponding author. Tel.: +61 3 8344 9726; fax: +61 3 8344 6678.
E-mail addresses: marco.campi@ing.unibs.it (M.C. Campi), sanghoko@kau.ac.kr (S. Ko), ewey@unimelb.edu.au (E. Weyer).

0005-1098/$ – see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.automatica.2009.05.028

of Hartigan (1969, 1970), provides *non-asymptotic* confidence sets with a user specified probability for the case where the true transfer functions from the input signal to the output signal and from the noise to the output signal both belong to the model class. In Campi and Weyer (2006b) and Campi and Weyer (in press), the assumption that the noise model belongs to the model class was removed, and in this paper we go one step further and also allow for unmodelled dynamics in the transfer function from the input to the output. The unmodelled dynamics is dealt with by suitable input design and the application of the *sign*-function in the computations of the correlation functions.

The main contribution of this paper is the constructive method for providing *non-asymptotic* confidence regions for the coefficients of the basis functions. A different approach for generating confidence sets is based on asymptotic theory for system identification (see e.g., Ljung (1999) or Söderström and Stoica (1989)). This is a well-matured approach and the confidence regions can be computed relatively easily. However, in some cases the asymptotic approach may lead to unreliable results when applied to a finite number of data points, as described in Garatti, Campi and Bittanti (2004) and Campi and Bittanti (2006). Moreover, asymptotic approaches have little validity when the number of data points is small, and hence finite sample methods as developed in this paper are of great interest. For further discussions on model quality evaluation and confidence sets for the parameters of dynamical systems, the readers are referred to Campi, Ooi, and Weyer (2004), Campi and Weyer (2002), Weyer and Campi (2002), Douma and Van den Hof (2006), Hjalmarson and Ninness (2006) and den Dekker, Bombois and Van den Hof (2008).

In the next subsection we give a simple preview example of the developed procedure which illustrates the main ideas and shows the generality of the approach.

1.1. A preview example

In order to illustrate the main ideas of the paper, we present an introductory toy-example. Suppose that the true system is given by

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

where $\theta_0^0 = 1$, $\theta_1^0 = 0.1$. The noise has been indicated with a generic n_t to signify that it can be arbitrary, and not just a white signal. u_t and n_t are independent. The output y_t of the true system has weaker dependence on the past input u_{t-1} than on the current input u_t, and we assume we want to estimate θ_0^0, the non-dynamical link between u_t and y_t.

Our task is to generate 25 input data and to construct a guaranteed confidence interval for θ_0^0. We first generate an input signal u_t, $t = 1, \ldots, 25$, which is independent and identically distributed (i.i.d.) with

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

and apply it to the system. The input–output data are shown in Fig. 1. We regard the term $\theta_1^0 u_{t-1}$ as unmodelled dynamics and construct a reduced-order predictor

$$\hat{y}_t(\theta) = \theta u_t.$$

The corresponding prediction error is given by

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

We then calculate

$$f_t(\theta) = \text{sign}[u_t \cdot \epsilon_t(\theta)], \quad t = 1, 2, \ldots, 25,$$

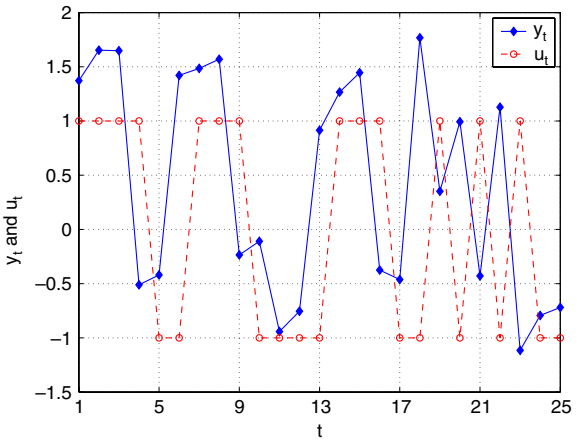

Fig. 1. Data for the preview example.

where the sign-function is defined as

$$\text{sign}[x] = \begin{cases} -1, & \text{for } x < 0, \\ +1, & \text{for } x > 0. \end{cases}$$

If $x = 0$ we let $\text{sign}[x] = 1$ or -1 with probability 0.5 each. Corresponding to the true parameter value, i.e. $\theta = \theta_0^0$, an inspection reveals that $\text{sign}[u_t \cdot \epsilon_t(\theta_0^0)] = \text{sign}[u_t \cdot (\theta_1^0 u_{t-1} + n_t)]$ is an independent and symmetrically distributed process which takes the values ± 1 with probability 0.5 each. Thus, based on this observation, we compute a number of estimates of $E\{\text{sign}[u_t \epsilon_t(\theta)]\}$ 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.

We select 20 subsets of data at random and compute the empirical estimates

$$\bar{g}_i(\theta) = \sum_{t=1}^{25} h_{i,t} f_t(\theta), \quad i = 0, \ldots, 19,$$

where $h_{i,t}$ are i.i.d. with the distribution

$$h_{i,t} = \begin{cases} 0, & \text{with probability 0.5} \\ 1, & \text{with probability 0.5,} \end{cases}$$

except for the first string which is given by $h_{0,t} = 0$ for all $t = 1, 2, \ldots, 25$ (hence $\bar{g}_0(\theta) = 0$). That is, $h_{i,t}$ determines if $f_t(\theta)$ is used when we compute the ith estimate of the correlation. Since it is very unlikely that all the $\bar{g}_i(\theta)$'s have the same sign for the true $\theta = \theta^0$, we discard the regions in parameter space where all functions but at most one are less than the zero function $\bar{g}_0(\theta)$ or greater than the zero function $\bar{g}_0(\theta)$, hence the name of the method: Leave-out Sign-dominant Correlation Regions (LSCR). In this procedure, however, we have neglected a detail which we now describe. Since $f_t(\theta) = \text{sign}[u_t \cdot \epsilon_t(\theta)]$ can only take on the values -1 and 1, it is possible that two or more of the $\bar{g}_i(\theta)$ functions take on the value zero on an interval. This tie and ambiguity can be broken by introducing a random ordering obtained by adding a random number ν_i, uniformly distributed between, say, -0.2 and 0.2 to the $\bar{g}_i(\theta)$ functions

$$g_i(\theta) = \bar{g}_i(\theta) + \nu_i, \quad i = 0, 1, \ldots, 19,$$

and by considering these $g_i(\theta)$ in place of the original $\bar{g}_i(\theta)$ functions. Next we plot $g_i(\theta), i = 0, 1, \ldots, 19$, as functions of θ and exclude the regions where at most one of the functions $g_i(\theta)$, $i = 1, 2, \ldots, 19$, is greater than $g_0(\theta)$ or at most one is smaller than $g_0(\theta)$. The obtained $g_i(\theta)$ functions and the confidence interval are shown in Fig. 2. The confidence interval for θ_0^0 is $\hat{\mathbf{\Theta}}_{25} = [0.80\ \ 1.12]$. It is a rigorous fact (stated in Theorem 1) that the

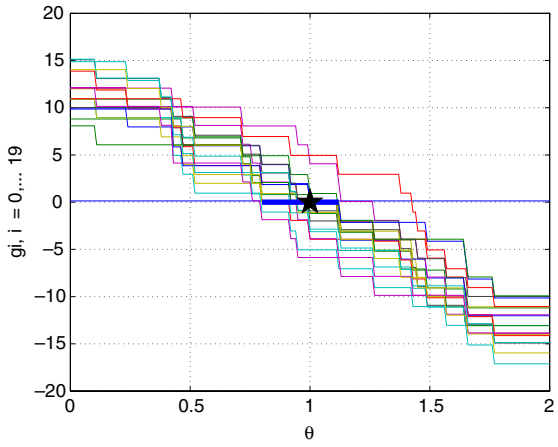

Fig. 2. The $g_i(\theta)$ functions for the preview example together with a 80% confidence interval (thick solid line) and the true parameter ($\star$).

confidence interval constructed this way has probability $1 - 2 \cdot 2/20 = 0.8$ of containing the true parameter value θ_0^0.

In this example, the noise sequence n_t was a realization of a biased independent Gaussian process with mean 0.5 and variance 0.1. However, no knowledge of the noise characteristics was used in the algorithm, nor did we make use of any knowledge about the unmodelled dynamics. Despite the facts that the system is not within the model set, the number of data points is finite, and the noise is biased and with unknown characteristics, the procedure has provided a confidence interval for the true parameter value with guaranteed exact probability.

1.2. The "essence" of the LSCR approach

LSCR is based on constructing data-based functions (usually correlations) that are independent and symmetrically distributed around zero corresponding to the true parameter value $\boldsymbol{\theta}^0$, while they are biased away from zero for $\boldsymbol{\theta} \neq \boldsymbol{\theta}^0$. The $f_t(\theta)$'s of the preview example are examples of such functions. These functions are then summed up in many different ways, as done in the $\bar{g}_i(\theta)$'s of the preview example, leading to sums that for $\boldsymbol{\theta} \neq \boldsymbol{\theta}^0$ are likely to be prevalently positive or prevalently negative. Thus, when constructing the confidence region one eliminates those $\boldsymbol{\theta}$ for which a prevalence of positive or negative values are observed, and this leads to confidence regions that concentrate around the true $\boldsymbol{\theta}^0$. One deep and fundamental aspect in this construction is that we are able to compute with minimal *a priori* assumptions, the (low) exact probability that the true $\boldsymbol{\theta} = \boldsymbol{\theta}^0$ will be erroneously excluded from the confidence region. This is made possible by the random sub-sampling employed in forming the sums (see e.g. the definition of the $\bar{g}_i(\theta)$'s) and by the fact that the data-based functions (the $f_t(\theta)$'s) are independent and symmetrically distributed around zero for $\boldsymbol{\theta} = \boldsymbol{\theta}^0$.

In the context of the present paper, the sign-function is used to secure the above-mentioned independence and symmetry properties in the presence of unmodelled dynamics. Indeed, inspecting the preview example one sees that $u_t \cdot \epsilon_t(\theta_0^0) = u_t \cdot (\theta_1^0 u_{t-1} + n_t)$ is not an independent sequence since e.g. $u_1 \cdot (\theta_1^0 u_0 + n_1)$ and $u_2 \cdot (\theta_1^0 u_1 + n_2)$ share u_1. However, as shown in Appendix A, the sign-function makes the sequence independent and symmetrically distributed.

1.3. Organization of the paper

In Section 2, we develop the algorithm for construction of the confidence regions for the case where the system is approximated by an FIR model. Then, in order to extend the algorithm to the case of generalized orthogonal basis functions, the results of Heuberger, Van den Hof, De Hoog and Wahlberg (2003) are briefly summarized in Section 3.1. In Section 3.2, we show that a system modelled by generalized orthonormal basis functions can be transformed into an FIR model to which the algorithm developed in Section 2 is applicable. The procedure is illustrated in two simulation examples using an FIR and a Laguerre model in Section 4. Finally, some concluding remarks are given in Section 5.

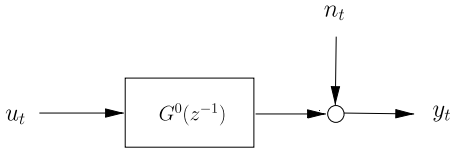

Fig. 3. The dynamical system.

Notations: In this paper, matrices will be denoted by upper case boldface (e.g. $\mathbf{A}$), vectors will be denoted by lower case boldface (e.g. $\mathbf{x}$), and scalars will be denoted by lower case (e.g. y) or upper case (e.g. Y). For a matrix $\mathbf{A}$, $\mathbf{A}^{\mathrm{T}}$ denotes its transpose. $\mathcal{H}_2$ is the space of scalar functions which are analytic for $|z| \geq 1$ and square-integrable on the unit circle.

