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

Marco C. Campi, Sangho Ko, Erik Weyer

Department of Electrical Engineering and Automation, University of Brescia, Via Branze 38, 25123 Brescia, Italy
School of Aerospace and Mechanical Engineering, Korea Aerospace University, 100, Hangdong-dong, Hwasun-gu, Deogyang-gu, Goyang, Gyeonggi-do, 412-791, South Korea
Department of Electrical and Electronic Engineering, The University of Melbourne, Parkville, VIC 3010, Australia

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

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

and the true system is given by

\[ G^0(z^{-1}) = \sum_{k=1}^{L} \theta_k^0 \mathcal{B}_k(z^{-1}), \]

where \( \mathcal{B}_k(z^{-1}) \) are the orthonormal basis functions, \( \theta_k, \ k = 1, 2, \ldots, L \), are the model parameters and \( \theta_k^0, k = 1, 2, \ldots, L \), are the true system parameters. The problem considered in this paper is how to construct a guaranteed confidence set for the true system parameters \( \theta_k^0, \theta_k^0, \ldots, \theta_k^0 \) 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 of Hjalmarsson and Gustafsson (1999), typically used orthonormal basis functions as the pulse basis functions corresponding to the FIR (Finite Impulse Response) models, the Laguerre functions (Wahlberg, 1991), 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

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

and the true system is given by

\[ G^0(z^{-1}) = \sum_{k=1}^{L} \theta_k^0 \mathcal{B}_k(z^{-1}), \]

where \( \mathcal{B}_k(z^{-1}) \) are the orthonormal basis functions, \( \theta_k, \ k = 1, 2, \ldots, L \), are the model parameters and \( \theta_k^0, k = 1, 2, \ldots, L \), are the true system parameters. The problem considered in this paper is how to construct a guaranteed confidence set for the true system parameters \( \theta_k^0, \theta_k^0, \ldots, \theta_k^0 \) 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.
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 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), Hjalmarsson 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 \( \theta_0^0 \), 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 \( \theta_0^0 \). 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

\[ \hat{y}_t(\theta) = \delta u_t. \]

The corresponding prediction error is given by

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

We then calculate

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

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 \( \text{sign}[u_t \cdot \epsilon_t(\theta_0^0)] = \text{sign}[u_t \cdot (\theta_0^0 u_{t-1} + n_t)] \) is an independent and symmetrically distributed process which takes the values \( \pm 1 \) 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

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

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 \( \tilde{g}_0(\theta) = 0 \)). That is, \( h_{i,t} \) determines if \( f_t(\theta) \) is used when we compute the \( i \)-th estimate of the confidence interval. Since it is very unlikely that all the \( \tilde{g}_i(\theta) \)'s have the same sign for the true \( \theta = \theta_0^0 \), we discard the regions in parameter space where all functions but at most one are less than the zero function \( \tilde{g}_0(\theta) \) or greater than the zero function \( \tilde{g}_0(\theta) \), 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 \( g_i(\theta) \) 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 \( \nu_t \), uniformly distributed between, say, \(-0.2 \) and \( 0.2 \) to the \( \tilde{g}_i(\theta) \) functions

\[ g_i(\theta) = \tilde{g}_i(\theta) + \nu_t, \quad i = 0, 1, \ldots, 19, \]

and by considering these \( g_i(\theta) \) in place of the original \( \tilde{g}_i(\theta) \) functions. Next we plot \( g_i(\theta), i = 0, 1, \ldots, 19 \), as functions of \( \theta \) 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 \( \theta_0^0 \) is \( \hat{\Theta}_{25} = [0.80, 1.12] \). It is a rigorous fact (stated in Theorem 1) that the
confidence interval constructed this way has probability $1 - 2 \cdot 2/20 = 0.80$ of containing the true parameter value $\theta^0$.

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 $\theta^0$, while they are biased away from zero for $\theta \neq \theta^0$. The $f_i(\theta)$’s of the preview example are examples of such functions. These functions are then summed up in many different ways, as done in the $g_i(\theta)$’s of the preview example, leading to sums that for $\theta \neq \theta^0$ are likely to be prevalently positive or prevalently negative. Thus, when constructing the confidence region one eliminates those $\theta$ for which a prevalence of positive or negative values are observed, and this leads to confidence regions that concentrate around the true $\theta^0$. 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 $\theta = \theta^0$ 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 $g_i(\theta)$’s) and by the fact that the data-based functions (the $f_i(\theta)$’s) are independent and symmetrically distributed around zero for $\theta = \theta^0$.