2. Confidence regions with undermodelling

In this section we present the algorithm for FIR models.

2.1. Problem definition

Data generating system: Consider the following linear time-invariant stable[1] discrete-time system with additive noise as shown in Fig. 3

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

The transfer function $G^0(z^{-1})$ is represented by

$$G^0(z^{-1}) = \sum_{k=1}^{\infty} \theta_k^0 z^{-k}, \tag{2}$$

where θ_k^0, $k = 1, 2, \ldots$, is the sequence of Markov parameters.

We assume that we can choose the input u_t. Moreover, the choice of u_t does not affect n_t, that is $G^0(z^{-1})$ indeed represents the input–output link, whereas n_t represents all other sources of variations in y_t besides u_t. This is formally expressed in the following assumption.

Assumption:

(A1) The noise sequence n_t is independent of u_t.

Model class: For estimation purposes, we consider the following predictor corresponding to an Lth order FIR model

$$\hat{y}_t(\boldsymbol{\theta}) = G(z^{-1}, \boldsymbol{\theta})u_t = \sum_{k=1}^{L} \theta_k z^{-k} u_t = \boldsymbol{\phi}_t^{\mathrm{T}} \boldsymbol{\theta}, \tag{3}$$

where $\boldsymbol{\phi}_t \triangleq [u_{t-1}, u_{t-2}, \ldots, u_{t-L}]^{\mathrm{T}}$ and $\boldsymbol{\theta} \triangleq [\theta_1, \theta_2, \ldots, \theta_L]^{\mathrm{T}}$.

Objective: The goal is to design the input sequence u_t and to construct a confidence set for $\boldsymbol{\theta}^0 \triangleq [\theta_1^0, \theta_2^0, \ldots, \theta_L^0]^{\mathrm{T}}$ based on N observed input–output data. The algorithm for construction of the confidence set should not require any knowledge about the noise characteristics, and the confidence set should contain $\boldsymbol{\theta}^0$ with guaranteed user-chosen probability p.

Not having any assumptions on the noise characteristics is a desired property in applications since in practice the properties of the noise sequence are rarely known.

[1] $G^0(z^{-1}) = \sum_{k=1}^{\infty} \theta_k^0 z^{-k}$ is stable if $\sum_{k=1}^{\infty} |\theta_k^0| < \infty$.

2.2. Construction of confidence regions

We determine confidence regions for θ^0 based on the sign of the correlation between the prediction error $\epsilon_t(\theta)$ and the inputs u_{t-s} for $s = 1, 2, \ldots, L$.

Input design:

(D1) The input signal u_t, for $t = 1, 2, \ldots, N$, is an independent sequence and has equal probability 0.5 of being larger or smaller than zero (see Remark 2 for more general choices of u_t).

We next describe a procedure for constructing confidence regions Θ_s for $s = 1, 2, \ldots, L$. The final confidence region is obtained later on as the intersection of the regions Θ_s.

Procedure for the construction of confidence region Θ_s:

(1) Compute the prediction errors

$$\epsilon_t(\theta) = y_t - \hat{y}_t(\theta) = y_t - \varphi_t^{\mathrm{T}}\theta \tag{4}$$

for $t = 1 + L, 2 + L, \ldots, K + L = N$.

(2) Select an integer $s \in \{1, 2, \ldots, L\}$ and compute

$$f_{t-s,s}(\theta) = \operatorname{sign}[u_{t-s} \cdot \epsilon_t(\theta)]$$

for $t = 1 + L, 2 + L, \ldots, K + L = N$, where sign is defined as

$$\operatorname{sign}[x] = \begin{cases} -1, & \text{if } x < 0 \\ -1, & \text{with probability 0.5 if } x = 0 \\ 1, & \text{with probability 0.5 if } x = 0 \\ 1, & \text{if } x > 0. \end{cases}$$

(3) Calculate empirical correlation functions through the following procedure. Select an integer M and construct M binary, i.e. $\{0, 1\}$-valued, stochastic strings of length K as follows: Let $h_{0,1}, h_{0,2}, \ldots, h_{0,K}$ be the string of all zeros. Every element of the remaining strings takes the value 0 or 1 with probability 0.5 each, and the elements are independent of each other. 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 $h_{1,1}, h_{1,2}, \ldots, h_{1,K}; h_{2,1}, h_{2,2}, \ldots, h_{2,K}; \ldots; h_{M-1,1}, h_{M-1,2}, \ldots, h_{M-1,K}$. Compute

$$\bar{g}_{i,s}(\theta) = \sum_{t=1+L}^{N} h_{i,t-L} \cdot f_{t-s,s}(\theta), \quad i = 0, 1, \ldots, M-1.$$

(4) Add a small random number $\nu_{i,s}$ uniformly distributed on $[-\alpha, \alpha]$ with $0 < \alpha < 0.5$ to each correlation function:

$$g_{i,s}(\theta) = \bar{g}_{i,s}(\theta) + \nu_{i,s}, \quad i = 0, 1, \ldots, M-1.$$

The addition of $\nu_{i,s}$ prevents ties from occurring in the next step.

(5) Select an integer q in the interval $[1, M/2]$ and find the region Θ_s such that at least q of the $g_{i,s}(\theta)$ functions are greater than the function $g_{0,s}(\theta) = \nu_{0,s}$ and at least q are smaller than $g_{0,s}(\theta) = \nu_{0,s}$.

The final confidence set $\hat{\Theta}_N$ is obtained by intersecting the sets Θ_s for $s = 1, 2, \ldots, L$, i.e.

$$\hat{\Theta}_N = \bigcap_{s=1}^{L} \Theta_s.$$

One implementation aspect which is important to note is that constructing the set Θ_s as indicated in the procedure can be very hard. What instead is easy is to pick a θ value and verify through points (1)–(5) whether this θ belongs to Θ_s. This suggests that Θ_s can be constructed by first gridding the θ space and then verify one by one whether or not the θ's on the grid are in Θ_s. While this method is practical for problems of low dimensionality (θ has few elements, i.e. L is small), it becomes computationally intensive for problems where θ has many elements. This issue must necessarily be given attention in future research.

2.3. Properties of the confidence regions

In this section we identify several important properties of the confidence regions.

As in Campi and Weyer (2006b); Campi and Weyer (in press), there are no assumptions on the noise. The main distinguishing feature of the current problem setting is that the true system does not belong to the model class. Therefore, while the correlation function $u_{t-s} \cdot \epsilon_t(\theta^0)$ evaluated at the true parameter was an independent and symmetrically distributed sequence in the settings in Campi and Weyer (2005); Campi and Weyer (2006b); Campi and Weyer (in press), this is no longer the case in the present setting. However, by taking the *sign*-function of the correlation function, $\operatorname{sign}[u_{t-s} \cdot \epsilon_t(\theta)]$ evaluated at $\theta = \theta^0$ becomes a sequence of independent random variables symmetrically distributed around zero (see Appendix A). Therefore, for the true value $\theta = \theta^0$ it is unlikely that nearly all of the correlation functions $g_{i,s}(\theta)$ are positive or negative, and those regions in parameter space where this happens are therefore excluded from the confidence sets in point (5) of the procedure. This fact in conjunction with the way the $g_{i,s}(\theta)$ functions are constructed in point (4) of the procedure results in the following theorem.

Theorem 1. *The set Θ_s constructed in point* (5) *of the procedure has the property that*

$$\Pr\left\{\theta^0 \in \Theta_s\right\} = 1 - \frac{2q}{M},$$

where M and q are chosen by the user in points (3) *and* (5) *of the procedure.*

Proof. See Appendix A.

$1 - 2q/M$ is the exact probability that $\theta^0 \in \Theta_s$, therefore $\theta^0 \notin \Theta_s$ with probability $2q/M$. As $\hat{\Theta}_N$ is obtained by intersecting Θ_s, $s = 1, \ldots, L$, it follows that $\theta^0 \notin \hat{\Theta}_N$ with probability at most $2qL/M$, and we have the following theorem.

Theorem 2.

$$\Pr\left\{\theta^0 \in \hat{\Theta}_N\right\} \geq 1 - \frac{2Lq}{M}.$$

It can be shown, as formally stated in Theorem 3, that the constructed region concentrates around the true parameter θ^0 in the sense that any $\theta \neq \theta^0$ will eventually be excluded from the confidence set as the number of data points increases, provided that the following additional assumptions hold.

(A2) The input u_t and noise n_t have probability density functions such that both can be arbitrarily small with non-zero probabilities.

(A3) The input u_t and noise n_t sequences are strict-sense stationary and strict ergodic.[2]

Assumption (A3) ensures that sample means converge to expected values, and Assumption (A2) ensures that the mismatch between θ and θ^0 for $\theta \neq \theta^0$ will give rise to a bias also after the sign-function is applied in point (2) of the procedure.

Theorem 3. *Under assumptions* (A1)–(A3)*, for every fixed $\theta \neq \theta^0$*

$$\Pr\left\{\exists \bar{N} | \theta \notin \hat{\Theta}_N, \forall N > \bar{N}\right\} = 1. \tag{5}$$

[2] Independent and identically distributed sequences are strict sense stationary and strict ergodic (Stout, 1974, Lemma 3.5.8.).

Proof. See Appendix B. □

Remark 1 (*Correlation with Generalized Instruments*). Point (2) in the procedure for the construction of confidence regions can be generalized: Instead of correlating the prediction errors with delayed inputs we can correlate them with an independent sequence ξ_t, i.e. $f_{t-s,s}(\boldsymbol{\theta}) = \text{sign}[\xi_{t-s} \cdot \epsilon_t(\boldsymbol{\theta})]$ where $\epsilon_t(\boldsymbol{\theta})$ is as before given by (4). Provided that ξ_t is independent of the noise and has equal probability of being larger and smaller than 0, Theorems 1 and 2 remain valid. In order for the confidence set to shrink around $\boldsymbol{\theta}^0$, ξ_t must also be sufficiently correlated with the input. This accommodates situations where we cannot choose the system input u_t, but a signal correlated with u_t is available. ■

Remark 2 (*Non-white Input Case*). The input designed in (D1) is white. At times having a smoother signal is more advisable for real systems. In this case, we can consider applying a filtered input $F(z^{-1})u_t$ where $F(z^{-1})$ is a known stable filter with a stable inverse. We then obtain

$$y_t = G^0(z^{-1}) \cdot F(z^{-1})u_t + n_t.$$

The output y_t is then filtered with the inverse filter $F^{-1}(z^{-1})$, so that

$$F^{-1}(z^{-1})y_t = G^0(z^{-1})u_t + F^{-1}(z^{-1})n_t. \tag{6}$$

We can then employ the LSCR method using the input signal u_t and the filtered output signal $F^{-1}(z^{-1})y_t$. The fact that the noise is filtered through $F^{-1}(z^{-1})$ in (6) is not a problem since the procedure holds without any assumptions on the noise sequence (apart from being independent of u_t). ■

2.4. Extension to the multi-variable case

The procedure described in the previous section can easily be extended to the MIMO (Multi-Input–Multi-Output) case by considering each output separately. Consider the following MIMO (m-input p-output) system

$$\mathbf{y}_t = \sum_{k=1}^{\infty} \boldsymbol{\Xi}_k^0 \mathbf{u}_{t-k} + \mathbf{n}_t,$$

where

$$\mathbf{y}_t = \begin{bmatrix} y_{1,t} \\ \vdots \\ y_{p,t} \end{bmatrix}, \qquad \mathbf{u}_t = \begin{bmatrix} u_{1,t} \\ \vdots \\ u_{m,t} \end{bmatrix},$$

$$\mathbf{n}_t = \begin{bmatrix} n_{1,t} \\ \vdots \\ n_{p,t} \end{bmatrix}, \qquad \boldsymbol{\Xi}_k^0 = \begin{bmatrix} \theta_{11,k}^0 & \cdots & \theta_{1m,k}^0 \\ \vdots & \ddots & \vdots \\ \theta_{p1,k}^0 & \cdots & \theta_{pm,k}^0 \end{bmatrix}.$$

We use a predictor corresponding to an FIR model

$$\hat{\mathbf{y}}_t(\boldsymbol{\Xi}) = \sum_{k=1}^{L} \boldsymbol{\Xi}_k \mathbf{u}_{t-k}.$$

The prediction error $\boldsymbol{\epsilon}_t(\boldsymbol{\Xi}) = \left[\epsilon_{1,t}, \epsilon_{2,t}, \ldots, \epsilon_{p,t}\right]^{\mathrm{T}} = \mathbf{y}_t - \hat{\mathbf{y}}_t(\boldsymbol{\Xi})$ is given by

$$\epsilon_{j,t}(\boldsymbol{\Xi}) = \sum_{k=1}^{L} \left(\tilde{\theta}_{j1,k}u_{1,t-k} + \cdots + \tilde{\theta}_{jm,k}u_{m,t-k}\right) + \sum_{k=L+1}^{\infty} \left(\theta_{j1,k}^0 u_{1,t-k} + \cdots + \theta_{jm,k}^0 u_{m,t-k}\right) + n_{j,t},$$

for $j = 1, 2, \ldots, p$.