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 $u_t \cdot e_i(\theta^0) = u_t \cdot (\theta^0 u_{t-1} + n_t)$ is not an independent sequence since e.g. $u_t \cdot (\theta^0 u_{t-1} + n_t)$ and $u_{t-2} \cdot (\theta^0 u_{t-1} + n_t)$ share $u_t$. 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.

Notations: In this paper, matrices will be denoted by upper case boldface (e.g. $A$), vectors will be denoted by lower case boldface (e.g. $x$), and scalars will be denoted by lower case (e.g. $y$) or upper case (e.g. $Y$). For a matrix $A$, $A^t$ 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

$$y_t = G^i(z^{-1})u_t + n_t.$$  \hspace{1cm} (1)

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

$$G^i(z^{-1}) = \sum_{k=1}^{\infty} \theta_k^i z^{-k},$$ \hspace{1cm} (2)

where $\theta_k^i$, $k=1,2,\ldots,$ 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 $G^i(z^{-1})$ 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 $l$th order FIR model

$$\hat{y}_t(\theta) = G(z^{-1}, \theta) u_t = \sum_{k=1}^{l} \theta_k z^{-k} \phi_k^l \theta,$$ \hspace{1cm} (3)

where $\phi_k^l \triangleq [u_{t-1}, u_{t-2}, \ldots, u_{t-l}]^T$ and $\theta \triangleq [\theta_1, \theta_2, \ldots, \theta_l]^T$.

Objective: The goal is to design the input sequence $u_t$ and to construct a confidence set for $\theta^0 \triangleq [\theta_1^0, \theta_2^0, \ldots, \theta_l^0]^T$ 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 $\theta^0$ 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.

\(^1\) $G^i(z^{-1}) = \sum_{k=1}^{\infty} \theta_k^i z^{-k}$ is stable if $\sum_{k=1}^{\infty} |\theta_k^i| < \infty$.\n
Fig. 2. The $g_i(\theta)$ functions for the preview example together with a 80% confidence interval (thick solid line) and the true parameter ($\star$).

Fig. 3. The dynamical system.
2.2. Construction of confidence regions

We determine confidence regions for $\theta^0$ 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 $\Theta_s$ for $s = 1, 2, \ldots, L$. The final confidence region is obtained later on as the intersection of the regions $\Theta_s$.

**Procedure for the construction of confidence region $\Theta_s$:**

1. Compute the prediction errors
   \[
   \epsilon_t(\theta) = y_t - \hat{y}_t(\theta) = y_t - \phi_t^\top \theta
   \]
   (4) for $t = 1 + L, 2 + L, \ldots, K + L = N$.

2. Select an integer $s \in \{1, 2, \ldots, L\}$ and compute
   \[
   f_{t-s,1}(\theta) = \text{sign}[u_{t-s} \cdot \epsilon_t(\theta)]
   \]
   for $t = 1 + L, 2 + L, \ldots, K + L = N$, where sign is defined as
   \[
   \text{sign}(x) = \begin{cases} 
   -1, & \text{if } x < 0 \\
   +1, & \text{if } x > 0 \\
   0, & \text{with probability 0.5 if } x = 0
   \end{cases}
   \]

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 $h_{1,1}, h_{1,2}, \ldots, h_{1,K}; h_{2,1}, h_{2,2}, \ldots, h_{2,2}; \ldots; h_{M-1,1}, h_{M-1,2}, \ldots, h_{M-1,K}$. Compute
   \[
   \tilde{g}_{s,i}(\theta) = \sum_{t=1}^N h_{t-s} \cdot f_{t-s,1}(\theta), \quad i = 0, 1, \ldots, M - 1.
   \]

4. Add a small random number $v_{s,i}$ uniformly distributed on $[-\alpha, \alpha]$ with $0 < \alpha < 0.5$ to each correlation function:
   \[
   g_{s,i}(\theta) = \tilde{g}_{s,i}(\theta) + v_{s,i}, \quad i = 0, 1, \ldots, M - 1.
   \]

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

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

The final confidence set $\hat{\Theta}_N$ is obtained by intersecting the sets $\Theta_s$ 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 $\Theta_s$ as indicated in the procedure can be very hard. What instead is easy is to pick a $\theta$ value and verify through points (1)–(5) whether this $\theta$ belongs to $\Theta_s$. This suggests that $\Theta_s$ can be constructed by first gridding the $\theta$ space and then verify one by one whether or not the $\theta$’s on the grid are in $\Theta_s$. While this method is practical for problems of low dimensionality ($\theta$ has few elements, i.e. $L$ is small), it becomes computationally intensive for problems where $\theta$ 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, $\text{sign}[u_{t-s} \cdot \epsilon_t(\theta)]$ 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 $g_{s,i}(\theta)$ 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 $g_{s,i}(\theta)$ functions are constructed in point (4) of the procedure results in the following theorem.