Under the assumption that the input vector sequence $\mathbf{u}_t$, $t = 1, 2, \ldots, N$, is independent in time and each element has equal probability 0.5 of being larger or smaller than 0 and considering the functions

$$f_{t-s,s}^{i,j}(\boldsymbol{\Xi}) = \text{sign}\left[u_{i,t-s} \cdot \epsilon_{j,t}(\boldsymbol{\Xi})\right],$$

for $i = 1, 2, \ldots, m$, $j = 1, 2, \ldots, p$, $s = 1, 2, \ldots, L$, we can construct guaranteed confidence regions for the parameters $\{\theta_{ji,k}^0\}$, for $i = 1, 2, \ldots, m$, $j = 1, 2, \ldots, p$, $k = 1, 2, \ldots, L$, by employing the LSCR algorithm developed in Section 2.2. Theorems 1 and 2 still hold with obvious modifications.

3. More general model classes

It is well known (Heuberger et al., 1995; Wahlberg, 1991) that using the pulse basis functions for approximation of moderately damped systems or systems with high sampling rates leads to approximations of high order. To deal with these situations, several orthonormal basis functions which incorporate prior system information have been suggested, e.g. the Laguerre functions (Wahlberg, 1991) and the Kautz functions (Wahlberg, 1994) which are both special cases of the generalized orthonormal basis functions introduced in Heuberger et al. (1995) and Van den Hof et al. (1995). In this section, we first briefly describe these generalized orthonormal basis functions and then extend the results from the previous section to cover series expansions using these basis functions.

3.1. Generalized orthonormal basis functions

The theorem below describes the generalized orthonormal basis functions. The theorem follows from the results and discussions leading up to Definition 3.1 in Heuberger et al. (2003).

Theorem 4. *Let $\mathcal{A}(z^{-1})$ be a stable all-pass transfer function with an internally balanced realization $(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D})$ of order $n > 0$. Denote*

$$\boldsymbol{\mathcal{B}}_k(z^{-1}) = \boldsymbol{\mathcal{B}}(z^{-1})\mathcal{A}^{k-1}(z^{-1}),$$

where $\boldsymbol{\mathcal{B}}(z^{-1}) = (z\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}$. Then, the set of all scalar elements of the vectors $\boldsymbol{\mathcal{B}}_k(z^{-1})$, $k = 1, 2, \ldots$, form an orthonormal set in $\mathcal{H}_2$, and for every strictly proper transfer function $H(z^{-1}) \in \mathcal{H}_2$, there exists a unique row vector sequence $\mathbf{L}_k$, $k = 1, 2, \ldots$, with $\sum_{k=1}^{\infty} \|\mathbf{L}_k\|^2 < \infty$, such that

$$H(z^{-1}) = \sum_{k=1}^{\infty} \mathbf{L}_k \boldsymbol{\mathcal{B}}_k(z^{-1}) = \sum_{k=1}^{\infty} \left(L_{k,1}\mathcal{B}_{k,1}(z^{-1}) + \cdots + L_{k,n}\mathcal{B}_{k,n}(z^{-1})\right).$$

One can construct an all-pass transfer function $\mathcal{A}(z^{-1})$ from any given set of stable poles, and thus the basis can incorporate dynamics of any complexity, combining, for example, both fast and slow dynamics in damped and resonant modes. This allows for the construction of simplified dynamic models by incorporating *a priori* knowledge about the system dynamics into the basis.

The pulse, Laguerre, and Kautz functions are special cases of the generalized orthonormal basis functions as shown next.

Pulse functions: Using the all-pass transfer function $\mathcal{A}(z^{-1}) = z^{-1}$ with balanced realization $(A, B, C, D) = (0, 1, 1, 0)$, we obtain the standard pulse basis

$$\mathcal{B}_k(z^{-1}) = \mathcal{A}^k(z^{-1}) = z^{-k}.$$

Laguerre functions: Using the all-pass transfer function $\mathcal{A}(z^{-1}) = (1-az)/(z-a)$ for some real-valued a with $|a| < 1$, and balanced realization

$$(A, B, C, D) = (a, \sqrt{1-a^2}, \sqrt{1-a^2}, -a),$$

the Laguerre basis is obtained (Wahlberg, 1991):

$$\mathcal{B}_k(z^{-1}) = \sqrt{1-a^2}z\frac{(1-az)^{k-1}}{(z-a)^k}.$$

Kautz functions: Using the all-pass transfer function $\mathcal{A}(z^{-1}) = \frac{-cz^2+b(c-1)z+1}{z^2+b(c-1)z-c}$ for some real-valued b, c with $|b|$, $|c| < 1$, and a balanced realization

$$\mathbf{A} = \begin{bmatrix} b & \sqrt{1-b^2} \\ c\sqrt{1-b^2} & -bc \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} 0 \\ \sqrt{1-c^2} \end{bmatrix},$$

$$\mathbf{C} = \begin{bmatrix} \gamma_1 & \gamma_2 \end{bmatrix}, \qquad D = -c$$

with $\gamma_1 = \sqrt{(1-c^2)(1-b^2)}$ and $\gamma_2 = -b\sqrt{1-c^2}$, we get

$$\boldsymbol{\mathcal{B}}_k(z^{-1}) = \frac{\sqrt{1-c^2}(-cz^2+b(c-1)z+1)^{k-1}}{(z^2+b(c-1)z-c)^k}\begin{bmatrix} \sqrt{1-b^2} \\ z-b \end{bmatrix},$$

and the Kautz functions follow (Wahlberg, 1994).

3.2. Generalized FIR models

In this section we convert the models, which are series expansions in the above generalized basis functions, into FIR models using a filtering procedure. After the conversion, the results in Section 2 can be used to generate confidence regions.

In the approach so far, data has been considered as a sequence in time. Specifically the output data points were

$$y_1, y_2, \ldots, y_{N-1}, y_N.$$

We next introduce a notation which clearly distinguishes between a sequence and a specific element of the sequence. We let $\{y_t\}$ denote a sequence and $\{y_t\}_{t=k}$ an element of the sequence. With this notation the above data points can be written as

$$\{y_t\}_{t=1}, \{y_t\}_{t=2}, \ldots, \{y_t\}_{t=N-1}, \{y_t\}_{t=N}$$

or

$$\{y_{t-N+1}\}_{t=N}, \{y_{t-N+2}\}_{t=N}, \ldots, \{y_{t-1}\}_{t=N}, \{y_t\}_{t=N}, \tag{7}$$

where $\{y_{t-k}\}$ is the sequence $\{y_t\}$ shifted by k, i.e. $\{y_{t-k}\} = z^{-k}\{y_t\}$. (7) can also be written as

$$\{z^{-N+1}\{y_t\}\}_{t=N}, \{z^{-N+2}\{y_t\}\}_{t=N}, \ldots, \{z^{-1}\{y_t\}\}_{t=N}, \{y_t\}_{t=N}. \tag{8}$$

Note that z^{-1} is an all-pass filter, and (8) is a special case of data obtained by repeated filtering with an all-pass filter $\mathcal{A}(z^{-1}) = z^{-1}$. More generally an arbitrary all-pass filter $\mathcal{A}(z^{-1})$ can be used leading to the data

$$\{\mathcal{A}^{N-1}(z^{-1})\{y_t\}\}_{t=N}, \{\mathcal{A}^{N-2}(z^{-1})\{y_t\}\}_{t=N}, \ldots, \{\mathcal{A}(z^{-1})\{y_t\}\}_{t=N}, \{y_t\}_{t=N}. \tag{9}$$

This observation forms the starting point for rewriting the models in the orthonormal basis functions as generalized FIR models. Note that from Theorem 4, $G^0(z^{-1})$ can be written as

$$G^0(z^{-1}) = \sum_{k=1}^{\infty}\boldsymbol{\theta}_k^0\boldsymbol{\mathcal{B}}_k(z^{-1}) = \sum_{k=1}^{\infty}\boldsymbol{\theta}_k^0\boldsymbol{\mathcal{B}}(z^{-1})\mathcal{A}^{k-1}(z^{-1}).$$

Therefore, the last datum in (9) is given by

$$\begin{aligned}\tilde{y}_N &:= \{y_t\}_{t=N} = \boldsymbol{\theta}_1^0\{\boldsymbol{\mathcal{B}}(z^{-1})\{u_t\}\}_{t=N} \\ &\quad + \boldsymbol{\theta}_2^0\{\boldsymbol{\mathcal{B}}(z^{-1})\mathcal{A}(z^{-1})\{u_t\}\}_{t=N} + \cdots + \{n_t\}_{t=N} \\ &= \boldsymbol{\theta}_1^0\tilde{\mathbf{u}}_{N-1} + \boldsymbol{\theta}_2^0\tilde{\mathbf{u}}_{N-2} + \cdots + \tilde{n}_N,\end{aligned}$$

where

$$\tilde{\mathbf{u}}_j = \{\boldsymbol{\mathcal{B}}(z^{-1})\mathcal{A}^{N-(j+1)}(z^{-1})\{u_t\}\}_{t=N}.$$

Similarly

$$\begin{aligned}\tilde{y}_{N-1} &:= \{\mathcal{A}(z^{-1})\{y_t\}\}_{t=N} \\ &= \boldsymbol{\theta}_1^0\{\boldsymbol{\mathcal{B}}(z^{-1})\mathcal{A}(z^{-1})\{u_t\}\}_{t=N} + \boldsymbol{\theta}_2^0\{\boldsymbol{\mathcal{B}}(z^{-1})\mathcal{A}^2(z^{-1})\{u_t\}\}_{t=N} \\ &\quad + \cdots + \{\mathcal{A}(z^{-1})\{n_t\}\}_{t=N} \\ &= \boldsymbol{\theta}_1^0\tilde{\mathbf{u}}_{N-2} + \boldsymbol{\theta}_2^0\tilde{\mathbf{u}}_{N-3} + \cdots + \tilde{n}_{N-1}\end{aligned}$$

and, in general,

$$\begin{aligned}\tilde{y}_j &:= \{\mathcal{A}^{N-j}(z^{-1})\{y_t\}\}_{t=N} = \boldsymbol{\theta}_1^0\tilde{\mathbf{u}}_{j-1} + \boldsymbol{\theta}_2^0\tilde{\mathbf{u}}_{j-2} + \cdots + \tilde{n}_j \\ &= \sum_{k=1}^{\infty}\boldsymbol{\theta}_k^0\tilde{\mathbf{u}}_{j-k} + \tilde{n}_j\end{aligned} \tag{10}$$

where

$$\tilde{n}_j = \{\mathcal{A}^{N-j}(z^{-1})\{n_t\}\}_{t=N}.$$

Expression (10) generalizes (1) and (2) to the case where a generic all-pass filter is used in place of the shift operator. Fig. 4 illustrates how the data $\tilde{\mathbf{u}}_j, \tilde{y}_j$, $j = 1, \ldots, N$, are obtained through the successive filtering with the all-pass filter $\mathcal{A}(z^{-1})$.