**Theorem 1.** The set $\hat{\Theta}_N$ constructed in point (5) of the procedure has the property that

\[
\Pr \{ \theta^0 \in \hat{\Theta}_N \} = 1 - \frac{2q}{M},
\]

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

**Proof.** See Appendix A.

1. $1 - 2q/M$ is the exact probability that $\theta^0 \in \hat{\Theta}_N$, therefore $\theta^0 \not\in \Theta_s$, with probability 2q/M. As $\Theta_N$ is obtained by intersecting $\Theta_s$, $s = 1, \ldots, L$, it follows that $\theta^0 \not\in \Theta_N$ with probability at most $2qL/M$, and we have the following theorem.

**Theorem 2.**

\[
\Pr \{ \theta^0 \in \Theta_N \} \geq 1 - \frac{2Lq}{M}.
\]

It can be shown, as formally stated in Theorem 3, that the constructed region concentrates around the true parameter $\theta^0$ 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.

Assumption (A3) ensures that sample means converge to expected values, and Assumption (A2) ensures that the mismatch between $\theta$ and $\theta^0$ 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 $\theta \neq \theta^0$

\[
\Pr \left\{ 3\bar{N} \not\in \hat{\Theta}_N, \forall N > \bar{N} \right\} = 1.
\]

\footnote{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 \( \xi_t \), i.e. \( f_{t-s,t}(\theta) = \text{sign}[\xi_{t-s} \cdot \epsilon_t(\theta)] \) where \( \epsilon_t(\theta) \) is as before given by (4). Provided that \( \xi_t \) 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 \( \theta^0 \), \( \xi_t \) 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

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

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

\[
y_t = \sum_{k=1}^{\infty} \sum_{l=-k}^{k} u_{t-l-k} + n_t,
\]

where

\[
y_t = \begin{bmatrix} y_{1,t} \\ \vdots \\ y_{p,t} \end{bmatrix}, \quad u_t = \begin{bmatrix} u_{1,t} \\ \vdots \\ u_{m,t} \end{bmatrix}, \\
\xi_t = \begin{bmatrix} n_{1,t} \\ \vdots \\ n_{p,t} \end{bmatrix}, \quad \Xi_k = \begin{bmatrix} \theta_{0,11,k} & \cdots & \theta_{0,1m,k} \\ \vdots & \ddots & \vdots \\ \theta_{0,p1,k} & \cdots & \theta_{0,pm,k} \end{bmatrix}.
\]

We use a predictor corresponding to an FIR model

\[
\hat{y}_t(\Xi) = \sum_{k=1}^{L} \sum_{l=-k}^{k} \Xi_k u_{t-l-k}.
\]

The prediction error \( \epsilon_t(\Xi) = [\epsilon_{1,t}, \epsilon_{2,t}, \ldots, \epsilon_{p,t}]^T = y_t - \hat{y}_t(\Xi) \) is given by

\[
\epsilon_{j,t}(\Xi) = \sum_{k=1}^{L} \left( \theta_{0,j,k} u_{t-l-k} + \cdots + \theta_{0,m,k} u_{m,t-k} \right) + n_{j,t},
\]

for \( j = 1, 2, \ldots, p \).

Under the assumption that the input vector sequence \( u_t, \ t = 1, 2, \ldots, N \), is independent in time and each element has equal probability 0.5 of being larger or smaller than 0 and considering the functions

\[
f_{t-s,t}^j(\Xi) = \text{sign}[u_{t-s} \cdot \epsilon_t(\Xi)],
\]

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 \( \{\theta_{j,k}^0\} \), 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 \( A(z^{-1}) \) be a stable all-pass transfer function with an internally balanced realization \((A, B, C, D)\) of order \( n > 0 \). Denote

\[
B_k(z^{-1}) = B(z^{-1}) A^{k-1}(z^{-1}),
\]

where \( B(z^{-1}) = (zI - A)^{-1} B \). Then, the set of all scalar elements of the vectors \( B_k(z^{-1}) \), \( k = 1, 2, \ldots, \) 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 \( B_k \), \( k = 1, 2, \ldots, \) with

\[
\sum_{k=1}^{\infty} \|B_k\|^2 < \infty, \text{ such that}
\]

\[
H(z^{-1}) = \sum_{k=1}^{\infty} L_k B_k(z^{-1})
\]

\[
= \sum_{k=1}^{\infty} \left( L_{k,1} B_{k,1}(z^{-1}) + \cdots + L_{k,n} B_{k,n}(z^{-1}) \right).
\]

One can construct an all-pass transfer function \( A(z^{-1}) \) 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 \( A(z^{-1}) = z^{-1} \) with balanced realization \((A, B, C, D) = (0, 1, 1, 0)\), we obtain the standard pulse basis

\[
B_k(z^{-1}) = A^k(z^{-1}) = z^{-k}.
\]
Laguerre functions: Using the all-pass transfer function \( A(z^{-1}) = (1 - az)/(z - a) \) 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):

\[
B_k(z^{-1}) = \sqrt{1 - a^2} (1 - az)^{k-1} \frac{1}{(z - a)^k}.
\]

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

\[
A = \begin{bmatrix} \gamma_1 & \gamma_2 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & \sqrt{1 - b^2} \\ c \sqrt{1 - b^2} & -bc \end{bmatrix},
\]

\[
C = \begin{bmatrix} \gamma_1 & -c \end{bmatrix}, \quad D = -c
\]

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

\[
B_k(z^{-1}) = \frac{-\sqrt{1 - c^2} + \theta_k}{\sqrt{1 - b^2} + \frac{1}{2} \sum_{\nu=1}^{k-1} (y_k - \nu b)}.
\]

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

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

Next introduce a notation which clearly distinguishes between a sequence and a specific element of the sequence. We let \( \{y_i\} \) denote a sequence and \( \{y_i\}_{i=0}^n \) an element of the sequence. With this notation the above data points can be written as

\[
\{y_i\}_{i=1}^N, \{y_i\}_{i=2}^N, \ldots, \{y_i\}_{i=N-1}, \{y_i\}_{i=N}
\]

or

\[
\{y_{i-N}\}_{i=1}^N, \{y_{i-N+2}\}_{i=1}^N, \ldots, \{y_{i-1}\}_{i=1}^N, \{y_i\}_{i=N},
\]

where \( \{y_{i-k}\} \) is the sequence \( \{y_i\} \) shifted by \( k \), i.e. \( y_{i-1} = y^{-1} \).

(7) can also be written as

\[
\{z^{-N+1} y_{i}\}_{i=1}^N, \{z^{-N+2} y_{i}\}_{i=1}^N, \ldots, \{z^{-1} y_{i}\}_{i=1}^N, \{y_i\}_{i=N}.
\]

(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 \( A(z^{-1}) = z^{-1} \). More generally an arbitrary all-pass filter \( A(z^{-1}) \) can be used leading to the data

\[
\{A^{-N-1}(z^{-1}) y_{i}\}_{i=1}^N, \{A^{-N-2}(z^{-1}) y_{i}\}_{i=1}^N, \ldots, \{A(z^{-1}) y_{i}\}_{i=1}^N, \{y_i\}_{i=N}.
\]

(9)

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

\[
G^0(z^{-1}) = \sum_{k=1}^{\infty} \theta_k B_k(z^{-1}) = \sum_{k=1}^{\infty} \theta_k B(z^{-1}) A^{k-1}(z^{-1}).
\]

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

\[
\tilde{y}_N := \begin{cases} 0 \{B(z^{-1}) y_i\}_{i=1}^N \end{cases} + \theta_1^0 B(z^{-1}) A(z^{-1}) y_i \begin{cases} 0 \{y_i\}_{i=1}^N \end{cases} + \cdots + \{n_i\}_{i=1}^N
\]

\[
= \theta_1^0 \tilde{y}_{N-1} + \theta_2^0 \tilde{y}_{N-2} + \cdots + \tilde{n}_N,
\]

where

\[
\tilde{u}_j = \{B(z^{-1}) A^{N-(j+1)}(z^{-1}) y_i\}_{i=1}^N.
\]