A reduced-order predictor is now given by

$$\begin{aligned}\hat{\tilde{y}}_j(\boldsymbol{\theta}) &= \sum_{k=1}^{L}\boldsymbol{\theta}_k\tilde{\mathbf{u}}_{j-k} \\ &= \sum_{k=1}^{L}(\theta_{1,k}\tilde{u}_{1,j-k} + \theta_{2,k}\tilde{u}_{2,j-k} + \cdots + \theta_{n,k}\tilde{u}_{n,j-k}),\end{aligned}$$

where $\theta_{l,k}$ and $\tilde{u}_{l,j-k}$ denote the elements of the $\boldsymbol{\theta}_k$ and $\tilde{\mathbf{u}}_{j-k}$ vectors.

The filtered inputs $\tilde{\mathbf{u}}_j$ should have properties similar to those stated in (D1), (A2) and (A3) for u_t in the case of pulse basis functions. For this to happen, we modify the design of u_t as given next in (D2). Proposition 1 below then proves that $\tilde{\mathbf{u}}_j$ does exhibit the desired properties.

(D2) The input u_t is a zero-mean white Gaussian signal with spectral density $\Phi_u > 0$.

The main difference between this input design (D2) and (D1) is that in (D2) u_t is Gaussian. Thanks to Gaussianity, independence is preserved after the filtering with an all-pass filter, as done in this section.

Proposition 1. *With the input design in* (D2)*, the filtered inputs* $\tilde{\mathbf{u}}_j$, $j = 1, \ldots, N$, *are independent and identically distributed zero-mean Gaussian, and, moreover, the elements of* $\tilde{\mathbf{u}}_j$ *are mutually independent and each element has probability* 0.5 *of being larger or smaller than zero.*

Proof. See Appendix C.

We further assume that the noise $\tilde{n}_j$ satisfies

(A5) $\tilde{n}_j$ is strict sense stationary and strict ergodic and independent of $\tilde{\mathbf{u}}_j$. Moreover, $\tilde{n}_j$ has a probability density such that it can be arbitrarily small with non-zero probability.

From Proposition 1 and Assumption (A5) it follows that the new input $\tilde{\mathbf{u}}_j$ and noise $\tilde{n}_j$ satisfy the assumptions on u_t and n_t in Section 2 and therefore the theory of Section 2 can be applied to the present situation. One aspect deserves to be explicitly mentioned though. $\tilde{\mathbf{u}}_j$ is a vector in the general case, so all the functions in the algorithm, and specifically $g_{i,s}(\boldsymbol{\theta})$, become vector functions. The confidence set $\boldsymbol{\Theta}_s$ is now obtained by taking the intersection of the confidence sets obtained by considering the n scalar elements of $g_{i,s}(\boldsymbol{\theta})$ separately.

The following corollary summarizes the properties of the confidence regions.

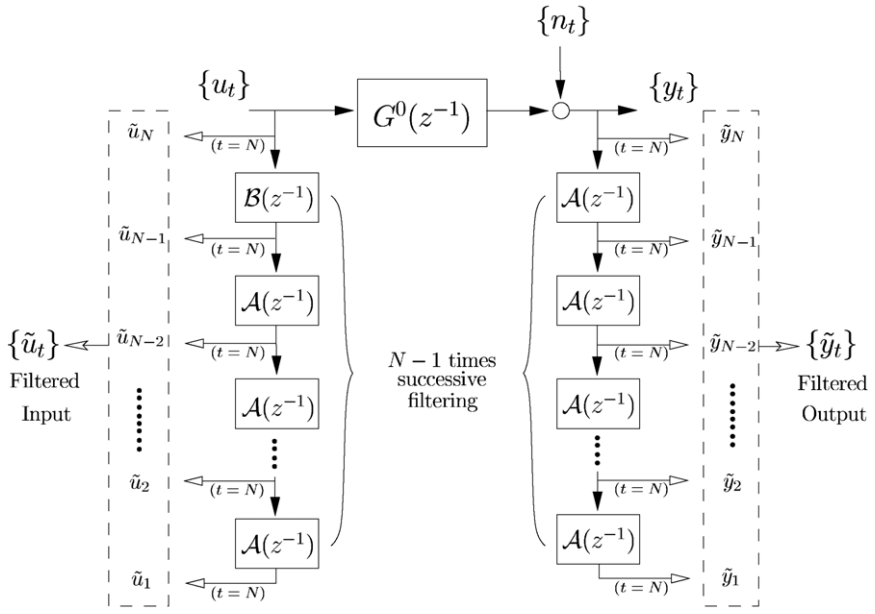

Fig. 4. Generation of the filtered data set $\{\tilde{\mathbf{u}}_t, \tilde{y}_t\}_{t=1,2,\ldots,N}$ from the original data set $\{u_t, y_t\}_{t=1,2,\ldots,N}$.

Corollary 1. *Assume that the input has been designed according to* (D2) *and that assumption* (A5) *holds. Then the set* $\boldsymbol{\Theta}_s$ *has the property*

$$\Pr\left\{\boldsymbol{\theta}^0 \in \boldsymbol{\Theta}_s\right\} \geq 1 - 2qn/M,$$

and the set $\hat{\boldsymbol{\Theta}}_N = \bigcap_{s=1}^{L} \boldsymbol{\Theta}_s$ *has the property that*

$$\Pr\left\{\boldsymbol{\theta}^0 \in \hat{\boldsymbol{\Theta}}_N\right\} \geq 1 - 2Lqn/M.$$

Furthermore, for every fixed $\boldsymbol{\theta} \neq \boldsymbol{\theta}^0$,

$$\Pr\left\{\exists \bar{N} | \boldsymbol{\theta} \notin \hat{\boldsymbol{\Theta}}_N, \forall N > \bar{N}\right\} = 1.$$

For simplicity we have stated the Corollary for the case where the number of scalar basis functions and hence the number of parameters, is a multiple of n, the order of the all-pass filter. The Corollary holds with obvious modifications when a model class for which the number of scalar basis functions is not a multiple of n is used.

Remark 3 (*Effects of Initial Conditions*). Performing the successive filtering illustrated in Fig. 4 requires information about past input and output $u_t, y_t,\ t \leq 0$, and, with an arbitrary initialization of the filters, the results of this section do not hold true rigorously. If, for example, the filters generating $\tilde{\mathbf{u}}_j$ are arbitrary initialized, then $\tilde{\mathbf{u}}_j$ is not an independent sequence. However, after a transient, the magnitude of the tail becomes negligible and the filtered input can in practice often be treated as independent. Note however that the transient reduces the number of data points that in practice can be used, and the longer the transient (e.g., for all-pass filters with poles close to the unit circle) the fewer the available data points. Note also that the initial conditions do not affect the results when the pulse basis functions are used since they have a finite impulse response. ■

4. Numerical example

In this section, we present two simulation examples which illustrate the algorithms developed in this paper.

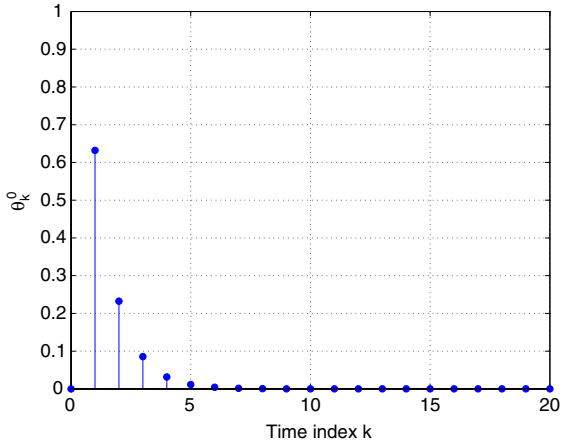

Fig. 5. First 20 coefficients of pulse basis functions.

4.1. FIR model

Consider the linear time-invariant discrete-time system

$$y_t = G^0(z^{-1})u_t + n_t, \tag{11}$$

where the true transfer function

$$G^0(z^{-1}) = \frac{0.6321}{z - 0.3679} \tag{12}$$

was obtained from a continuous-time transfer function

$$G_c^0(s) = \frac{1}{s+1}$$

by discretizing it with a zero-order-hold and sampling period 1 second. The transfer function (12) has a rapidly decaying impulse response sequence as shown in Fig. 5. The measurement noise sequence n_t was a biased white Gaussian sequence with mean 0.2 and variance 0.05. The above information is only given for completeness and it was not used in the algorithm.

The input sequence u_t was chosen as a zero-mean white Gaussian sequence with variance 1. As a model class, we used second-order FIR models

$$\hat{y}_t(\boldsymbol{\theta}) = \theta_1 u_{t-1} + \theta_2 u_{t-2}, \quad \boldsymbol{\theta} \triangleq [\theta_1 \quad \theta_2]^{\mathrm{T}}.$$

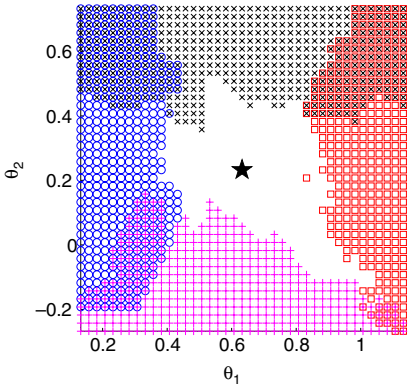

Fig. 6. Non-asymptotic 95% confidence region for (θ_1^0, θ_2^0) (blank region) using 50 data points. ★ = true parameter.

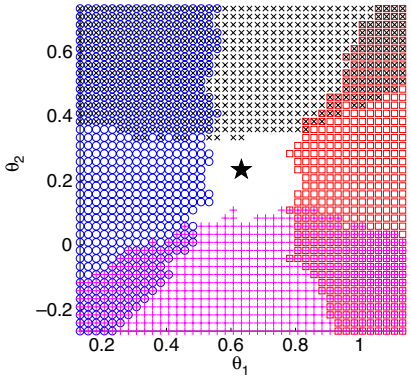

Fig. 7. Non-asymptotic 95% confidence region for (θ_1^0, θ_2^0) (blank region) using 100 data points. ★ = true parameter.