Similarly

\[
\tilde{y}_{N-1} := \begin{cases} 0 \{B(z^{-1}) y_i\}_{i=1}^N \end{cases} + \theta_1^0 B(z^{-1}) A(z^{-1}) y_i \begin{cases} 0 \{y_i\}_{i=1}^N \end{cases} + \cdots + \{A(z^{-1}) y_i\}_{i=1}^N
\]

\[
= \theta_1^0 \tilde{u}_{N-2} + \theta_2^0 \tilde{u}_{N-3} + \cdots + \tilde{n}_{N-1}
\]

and, in general,

\[
\tilde{y}_j := \{A^{N-j}(z^{-1}) y_i\}_{i=1}^N = \theta_1^0 \tilde{u}_{j-1} + \theta_2^0 \tilde{u}_{j-2} + \cdots + \tilde{n}_j
\]

\[
= \sum_{k=1}^{\infty} \theta_k \tilde{u}_{j-k} + \tilde{n}_j
\]

(10)

where

\[
\tilde{n}_j := \{A^{N-j}(z^{-1}) y_i\}_{i=1}^N.
\]

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 \( \tilde{u}_j, \tilde{y}_j, j = 1, \ldots, N \), are obtained through the successive filtering with the all-pass filter \( A(z^{-1}) \).

A reduced-order predictor is now given by

\[
\hat{y}_j(\theta) = \sum_{k=1}^{\infty} \theta_k \tilde{u}_{j-k}
\]

\[
= \sum_{k=1}^{\infty} (\theta_{1,k} \tilde{u}_{j-k} + \theta_{2,k} \tilde{u}_{2j-k} + \cdots + \theta_{n,k} \tilde{u}_{nj-k}),
\]

where \( \theta_{1,k} \) and \( \tilde{u}_{j-k} \) denote the elements of the \( \theta_j \) and \( \tilde{u}_j \) vectors.

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

(D2) The input \( u_i \) is a zero-mean white Gaussian signal with spectral density \( \phi_{\theta} > 0 \).

The main difference between this input design (D2) and (D1) is that in (D2) \( u_i \) 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 \( \tilde{u}_j, j = 1, \ldots, N \), are independent and identically distributed zero-mean Gaussian, and, moreover, the elements of \( \tilde{u}_j \) 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 \( \tilde{u}_j \). 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 \( \tilde{u}_j \) and noise \( \tilde{n}_j \) satisfy the assumptions on \( u_i \) and \( n_i \) 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. \( \tilde{u}_j \) is a vector in the general case, so all the functions in the algorithm, and specifically \( g_k(\theta) \), become vector functions. The confidence set \( \Theta \) is now obtained by taking the intersection of the confidence sets obtained by considering the \( n \) scalar elements of \( g_k(\theta) \) separately.

The following corollary summarizes the properties of the confidence regions.
Fig. 4. Generation of the filtered data set $\{\tilde{u}_t, \tilde{y}_t\}_{t=1,2,...,N}$ from the original data set $\{u_t, y_t\}_{t=1,2,...,N}$.

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

$$\Pr\{\theta^0 \in \Theta_s\} \geq 1 - 2qn/M.$$  

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

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

Furthermore, for every fixed $\theta \neq \theta^0$, 

$$\Pr\{\exists \bar{N} | \theta \not\in \hat{\Theta}_N, \forall N > \bar{N}\} = 1.$$  

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 $\tilde{u}_t$ are arbitrary initialized, then $\tilde{u}_t$ 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.

**4.1. FIR model**

Consider the linear time-invariant discrete-time system

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

where the true transfer function

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

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

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

Fig. 5. First 20 coefficients of pulse basis functions.
Fig. 6. Non-asymptotic 95% confidence region for \((\theta_1^0, \theta_2^0)\) (blank region) using 50 data points. ⋆ = true parameter.

Fig. 7. Non-asymptotic 95% confidence region for \((\theta_1^0, \theta_2^0)\) (blank region) using 100 data points. ⋆ = true parameter.

Fig. 8. Non-asymptotic 95% confidence region for \((\theta_1^0, \theta_2^0)\) (blank region) using 500 data points. ⋆ = true parameter.

Fig. 9. Non-asymptotic 95% confidence region for \((\theta_1^0, \theta_2^0)\) (blank region) using 2000 data points. ⋆ = true parameter.