To construct confidence regions for θ_1^0 and θ_2^0, we considered $N = 50, 100, 500$ and 2000 data points generated according to (11) and (12), and computed

$$f_{t-1,1}(\boldsymbol{\theta}) = \text{sign}\left[u_{t-1} \cdot (y_t - \hat{y}_t(\boldsymbol{\theta}))\right], \quad t = 3, 4, \ldots, N,$$
$$f_{t-2,2}(\boldsymbol{\theta}) = \text{sign}\left[u_{t-2} \cdot (y_t - \hat{y}_t(\boldsymbol{\theta}))\right], \quad t = 3, 4, \ldots, N.$$

Then, we computed the following $M = 400$ empirical correlation functions

$$g_{i,1}(\boldsymbol{\theta}) = \sum_{t=3}^{N} h_{i,t-2} \cdot f_{t-1,1}(\boldsymbol{\theta}) + \nu_{i,1}, \quad i = 0, 1, \ldots, M-1,$$
$$g_{i,2}(\boldsymbol{\theta}) = \sum_{t=3}^{N} h_{i,t-2} \cdot f_{t-2,2}(\boldsymbol{\theta}) + \nu_{i,2}, \quad i = 0, 1, \ldots, M-1.$$

Here $\nu_{i,1}$ and $\nu_{i,2}$ were uniformly distributed on $[-0.1,\ 0.1]$. We excluded the regions in parameter space where at most 4 (out of the $M = 400$) $g_{i,1}(\boldsymbol{\theta})$ (or $g_{i,2}(\boldsymbol{\theta})$) functions were smaller or greater than the $g_{0,1}(\boldsymbol{\theta})$ (or $g_{0,2}(\boldsymbol{\theta})$) function. The obtained confidence regions are the blank areas in Figs. 6–9. The regions constructed this way have a probability of at least $1 - 2 \cdot 2 \cdot 5/400 = 0.95$ of containing the true parameter. The true value is marked with ★. The regions where at most 4 values of $g_{i,1}(\boldsymbol{\theta})$ functions were smaller than $g_{0,1}(\boldsymbol{\theta})$ are marked with ◯, and the region where at most 4 were greater than $g_{0,1}(\boldsymbol{\theta})$ are marked with □. Likewise for $g_{i,2}(\boldsymbol{\theta})$, where + and × represent the regions where at most 4 values of $g_{i,2}(\boldsymbol{\theta})$ were smaller or greater than $g_{0,2}(\boldsymbol{\theta})$. As we can see, each step in the construction of the confidence region excludes a particular region. As expected, the size of the region decreases as the number of data points gets larger. Note that the axes have been re-scaled between Figs. 7 and 8.

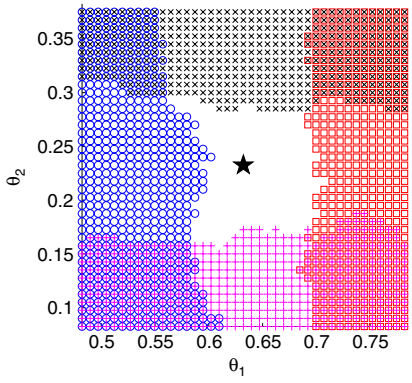

Fig. 8. Non-asymptotic 95% confidence region for (θ_1^0, θ_2^0) (blank region) using 500 data points. ★ = true parameter.

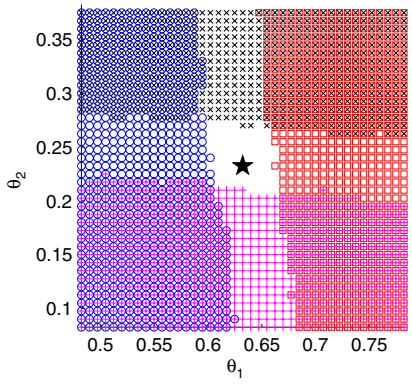

Fig. 9. Non-asymptotic 95% confidence region for (θ_1^0, θ_2^0) (blank region) using 2000 data points. ★ = true parameter.

4.2. Laguerre model

We now consider a system with a more slowly decaying impulse response sequence. The transfer function $G^0(z^{-1})$ is now given by

$$G^0(z^{-1}) = \frac{0.155z + 0.094}{(z - 0.607)(z - 0.368)} = \frac{n_1}{z - p_1} + \frac{n_2}{z - p_2}$$

where $n_1 = 0.787, n_2 = -0.632, p_1 = 0.607$, and $p_2 = 0.368$. This system was obtained from a continuous-time transfer function

$$G_c^0(s) = \frac{1}{(2s + 1)(s + 1)}$$

by discretizing it with a zero-order-hold and a sampling period of 1 s. The input u_t and the measurement noise n_t were zero-mean white Gaussian with variance 1 and 0.05 respectively. $N = 4000$ input–output data points were generated.

The transfer function $G^0(z^{-1})$ can be expanded using the pulse and Laguerre basis functions. The coefficients of the respective basis functions are given by

$$\theta_k^0(\text{pulse}) = \sum_{\ell=1}^{2} n_\ell p_\ell^{k-1}, \quad \text{for } k > 0$$

$$\theta_k^0(\text{Laguerre}) = \sum_{\ell=1}^{2} \frac{n_\ell \sqrt{(1 - a^2)}}{1 - a \cdot p_\ell} \left[\frac{p_\ell - a}{1 - a \cdot p_\ell}\right]^{k-1}, \quad \text{for } k > 0,$$

where $a = 0.5$ is the pole location of the Laguerre basis functions. This pole lies between the true system poles. The first 20 coefficients are shown in Fig. 10, where we observe that for the Laguerre basis functions the first two terms are dominant. The coefficients of the pulse basis functions on the other hand are slowly decaying.

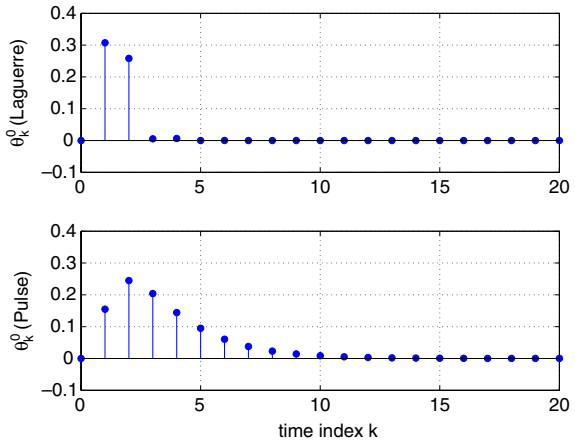

Fig. 10. First 20 coefficients of Laguerre basis functions with $a = 0.5$ (top) and pulse basis functions (bottom).

We chose a second-order Laguerre model for the system. Its predictor is given by

$$\hat{y}_t(\boldsymbol{\theta}) = \theta_1 \mathcal{B}_1(z^{-1})u_t + \theta_2 \mathcal{B}_2(z^{-1})u_t,$$

where

$$\mathcal{B}_k(z^{-1}) = \mathcal{B}(z^{-1})\mathcal{A}^{k-1}(z^{-1})$$

with

$$\mathcal{B}(z^{-1}) = \frac{\sqrt{1-a^2}}{z-a}, \qquad \mathcal{A}(z^{-1}) = \frac{1-az}{z-a}, \quad a = 0.5.$$

To obtain the confidence region for θ_1^0 and θ_2^0, we applied the successive filtering explained in Section 3.2 to the original data. Zero initial conditions were used in the filters.

We used the last 1000 filtered data points and computed, for $j = 3003, 3004, \ldots, 4000$,

$$f_{j-1,1}(\boldsymbol{\theta}) = \operatorname{sign}\left[\tilde{u}_{j-1} \cdot \epsilon_j(\boldsymbol{\theta})\right],$$
$$f_{j-2,2}(\boldsymbol{\theta}) = \operatorname{sign}\left[\tilde{u}_{j-2} \cdot \epsilon_j(\boldsymbol{\theta})\right],$$

and obtained $M = 960$ empirical correlation functions

$$g_{i,1}(\boldsymbol{\theta}) = \sum_{j=3003}^{4000} h_{i,j-3002} \cdot f_{j-1,1}(\boldsymbol{\theta}) + \nu_{i,1}, \quad i = 0, 1, \ldots, 959,$$

$$g_{i,2}(\boldsymbol{\theta}) = \sum_{j=3003}^{4000} h_{i,j-3002} \cdot f_{j-2,2}(\boldsymbol{\theta}) + \nu_{i,2}, \quad i = 0, 1, \ldots, 959.$$

Here $\nu_{i,1}$ and $\nu_{i,2}$ were uniformly distributed on $[-0.1,\ 0.1]$. We excluded the regions in parameter space where at most 11 (out of $M = 960$) $g_{i,1}(\boldsymbol{\theta})$ (or $g_{i,2}(\boldsymbol{\theta})$) functions were smaller or greater than the $g_{0,1}$ (or $g_{0,2}$) function. The obtained confidence region is the blank area in Fig. 11, and it contains the true parameter with probability at least 0.95. The region where at most 11 values of $g_{i,1}(\boldsymbol{\theta})$ functions were smaller than $g_{0,1}(\boldsymbol{\theta})$ is marked with ◯, and the region where at most 11 were greater than $g_{0,1}(\boldsymbol{\theta})$ is marked with □. Likewise for $g_{i,2}(\boldsymbol{\theta})$, where + and × represent the regions where at most 11 values of $g_{i,2}(\boldsymbol{\theta})$ were smaller or greater than $g_{0,2}(\boldsymbol{\theta})$.

5. Concluding remarks

In this paper, we have extended the LSCR algorithm for construction of non-asymptotic confidence regions to the case where undermodelling is present. The systems are approximated by generalized orthonormal basis functions models, and by applying the sign-function in the computations of the correlation functions, guaranteed non-asymptotic confidence regions can be constructed. Remarkably, the method does not make use of any assumptions on the noise or on the decay rate of the unmodelled dynamics. The method was first developed for FIR models and then extended to models represented by generalized orthonormal basis functions through a filtering procedure.

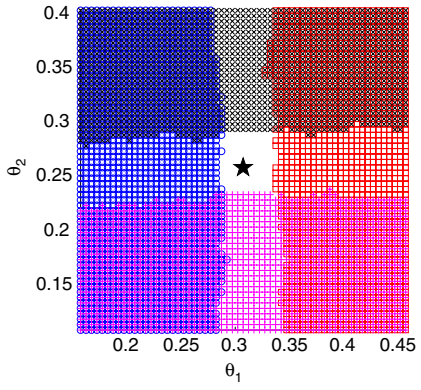

Fig. 11. Non-asymptotic 95% confidence region for (θ_1^0, θ_2^0) (blank region) using 1000 filtered data points. ★ = true parameter.

Appendix A. Proof of Theorem 1

We start by establishing some preliminary propositions.

Proposition 2. *Under assumption* (A1), *for all* $s \in \{1, 2, \ldots, L\}$, $v_t \triangleq \operatorname{sign}[u_{t-s} \cdot \epsilon_t(\boldsymbol{\theta}^0)]$ *is an independent sequence and* v_t *takes the values* 1 *and* -1 *with probability* 0.5 *each.*

Proof. To simplify notation, we write ϵ_t^0 for $\epsilon_t(\boldsymbol{\theta}^0)$.

Note first that ϵ_t^0 is given by

$$\epsilon_t^0 = \sum_{k=L+1}^{\infty} \theta_k^0 u_{t-k} + n_t,$$

from which it follows that

(a) u_{t-s} and ϵ_t^0 are independent for all $s \in \{1, 2, \ldots, L\}$,
(b) for any $\tau > t$ and for all $s \in \{1, 2, \ldots, L\}$, $u_{\tau-s}$ is independent of the pair (v_t, ϵ_τ^0),

since u_t is an independent sequence.

We first prove that v_t takes the values ± 1 with probability 0.5 each. From the input design, it follows that $\Pr\{u_{t-s} = 0\} = 0$ and using the properties of the sign-function we have that

$$\Pr\{v_t = 1\} = \Pr\left\{u_{t-s} > 0, \epsilon_t^0 > 0\right\} + \Pr\left\{u_{t-s} < 0, \epsilon_t^0 < 0\right\} + 0.5\Pr\left\{\epsilon_t^0 = 0\right\}.$$

Utilizing fact (a) above and that $\Pr\{u_{t-s} > 0\} = \Pr\{u_{t-s} < 0\} = 0.5$, we get

$$\begin{aligned}\Pr\{v_t = 1\} &= \Pr\{u_{t-s} > 0\} \cdot \Pr\left\{\epsilon_t^0 > 0\right\} \\ &\quad + \Pr\{u_{t-s} < 0\} \cdot \Pr\left\{\epsilon_t^0 < 0\right\} + 0.5\Pr\left\{\epsilon_t^0 = 0\right\} \\ &= 0.5\left[\Pr\left\{\epsilon_t^0 > 0\right\} + \Pr\left\{\epsilon_t^0 < 0\right\} + \Pr\left\{\epsilon_t^0 = 0\right\}\right] \\ &= 0.5.\end{aligned}$$

$\Pr\{v_t = -1\} = 0.5$ is proved similarly.