To construct confidence regions for \(\theta_1^0\) and \(\theta_2^0\), we considered \(N = 50, 100, 500\) and 2000 data points generated according to (11) and (12), and computed

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

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

Then, we computed the following \(M = 400\) empirical correlation functions

\[
g_{t,1}(\theta) = \sum_{i=3}^{N} h_{t-i-2} \cdot f_{t-1,1}(\theta) + v_{1,1}, \quad i = 0, 1, \ldots, M - 1,
\]

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

Here \(v_{1,1}\) and \(v_{1,2}\) were uniformly distributed on \([-0.1, 0.1]\). We excluded the regions in parameter space where at most 4 (out of the \(M = 400\)) \(g_{t,1}(\theta)\) (or \(g_{t,2}(\theta)\)) functions were smaller or greater than the \(g_{0,1}(\theta)\) (or \(g_{0,2}(\theta)\)) 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 \(g_{t,1}(\theta)\) functions were smaller than \(g_{0,1}(\theta)\) are marked with ○, and the region where at most 4 were greater than \(g_{0,1}(\theta)\) are marked with □. Likewise for \(g_{t,2}(\theta)\), where + and x represent the regions where at most 4 values of \(g_{t,2}(\theta)\) were smaller or greater than \(g_{0,2}(\theta)\). 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.

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

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

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

\[
G_c(s) = \frac{1}{(2s + 1)(s + 1)}
\]

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

\[
\theta_k^0(\text{pulse}) = \sum_{i=1}^{2} n_i \rho_{i,k}^{k-1}, \quad \text{for } k > 0
\]

\[
\theta_k^0(\text{Laguerre}) = \sum_{i=1}^{2} n_i \sqrt{(1 - a^2)} \left[ p_1 - a \cdot p_i \right]^{k-1}, \quad \text{for } k > 0,
\]

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.
We chose a second-order Laguerre model for the system. Its predictor is given by
\[ \hat{y}_t(\theta) = \theta_1 \mathcal{B}_1(z^{-1}) u_t + \theta_2 \mathcal{B}_2(z^{-1}) u_t, \]
where
\[ \mathcal{B}_k(z^{-1}) = B(z^{-1}) A^{k-1}(z^{-1}) \]
with
\[ B(z^{-1}) = \frac{1 - a^2}{z - a}, \quad A(z^{-1}) = \frac{1 - az}{z - a}, \quad a = 0.5. \]

To obtain the confidence region for \( \theta_0^1 \) and \( \theta_0^2 \), we applied the successive filtering explained in Section 3.2 to the original data. 

We used the last 1000 filtered data points and computed, for \( j = 3003, 3004, \ldots, 4000 \),
\[ f_{j-1,1}(\theta) = \text{sign} \left[ \hat{u}_{j-1} \cdot e_j(\theta) \right], \]
\[ f_{j-2,2}(\theta) = \text{sign} \left[ \hat{u}_{j-2} \cdot e_j(\theta) \right], \]
and obtained \( M = 960 \) empirical correlation functions
\[ g_{i1}(\theta) = \sum_{j=3003}^{4000} h_{j-3002} \cdot f_{j-1,1}(\theta) + v_i, \quad i = 0, 1, \ldots, 959, \]
\[ g_{i2}(\theta) = \sum_{j=3003}^{4000} h_{j-3002} \cdot f_{j-2,2}(\theta) + v_i, \quad i = 0, 1, \ldots, 959. \]

Here \( v_{i1} \) and \( v_{i2} \) were uniformly distributed on \([-0.1, 0.1]\). We excluded the regions in parameter space where at most 11 (out of \( M = 960 \)) \( g_{i1}(\theta) \) (or \( g_{i2}(\theta) \)) functions were smaller or greater than the \( g_{01}(\theta) \) (or \( g_{02}(\theta) \)) 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 \( g_{i1}(\theta) \) functions were smaller than \( g_{01}(\theta) \) is marked with \( \bigcirc \), and the region where at most 11 were greater than \( g_{01}(\theta) \) is marked with \( \Box \). Likewise for \( g_{i2}(\theta) \), where + and \( \times \) represent the regions where at most 11 values of \( g_{i2}(\theta) \) were smaller or greater than \( g_{02}(\theta) \).

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.

Appendix A. Proof of Theorem 1

We start by establishing some preliminary propositions.

**Proposition 2.** Under assumption (A1), for all \( s \in \{1, 2, \ldots, L\} \),
\[ v_t = \text{sign}(u_{t-s} \cdot e_s(\theta^0)) = \text{sign}(u_{t-s} \cdot e_s(\theta^0)) \]
and \( v_t \) takes the values 1 and \(-1\) with probability 0.5 each.