Next we prove that v_t is an independent sequence. Observe that, for any two positive integers t and τ with $\tau > t$,

$$\begin{aligned}&\Pr\{v_t = 1, v_\tau = \operatorname{sign}[u_{\tau-s} \cdot \epsilon_\tau^0] = 1\} \\ &= \Pr\{v_t = 1, u_{\tau-s} > 0, \epsilon_\tau^0 > 0\} \\ &\quad + \Pr\{v_t = 1, u_{\tau-s} < 0, \epsilon_\tau^0 < 0\} \\ &\quad + 0.5\Pr\{v_t = 1, \epsilon_\tau^0 = 0\}\end{aligned}$$

$$
\begin{aligned}
&= \Pr\{u_{\tau-s} > 0 | v_t = 1, \epsilon_\tau^0 > 0\} \cdot \Pr\{v_t = 1, \epsilon_\tau^0 > 0\} \\
&\quad + \Pr\{u_{\tau-s} < 0 | v_t = 1, \epsilon_\tau^0 < 0, \} \cdot \Pr\{v_t = 1, \epsilon_\tau^0 < 0\} \\
&\quad + 0.5 \Pr\{v_t = 1, \epsilon_\tau^0 = 0\} \\
&= [\text{using fact (b)}] \\
&= \Pr\{u_{\tau-s} > 0\} \cdot \Pr\{v_t = 1, \epsilon_\tau^0 > 0\} \\
&\quad + \Pr\{u_{\tau-s} < 0\} \cdot \Pr\{v_t = 1, \epsilon_\tau^0 < 0\} \\
&\quad + 0.5 \Pr\{v_t = 1, \epsilon_\tau^0 = 0\} \\
&= 0.5 \Pr\{v_t = 1\} \\
&= 0.25,
\end{aligned}
$$

that is $\Pr\{v_t = 1, v_\tau = 1\} = 0.5 \cdot 0.5 = \Pr\{v_t = 1\} \cdot \Pr\{v_\tau = 1\}$. Treating all other combinations similarly we obtain $\Pr\{v_t = 1, v_\tau = -1\} = \Pr\{v_t = -1, v_\tau = 1\} = \Pr\{v_t = -1, v_\tau = -1\} = 0.25$. Hence, the independence of v_t and v_τ has been proved. ■

Proposition 3. *Let* $\mathbf{H}$ *be a stochastic* $M \times K$ *matrix with elements* $h_{i,t}$, $i = 0, 1, \ldots, M-1$, $t = 1, 2, \ldots, K$, *constructed according to point* (3) *in Section 2.2, and further let* $\boldsymbol{\xi} \triangleq [\xi_1, \xi_2, \ldots, \xi_K]^{\mathrm{T}}$ *be a vector independent of* $\mathbf{H}$ *of mutually independent random variables symmetrically distributed around* 0. *Given an* $\bar{\imath} \in \{0, 1, \ldots M-1\}$, *let* $\mathbf{H}_{\bar{\imath}}$ *be the* $M \times K$ *matrix whose rows are all equal to the* $\bar{\imath}$*th row of* $\mathbf{H}$. *Then,* $\mathbf{H}\boldsymbol{\xi}$ *and* $(\mathbf{H} - \mathbf{H}_{\bar{\imath}})\boldsymbol{\xi}$ *have the same* M*-dimensional distribution provided that the* $\bar{\imath}$*th element of* $(\mathbf{H} - \mathbf{H}_{\bar{\imath}})\boldsymbol{\xi}$ *(which is* 0*) is repositioned as the first entry of the vector.*

Proof. Let $\mathcal{H}$ be the set of all deterministic $\{0, 1\}$-valued $M \times K$ matrices whose first row is all zeros and where the rows are all different from each other. An inspection of point (3) of the algorithm in Section 2.2 reveals that the stochastic matrix $\mathbf{H}$ 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 $\mathbf{H} \in \mathcal{H}$, introduce the notation $|\mathbf{H} - \mathbf{H}_{\bar{\imath}}|$ to denote the matrix where each entry of $\mathbf{H} - \mathbf{H}_{\bar{\imath}}$ is substituted by its absolute value. Consider the following map:

$$\text{map} : \mathbf{H} \to |\mathbf{H} - \mathbf{H}_{\bar{\imath}}|$$

and further reposition the $\bar{\imath}$-th row as the first row.

It is easy to verify that this map transforms elements $\mathbf{H} \in \mathcal{H}$ into elements of $\mathcal{H}$ and, moreover, if $\mathbf{H}_1 \neq \mathbf{H}_2$, then $\text{map}(\mathbf{H}_1) \neq \text{map}(\mathbf{H}_2)$. That is, the map is one-to-one on $\mathcal{H}$.

Now, from the fact that the map is one-to-one and the stochastic matrix $\mathbf{H}$ constructed in point (3) takes on all possible matrices in $\mathcal{H}$ with the same probability, it turns out that $\text{map}(\mathbf{H})$ has the same probability distribution as $\mathbf{H}$.

Introduce next the new variables

$$\tilde{\xi}_t \triangleq \begin{cases} \xi_t, & \text{if } h_{\bar{\imath},t} = 0 \\ -\xi_t, & \text{if } h_{\bar{\imath},t} = 1, \end{cases}$$

and let $\tilde{\boldsymbol{\xi}}$ be the vector with elements $\tilde{\xi}_t$. We show below that (i) the vector $\tilde{\boldsymbol{\xi}}$ is independent of $\mathbf{H}$ (so that $\tilde{\boldsymbol{\xi}}$ is also independent of $\text{map}(\mathbf{H})$); and (ii) the vector $\tilde{\boldsymbol{\xi}}$ has the same distribution as $\boldsymbol{\xi}$.

To verify these two properties without too much notational clutter, suppose that $K = 2$. Fix a specific matrix $\bar{\mathbf{H}} \in \mathcal{H}$ and consider the event where $\mathbf{H} = \bar{\mathbf{H}}$. The entries of the $\bar{\imath}$th row of $\bar{\mathbf{H}}$ will take on fixed numerical values, for the sake of concreteness say $(h_{\bar{\imath},1}, h_{\bar{\imath},2}) = (0, 1)$. Then, over the event where $\mathbf{H} = \bar{\mathbf{H}}$, for given sets $\mathbf{E}_1$ and $\mathbf{E}_2$, we have:

$$
\begin{aligned}
&\Pr\{\mathbf{H} = \bar{\mathbf{H}}, \tilde{\xi}_1 \in \mathbf{E}_1, \tilde{\xi}_2 \in \mathbf{E}_2\} \\
&\quad = \Pr\{\mathbf{H} = \bar{\mathbf{H}}, \xi_1 \in \mathbf{E}_1, -\xi_2 \in \mathbf{E}_2\} \\
&\quad = [\text{since } \mathbf{H} \text{ and } \boldsymbol{\xi} \text{ are independent}] \\
&\quad = \Pr\{\mathbf{H} = \bar{\mathbf{H}}\} \cdot \Pr\{\xi_1 \in \mathbf{E}_1, -\xi_2 \in \mathbf{E}_2\} \\
&\quad = [\text{since } \xi_1 \text{ and } \xi_2 \text{ are independent}] \\
&\quad = \Pr\{\mathbf{H} = \bar{\mathbf{H}}\} \cdot \Pr\{\xi_1 \in \mathbf{E}_1\} \cdot \Pr\{-\xi_2 \in \mathbf{E}_2\} \\
&\quad = [\text{since } \xi_2 \text{ is symmetrically distributed}] \\
&\quad = \Pr\{\mathbf{H} = \bar{\mathbf{H}}\} \cdot \Pr\{\xi_1 \in \mathbf{E}_1\} \cdot \Pr\{\xi_2 \in \mathbf{E}_2\} \\
&\quad = \Pr\{\mathbf{H} = \bar{\mathbf{H}}, \xi_1 \in \mathbf{E}_1, \xi_2 \in \mathbf{E}_2\},
\end{aligned}
$$

showing that $(\tilde{\xi}_1, \tilde{\xi}_2)$ and (ξ_1, ξ_2) have the same distribution, conditionally to that $\mathbf{H} = \bar{\mathbf{H}}$. Since the same holds for any other choice of $\bar{\mathbf{H}}$, the conclusion is drawn that $(\mathbf{H}, \tilde{\xi}_1, \tilde{\xi}_2)$ has the same joint distribution as $(\mathbf{H}, \xi_1, \xi_2)$, so that $(\mathbf{H}, \tilde{\xi}_1, \tilde{\xi}_2)$ carries the same distribution properties as $(\mathbf{H}, \xi_1, \xi_2)$. Generalizing to any K, we similarly get that $(\mathbf{H}, \tilde{\boldsymbol{\xi}})$ has the same joint distribution as $(\mathbf{H}, \boldsymbol{\xi})$. Now, property (i) that $\tilde{\boldsymbol{\xi}}$ is independent of $\mathbf{H}$ follows from that $\boldsymbol{\xi}$ is independent of $\mathbf{H}$, since the joint distribution of $\mathbf{H}$ and $\tilde{\boldsymbol{\xi}}$ is the same as that of $\mathbf{H}$ and $\boldsymbol{\xi}$. Moreover, the marginals of $\tilde{\boldsymbol{\xi}}$ and $\boldsymbol{\xi}$ are obviously the same, and this is property (ii).

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

Proposition 4. *Let* $\mathbf{H}$ *and* $\boldsymbol{\xi}$ *be as in Proposition 3. Let* $\boldsymbol{\mu} \triangleq [\mu_0, \mu_1, \ldots, \mu_{M-1}]^{\mathrm{T}}$ *be a vector of mutually independent and identically distributed random variables, independent of* $\boldsymbol{\xi}$ *and* $\mathbf{H}$. *Then, the random vector* $\mathbf{H}\boldsymbol{\xi} + \boldsymbol{\mu}$ *has the following property: each element of the vector* $\mathbf{H}\boldsymbol{\xi} + \boldsymbol{\mu}$ *has the same probability* $1/M$ *to be in the* j*th position (i.e. there are exactly* $j - 1$ *other elements in* $\mathbf{H}\boldsymbol{\xi} + \boldsymbol{\mu}$ *smaller than the variable under consideration) and this holds for any choice of* j *between* 0 *to* $M - 1$.