**Proof.** To simplify notation, we write \( e_t^0 \) for \( e_t(\theta^0) \).

Note first that \( e_t^0 \) is given by
\[ e_t^0 = \sum_{k=1}^{\infty} \theta_k^0 u_{t-k} + n_t, \]

from which it follows that
(a) \( u_{t-s} \) and \( e_t^0 \) are independent for all \( s \in \{1, 2, \ldots, L\} \),
(b) for any \( \tau > t \) and for all \( s \in \{1, 2, \ldots, L\} \), \( u_{t-s} \) is independent of the pair \( (u_t, e_t^0) \),
since \( u_t \) is an independent sequence.

We first prove that \( v_t \) takes the values \( \pm 1 \) 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
\[ \Pr\{v_t = 1\} = \Pr\{u_{t-s} > 0, e_t^0 > 0\} + \Pr\{u_{t-s} < 0, e_t^0 < 0\} + 0.5\Pr\{e_t^0 = 0\}. \]

Utilizing fact (a) above and that \( \Pr\{u_{t-s} > 0\} = \Pr\{u_{t-s} < 0\} = 0.5 \), we get
\[ \Pr\{v_t = 1\} = 0.5 \Pr\{u_{t-s} > 0\} \cdot \Pr\{e_t^0 > 0\} + \Pr\{u_{t-s} < 0\} \cdot \Pr\{e_t^0 < 0\} + 0.5\Pr\{e_t^0 = 0\} \]
\[ = 0.5 \Pr\{e_t^0 > 0\} + \Pr\{e_t^0 < 0\} + 0.5\Pr\{e_t^0 = 0\} \]
\[ = 0.5. \]

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

Next we prove that \( v_t \) is an independent sequence. Observe that, for any two positive integers \( t \) and \( \tau \) with \( \tau > t \),
\[ \Pr\{v_t = 1, v_\tau = \text{sign}(u_{t-\tau} \cdot e_\tau^0) = 1\} = \Pr\{v_t = 1, u_{t-\tau} > 0, e_\tau^0 > 0\} + \Pr\{v_\tau = 1, u_{t-\tau} < 0, e_\tau^0 < 0\} + 0.5\Pr\{v_t = 1, e_\tau^0 = 0\} \]
\[ = 0.5. \]
Let \( h_{i,t} \in \{0,1\} , \quad i = 0, \ldots, M - 1 \), \( t = 1, 2, \ldots, K \), constructed according to point (3) in Section 2.2, and further let \( \bar{\xi} = [\bar{\xi}_1, \bar{\xi}_2, \ldots, \bar{\xi}_M] \) be a vector independent of \( H \) of mutually independent random variables symmetrically distributed around 0. Given an \( i \in \{0,1, \ldots, M - 1\} \), let \( \tilde{H} \) be the \( M \times K \) matrix whose rows are all equal to the \( i \)-th row of \( H \). Then, \( \tilde{H} \) and \((H - \tilde{H})\xi\) have the same \( M \)-dimensional distribution provided that the \( i \)-th element of \((H - \tilde{H})\xi\) (which is 0) is repositioned as the first entry of the vector.

**Proposition 3.** Let \( H \) be a stochastic \( M \times K \) matrix with elements \( h_{i,t} \). Let \( \bar{\xi}_1 = 0,\ldots, \bar{\xi}_M \), the \( i \)-th row of \( H \) be repeated as an entry of \( \bar{H} \). The resulting matrix \( \bar{H} \) is independent of \( H \), so that the vector \((\bar{H} - H)\xi\) as the first entry is the same as the map \(H \xi\). Since \( \bar{H} \) is distributed as \( H \), \( \xi \) and \( (\bar{H} - H)\xi \) are independent, map \(H \xi\) has the same distribution as \( H \xi \) and the proposition is established.