Proof. Pick an element of the vector $\mathbf{H}\boldsymbol{\xi} + \boldsymbol{\mu}$, say $\sum_{t=1}^{K} h_{\bar{\imath},t} \cdot \xi_t + \mu_{\bar{\imath}}$. This variable is in the jth position if the inequality

$$\sum_{t=1}^{K} h_{\bar{\imath},t} \cdot \xi_t + \mu_{\bar{\imath}} > \sum_{t=1}^{K} h_{i,t} \cdot \xi_t + \mu_i \tag{A.1}$$

is satisfied for exactly $j - 1$ choices of $i \in \{0, 1, \ldots, M-1\}$. The relation (A.1) is equivalent to say that

$$\sum_{t=1}^{K} (h_{i,t} - h_{\bar{\imath},t}) \cdot \xi_t + \mu_i - \mu_{\bar{\imath}} < 0$$

holds for $j-1$ selections of $i \in \{0, 1, \ldots, M-1\}$. From Proposition 3 it follows that $\sum_{t=1}^{K} (h_{i,t} - h_{\bar{\imath},t}) \cdot \xi_t$, $i = 0, 1, \ldots, M-1$, with the $\bar{\imath}$th element repositioned as the first one, has the same joint M-dimensional distribution as $\sum_{t=1}^{K} h_{i,t} \cdot \xi_t$, $i = 0, 1, \ldots, M-1$, i.e. the joint distribution is independent of the chosen $\bar{\imath}$. Moreover, since μ_i is independent and identically distributed, the M-dimensional joint distribution of $\mu_i - \mu_{\bar{\imath}}$, $i = 0, 1, \ldots, M-1$, with the $\bar{\imath}$th element repositioned as the first element is also independent of the chosen $\bar{\imath}$. As $\boldsymbol{\mu}$ is independent of $\mathbf{H}$ and $\boldsymbol{\xi}$, it also follows that the joint M-dimensional distribution of

$$\sum_{t=1}^{K} (h_{i,t} - h_{\bar{\imath},t}) \cdot \xi_t + \mu_i - \mu_{\bar{\imath}}$$

with the $\bar{\imath}$th element repositioned as the first is independent of the chosen $\bar{\imath}$. Hence, the probability that $\bar{\imath}$ is in the jth position does not depend on the chosen $\bar{\imath}$, and since $\bar{\imath}$ can take on M values this probability is $1/M$. ■

Using these propositions, we now prove Theorem 1. Let $\mathbf{g}_s = [g_{0,s}(\boldsymbol{\theta}^0), g_{1,s}(\boldsymbol{\theta}^0), \ldots, g_{M-1,s}(\boldsymbol{\theta}^0)]^{\mathrm{T}}$ and note that $\mathbf{g}_s$ can be written as $\mathbf{H}\boldsymbol{\nu} + \boldsymbol{\nu}$ where $\mathbf{H}$ is the stochastic $M \times K$ matrix constructed in point (3) of the algorithm in Section 2.2, $\mathbf{v} = [v_{1+L}, \ldots, v_N]^{\mathrm{T}} = \left[\mathrm{sign}[u_{1+L-s} \cdot \epsilon_{1+L}(\boldsymbol{\theta}^0)], \ldots, \mathrm{sign}[u_{N-s} \cdot \epsilon_N(\boldsymbol{\theta}^0)]\right]^{\mathrm{T}}$ and $\boldsymbol{\nu} = [\nu_{0,s}, \nu_{1,s}, \ldots, \nu_{M-1,s}]^{\mathrm{T}}$. Observe that $\mathbf{H}$ is the same as in Proposition 4 while $\boldsymbol{\nu}$ satisfies the assumptions for $\boldsymbol{\xi}$ in Proposition 4 in view of the results in Proposition 2 and $\boldsymbol{\nu}$ satisfies the assumptions for μ in Proposition 4.

Consider now the event

$$\begin{aligned}\mathbf{A} = &\left\{g_{0,s}(\boldsymbol{\theta}^0) \text{ is in the 1st or 2nd or } \ldots \text{ or } q \text{ th position}\right\}\\ &\cup\{g_{0,s}(\boldsymbol{\theta}^0) \text{ is in the } M\text{th or } (M-1)\text{th or}\\ &\ldots \text{ or } (M-q+1)\text{th position}\}.\end{aligned}$$

In view of Proposition 4,

$$\Pr(\mathbf{A}) = \frac{2q}{M}. \tag{A.2}$$

Suppose that we have extracted a probabilistic outcome ω in $\mathbf{A}$. Then, either $g_{i,s}(\boldsymbol{\theta}^0) > g_{0,s}(\boldsymbol{\theta}^0)$ for at most $q-1$ selection of i or it is less than $g_{0,s}(\boldsymbol{\theta}^0)$ for at most $q-1$ selection of i, so that $\boldsymbol{\theta}^0 \notin \boldsymbol{\Theta}_s$ (recall the construction of $\boldsymbol{\Theta}_s$). Vice versa, if $\omega \notin A$, then $g_{i,s}(\boldsymbol{\theta}^0) > g_{0,s}(\boldsymbol{\theta}^0)$ for at least q selection of i and $g_{i,s}(\boldsymbol{\theta}^0) < g_{0,s}(\boldsymbol{\theta}^0)$ for at least q selection of i, yielding $\boldsymbol{\theta}^0 \in \boldsymbol{\Theta}_s$. Using (A.2), the conclusion is drawn that

$$\Pr\left\{\boldsymbol{\theta}^0 \in \boldsymbol{\Theta}_s\right\} = 1 - \frac{2q}{M}$$

and the proof is completed. ■

Appendix B. Proof of Theorem 3

Before proving the theorem, we show a uniqueness property of the true parameter $\boldsymbol{\theta} = \boldsymbol{\theta}^0$.

Proposition 5. *Let $\epsilon_t(\boldsymbol{\theta}) = y_t - \hat{y}_t(\boldsymbol{\theta})$ be the prediction error associated with the predictor (3). Under assumptions* (A1)–(A2), *$\boldsymbol{\theta} = \boldsymbol{\theta}^0$ is the unique solution to the set of equations*

$$E\{\mathrm{sign}[u_{t-s} \cdot \epsilon_t(\boldsymbol{\theta})]\} = 0, \quad s = 1, 2, \ldots, L.$$

Proof. The fact that $\boldsymbol{\theta}^0$ is a solution to

$$E\{\mathrm{sign}[u_{t-s} \cdot \epsilon_t(\boldsymbol{\theta})]\} = 0 \quad \text{for } s = 1, 2, \ldots, L$$

has already been shown in the proof of Proposition 2 under assumptions (A1). We prove here that it is the only solution.

$E\{\mathrm{sign}[u_{t-s} \cdot \epsilon_t(\boldsymbol{\theta})]\} = 0$ is satisfied if and only if

$$\Pr\{u_{t-s} \cdot \epsilon_t(\boldsymbol{\theta}) > 0\} = \Pr\{u_{t-s} \cdot \epsilon_t(\boldsymbol{\theta}) < 0\}. \tag{B.1}$$

We show by contradiction that, for each s, the condition (B.1) holds only when $\theta_s = \theta_s^0$. Fix a $\boldsymbol{\theta}$ such that $\tilde{\theta}_s \triangleq \theta_s^0 - \theta_s \neq 0$, then by dividing by $\tilde{\theta}_s$, (B.1) can be rewritten as

$$\Pr\left\{u_{t-s}^2 + u_{t-s} \cdot x_t^s > 0\right\} = \Pr\left\{u_{t-s}^2 + u_{t-s} \cdot x_t^s < 0\right\}, \tag{B.2}$$

where

$$x_t^s \triangleq \sum_{k \le L,\ k \neq s} (\tilde{\theta}_k/\tilde{\theta}_s) u_{t-k} + \sum_{k>L} (\theta_k^0/\tilde{\theta}_s) u_{t-k} + (1/\tilde{\theta}_s) n_t.$$

Now, using the given assumptions, we show that the probability on the left-hand side of (B.2) is larger than the one on the right-hand side due to the term u_{t-s}^2, arriving at a contradiction.

From the assumption (A1), x_t^s is independent of u_{t-s}. Hence,

$$\begin{aligned}&\Pr\left\{u_{t-s} \cdot x_t^s < 0\right\}\\ &= \Pr\left\{u_{t-s} < 0, x_t^s > 0\right\} + \Pr\left\{u_{t-s} > 0, x_t^s < 0\right\}\\ &= \Pr\{u_{t-s} < 0\} \Pr\left\{x_t^s > 0\right\} + \Pr\{u_{t-s} > 0\} \Pr\left\{x_t^s < 0\right\}\\ &= 0.5 \cdot \Pr\left\{x_t^s \neq 0\right\}\\ &= 0.5,\end{aligned}$$

where $\Pr\left\{x_t^s \neq 0\right\} = 1$ follows from condition (D1) and assumption (A2). Similarly, it can be shown that

$$\Pr\left\{u_{t-s} \cdot x_t^s > 0\right\} = 0.5,$$

so that

$$\Pr\left\{u_{t-s} \cdot x_t^s < 0\right\} = \Pr\left\{u_{t-s} \cdot x_t^s > 0\right\} = 0.5. \tag{B.3}$$

For any sets $\mathbf{A}$ and $\mathbf{B}$, it is true that $\Pr(\mathbf{A}) = \Pr(\mathbf{A} \cap \mathbf{B}) + \Pr(\mathbf{A} \cap \mathbf{B}^c)$. Using this relation and

$$\begin{aligned}\mathbf{B} &\triangleq \{\omega : u_{t-s}^2(\omega) + u_{t-s}(\omega) \cdot x_t^s(\omega) < 0\}\\ &\subseteq \left\{\omega : u_{t-s}(\omega) \cdot x_t^s(\omega) < 0\right\} \triangleq \mathbf{A},\end{aligned}$$

we have

$$\begin{aligned}\Pr\{u_{t-s} \cdot x_t^s < 0\} &= \Pr\left\{u_{t-s}^2 + u_{t-s} \cdot x_t^s < 0\right\}\\ &+ \Pr\left\{u_{t-s} \cdot x_t^s < 0, u_{t-s}^2 + u_{t-s} \cdot x_t^s \ge 0\right\}.\end{aligned} \tag{B.4}$$

The second term on the right-hand side of (B.4) is

$$\begin{aligned}&\Pr\left\{u_{t-s} \cdot x_t^s < 0, u_{t-s}^2 + u_{t-s} \cdot x_t^s \ge 0\right\}\\ &= \Pr\left\{u_{t-s} < 0, x_t^s > 0, u_{t-s}^2 + u_{t-s} \cdot x_t^s \ge 0\right\}\\ &\quad + \Pr\left\{u_{t-s} > 0, x_t^s < 0, u_{t-s}^2 + u_{t-s} \cdot x_t^s \ge 0\right\}\\ &= \Pr\left\{u_{t-s}^2 + u_{t-s} \cdot x_t^s \ge 0 | u_{t-s} < 0, x_t^s > 0\right\}\\ &\quad \times \Pr\left\{u_{t-s} < 0, x_t^s > 0\right\}\\ &\quad + \Pr\left\{u_{t-s}^2 + u_{t-s} \cdot x_t^s \ge 0 | u_{t-s} > 0, x_t^s < 0\right\}\\ &\quad \times \Pr\left\{u_{t-s} > 0, x_t^s < 0\right\}\\ &= \Pr\left\{-u_{t-s} \ge x_t^s | u_{t-s} < 0, x_t^s > 0\right\} \times \Pr\left\{u_{t-s} < 0, x_t^s > 0\right\}\\ &\quad + \Pr\left\{u_{t-s} \ge -x_t^s | u_{t-s} > 0, x_t^s < 0\right\} \times \Pr\left\{u_{t-s} > 0, x_t^s < 0\right\}.\end{aligned} \tag{B.5}$$

Since x_t^s is independent of u_{t-s} and can take on arbitrary small values with non-zero probability (assumption (A2)), at least one of the two terms on the right-hand side of (B.5) must be strictly positive. Therefore, from (B.5) we have $\Pr\{u_{t-s} \cdot x_t^s < 0, u_{t-s}^2 + u_{t-s} \cdot x_t^s \ge 0\} > 0$ which in turn from (B.4) implies that

$$\Pr\left\{u_{t-s} \cdot x_t^s < 0\right\} > \Pr\left\{u_{t-s}^2 + u_{t-s} \cdot x_t^s < 0\right\}.$$

Similarly, we can prove that

$$\Pr\left\{u_{t-s} \cdot x_t^s > 0\right\} < \Pr\left\{u_{t-s}^2 + u_{t-s} \cdot x_t^s > 0\right\}.$$

Hence, using (B.3),

$$\Pr\left\{u_{t-s}^2 + u_{t-s} \cdot x_t^s > 0\right\} > \Pr\left\{u_{t-s}^2 + u_{t-s} \cdot x_t^s < 0\right\},$$

which contradicts (B.2). This completes the proof of the proposition. ■

Now we prove Theorem 3.

We will prove that, with probability 1, the functions $\bar{g}_{i,s}(\boldsymbol{\theta})/N$, $i = 1, 2, \ldots, M-1$, tend to $0.5 \cdot E\{f_{t-s,s}(\boldsymbol{\theta})\}$ as N goes to infinity. Then, for $\boldsymbol{\theta} \neq \boldsymbol{\theta}^0$, it is known from Proposition 5 that $E\{f_{t-s,s}(\boldsymbol{\theta})\} = E\{\mathrm{sign}[u_{t-s} \cdot \epsilon_t(\boldsymbol{\theta})]\} \neq 0$ for some $s \in \{1, 2, \ldots, L\}$, and when $N \to \infty$ all the $\bar{g}_{i,s}(\boldsymbol{\theta})$ and $g_{i,s}(\boldsymbol{\theta})$, $i = 1, 2, \ldots, M-1$, will have the

same sign as $E\{f_{t-s,s}(\boldsymbol{\theta})\}$. Consequently, $\boldsymbol{\theta}$ will be discarded from $\hat{\boldsymbol{\Theta}}_N$ for N large enough, as stated in the theorem.

Pick any $\boldsymbol{\theta} \neq \boldsymbol{\theta}^0$ and an $s \in \{1, 2, \ldots, L\}$. The process $f_{t-s,s}(\boldsymbol{\theta}) = \text{sign}[u_{t-s} \cdot \epsilon_t(\boldsymbol{\theta})]$ inherits the properties of being strict sense stationary and strict ergodic from u_t and n_t (assumption (A3)), see e.g. Stout (1974, Theorem 3.5.8, p. 182). Thus, as $N \to \infty$, the Birkhoff–Khinchin theorem (see Theorem 1 in Section 3, Chapter 5 of Shiryaev (1995)) entails that $\sum_{t=1+L}^{N} f_{t-s,s}(\boldsymbol{\theta})/N$ converges to $E\{f_{t-s,s}(\boldsymbol{\theta})\}$ with probability one. Then,

$$\frac{\bar{g}_{i,s}(\boldsymbol{\theta})}{N} = \frac{1}{N}\sum_{t=1+L}^{N} h_{i,t-L} \cdot f_{t-L,s}(\boldsymbol{\theta}) \to 0.5 \cdot E\{f_{t-s,s}(\boldsymbol{\theta})\} \neq 0, \quad i = 1, 2, \ldots, M-1,$$

with probability one. It follows that all of the functions $g_{i,s}(\boldsymbol{\theta}) - g_{0,s}(\boldsymbol{\theta}) = \bar{g}_{i,s}(\boldsymbol{\theta}) + \nu_{i,s} - \nu_{0,s}$, $i = 1, 2, \ldots, M-1$, will have the same sign as $E\{f_{t-s,s}(\boldsymbol{\theta})\}$ for all $N > \bar{N}$ with a sufficiently large value of $\bar{N}$, and the $\boldsymbol{\theta}$ will be excluded from the confidence set $\boldsymbol{\Theta}_s$ (and hence from $\hat{\boldsymbol{\Theta}}_N$). Therefore (5) holds. ■

Appendix C. Proof of Proposition 1

From (D2), u_t is white with spectral density $\Phi_u > 0$ and hence

$$\begin{aligned} E\{\tilde{\mathbf{u}}_i\tilde{\mathbf{u}}_j^{\mathrm{T}}\} &= E\left\{\{\boldsymbol{\mathcal{B}}_{N-i}(z^{-1})\{u_t\}\}_{t=N} \cdot \{\boldsymbol{\mathcal{B}}_{N-j}^{\mathrm{T}}(z^{-1})\{u_t\}\}_{t=N}\right\} \\ &= \frac{\Phi_u}{2\pi}\int_{-\pi}^{\pi} \boldsymbol{\mathcal{B}}_{N-i}(\mathrm{e}^{-j\omega})\overline{\boldsymbol{\mathcal{B}}_{N-j}^{\mathrm{T}}(\mathrm{e}^{-j\omega})}\mathrm{d}\omega \\ &= \begin{cases} \Phi_u \cdot \mathbf{I}, & \text{for } i = j, \ (\mathbf{I}\text{: identity matrix}) \\ \mathbf{0}, & \text{for } i \neq j, \end{cases} \end{aligned} \tag{C.1}$$

where we have used Parseval's relationship and the orthogonality of the basis functions (Theorem 4), i.e.

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} \boldsymbol{\mathcal{B}}_n(\mathrm{e}^{-j\omega})\overline{\boldsymbol{\mathcal{B}}_m(\mathrm{e}^{-j\omega})}\mathrm{d}\omega = \begin{cases} \mathbf{I}, & n = m \\ \mathbf{0}, & n \neq m. \end{cases}$$

$\tilde{\mathbf{u}}_j$ is Gaussian since it is obtained by filtering u_t which is Gaussian. Moreover, uncorrelated Gaussian variables are independent, and hence the theorem follows from (C.1). ■

References

Campi, M. C., & Weyer, E. (2002). Finite sample properties of system identification methods. *IEEE Transactions on Automatic Control*, *47*(8), 1329–1334.

Campi, M. C., & Weyer, E. (2005). Guaranteed non-asymptotic confidence regions in system identification. *Automatica*, *41*(10), 1751–1764.

Campi, M. C., & Weyer, E. (2006a). Identification with finitely many data points: The LSCR approach. In *Semi-plenary presentation. Proc. symposium on system identification*.

Campi, M. C., & Weyer, E. (2006b). Non-asymptotic confidence sets for input–output transfer functions. In *Proc. 45th IEEE conference on decision and control* (pp. 157–162).

Campi, M. C., & Weyer, E. (2009). Non-asymptotic confidence sets for the parameters of linear transfer functions. *IEEE Transactions on Automatic Control* (in press).

Campi, M. C., Ooi, S. K., & Weyer, E. (2004). Non-asymptotic assessment of generalized FIR models with periodic inputs. *Automatica*, *40*(12), 2029–2041.

den Dekker, A. J., Bombois, X., & Van den Hof, P. M. J. (2008). Finite sample confidence regions for parameters in prediction error identification using output error models. In Proc. 17th IFAC world congress (pp. 5024–5029).

Douma, S. G., & Van den Hof, P. M. J. (2006). Probabilistic model uncertainty bounding: An approach with finite-time perspectives. In Proc. 14th IFAC symposium on system identification (pp. 1021–1026).

Garatti, S., Campi, M. C., & Bittanti, S. (2004). Assessing the quality of identified models through the asymptotic theory — When is the result reliable? *Automatica*, *40*(8), 1319–1332.

Garatti, S., Campi, M. C., & Bittanti, S. (2006). The asymptotic model quality assessment for instrumental variable identification revisited. *Systems & Control Letters*, *55*(6), 494–500.

Hartigan, J. A. (1969). Using subsample value as typical values. *Journal of the American Statistical Association*, *64*(328), 1303–1317.

Hartigan, J. A. (1970). Exact confidence intervals in regression problems with independent symmetric errors. *The Annals of Mathematical Statistics*, *41*(6), 1992–1998.

Heuberger, P. S. C., Van den Hof, P. M. J., & Bosgra, O. H. (1995). A generalized orthonormal basis for linear dynamical systems. *IEEE Transactions on Automatic Control*, *40*(3), 451–465.

Heuberger, P. S. C., De Hoog, T. J., Van den Hof, P. M. J., & Wahlberg, B. (2003). Orthonormal basis functions in time and frequency domain: HAMBO transform theory. *SIAM Journal on Control and Optimization*, *42*(4), 1347–1373.

Hjalmarson, H., & Ninness, B. (2006). Least-squares estimation of a class of frequency functions: A finite sample variance expression. *Automatica*, *42*(4), 589–600.

Ljung, L. (1999). *System identification: Theory for the user*. Upper Saddle River, NJ, USA: Prentice Hall.

Ninness, B., Hjalmarsson, H., & Gustafsson, F. (1999). The fundamental role of general orthonormal bases in system identification. *IEEE Transactions on Automatic Control*, *44*(7), 1384–1406.

Shiryaev, A. N. (1995). *Probability*. New York, USA: Springer-Verlag.

Söderström, T., & Stoica, P. (1989). *System identification*. Hertfordshire, UK: Prentice Hall.

Stout, W. F. (1974). *Almost sure convergence*. New York, USA: Academic Press.

Van den Hof, P. M. J., Heuberger, P. S. C., & Bokor, J. (1995). System identification with generalized orthonormal basis functions. *Automatica*, *31*(12), 1821–1834.

Wahlberg, B. (1991). System identification using Laguerre models. *IEEE Transactions on Automatic Control*, *36*(5), 551–562.

Wahlberg, B. (1994). System identification using Kautz models. *IEEE Transactions on Automatic Control*, *39*(6), 1276–1282.

Weyer, E., & Campi, M. C. (2002). Non-asymptotic confidence ellipsoids for the least-squares estimate. *Automatica*, *38*(9), 1539–1547.

Marco C. Campi is Professor of Automatic Control at the University of Brescia, Italy.

In 1988, he received the Doctor degree in electronic engineering from the Politecnico di Milano, Milano, Italy. From 1988 to 1989, he was a Research Assistant at the Department of Electrical Engineering of the Politecnico di Milano. From 1989 to 1992, he worked as a Researcher at the Centro di Teoria dei Sistemi of the National Research Council (CNR) in Milano and, in 1992, he joined the University of Brescia, Brescia, Italy. 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, USA; the Centre for Artificial Intelligence and Robotics, Bangalore, India; the University of Melbourne, Australia; the Kyoto University, Japan.

Prof. Campi is an Associate Editor of Systems and Control Letters, and a past Associate Editor of Automatica and the European Journal of Control. From 2002 to 2008, he served as the Chair of the Technical Committee IFAC on Stochastic Systems (SS) and he is currently the vice-chair for the Technical Committee IFAC on Modeling, Identification and Signal Processing (MISP). Moreover, he is a distinguished lecturer of the Control Systems Society. Marco Campi's doctoral thesis was awarded the "Giorgio Quazza" prize as the best original thesis for year 1988. In 2008, he received the IEEE CSS George S. Axelby outstanding paper award for the article "The Scenario Approach to Robust Control Design", co-authored with G. Calafiore.

The research interests of Marco Campi include: system identification, stochastic systems, adaptive and data-based control, robust convex optimization, robust control and estimation, and learning theory.

Sangho Ko is currently an assistant professor at the School of Aerospace and Mechanical Engineering, Korea Aerospace University, South Korea. He received his Ph.D. degree in Mechanical Engineering from the University of California, San Diego (UCSD) in 2005. From 2005 to 2006, he was a post-doctoral scholar in UCSD to research system identification for combustion instability problem of gas turbine engines. From March 2006 to February 2008 he was a research fellow in the Department of Electrical and Electronic Engineering of the University of Melbourne, Australia, where he worked on finite sample quality assessment of system identification. From 1992 to 1999, he was with Samsung Aerospace Industries, Ltd., Kyungnam, Korea, where he was involved in designing digital flight control system of the advanced jet trainer T-50 for the Republic of Korea Air Force.

Erik Weyer received the Siv. Ing. degree in 1988 and the Ph.D. in 1993, both from the Norwegian Institute of Technology, Trondheim, Norway. From 1994 to 1996 he was a Research Fellow at the University of Queensland, and since 1997 he has been with the Department of Electrical and Electronic Engineering, the University of Melbourne, where he is currently an Associate Professor. His research interests are in the area of system identification and control.