**Proposition 4.** Let \( H \) and \( \xi \) be as in Proposition 3. Let \( \mu = [\mu_0, \mu_1, \ldots, \mu_{M-1}]^T \) be a vector of mutually independent and identically distributed random variables, independent of \( \xi \) and \( H \). Then, the random vector \( H\xi + \mu \) has the following property: each element of the vector \( H\xi + \mu \) has the same probability 1/\( M \) to be in the \( j \)-th position (i.e. there are exactly \( j \)-1 other elements in \( H\xi + \mu \) 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 \( H\xi + \mu \), say \( \sum_{i=1}^K h_{i,t} \cdot \bar{\xi}_i + \mu_i \). This variable is in the \( j \)-th position if the inequality

\[
\sum_{i=1}^K h_{i,t} \cdot \bar{\xi}_i + \mu_i > \sum_{i=1}^K h_{i,t} \cdot \bar{\xi}_i + \mu_i 
\]

(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

\[
\sum_{i=1}^K (h_{i,t} - \bar{h}_{i,t}) \cdot \bar{\xi}_i + \mu_i - \mu_i < 0
\]

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

\[
\sum_{i=1}^K (h_{i,t} - \bar{h}_{i,t}) \cdot \bar{\xi}_i + \mu_i - \mu_i
\]

with the \( i \)-th element repositioned as the first is independent of the chosen \( i \). Hence, the probability that \( i \) is in the \( j \)-th position does not depend on the chosen \( i \), and since \( i \) can take on \( M \) values this probability is \( 1/M \).
Using these propositions, we now prove Theorem 1. Let \( g_i = [g_{0,i}(\theta^0), g_{1,i}(\theta^0), \ldots, g_{M-1,i}(\theta^0)]^T \) and note that \( g_i \) can be written as \( Hv + v \) where \( H \) is the stochastic \( M \times K \) matrix constructed in point (3) of the algorithm in Section 2.2, \( v \) is the true parameter \( E \). Appendix B. Proof of Proposition 4 while \( v \) satisfies the assumptions for \( \theta \) in Proposition 4 in view of the results in Proposition 2 and \( v \) satisfies the assumptions for \( \mu \) in Proposition 4.

Consider now the event

\[
A = \{g_{0,i}(\theta^0) \text{ is in the 1st or 2nd or \ldots or q th position}\} \\
\cup \{g_{0,i}(\theta^0) \text{ is in the Mth or } (M - 1)\text{th or \ldots or } (M - q + 1)\text{th position}\}.
\]

In view of Proposition 4, \( \Pr(A) = \frac{2q}{M} \).

Suppose that we have extracted a probabilistic outcome \( \omega \) in \( A \). Then, either \( g_{0,i}(\theta^0) > g_{0,i}(\theta^0) \) for at most \( q - 1 \) selection of \( i \) or it is less than \( g_{0,i}(\theta^0) \) for at most \( q - 1 \) selection of \( i \), so that \( \theta^0 \not\in \Theta_i \) (recall the construction of \( \Theta \)). Vice versa, if \( \omega \not\in A \), then \( g_{0,i}(\theta^0) > g_{0,i}(\theta^0) \) for at least \( q \) selection of \( i \) and \( g_{0,i}(\theta^0) < g_{0,i}(\theta^0) \) for at least \( q \) selection of \( i \), yielding \( \theta^0 \not\in \Theta_i \). Using (A.2), the conclusion is drawn that

\[
\Pr(\theta^0 \in \Theta \} = 1 - \frac{2q}{M}
\]

and the proof is completed. 

**Appendix B. Proof of Theorem 3**

Before proving the theorem, we show a uniqueness property of the true parameter \( \theta = \theta^0 \).

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

\[
E[\text{sign}(u_{i,s} \cdot e_i(\theta))] = 0, \quad s = 1, 2, \ldots, L.
\]

**Proof.** The fact that \( \theta^0 \) is a solution to

\[
E[\text{sign}(u_{i,s} \cdot e_i(\theta))] = 0 \quad \text{for } s = 1, 2, \ldots, L
\]

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

\[
E[\text{sign}(u_{i,s} \cdot e_i(\theta))] = 0 \quad \text{is satisfied if and only if}
\]

\[
\Pr(u_{i,s} \cdot e_i(\theta) > 0) = \Pr(u_{i,s} \cdot e_i(\theta) < 0).
\]

We show by contradiction that, for each \( s \), the condition (B.1) holds only when \( \theta_0 = \theta^0 \). Fix a \( \theta \) such that \( \theta_0 = \theta^0 \not\equiv \theta_0 = \theta \not\equiv \theta^0 \). Then, by dividing by \( \theta_0 \), (B.1) can be rewritten as

\[
\Pr(u_{i,s} \cdot e_i(\theta) > 0) = \Pr(u_{i,s} \cdot e_i(\theta) < 0),
\]

where

\[
X_i^t \equiv \sum_{k \geq 1} \frac{1}{k} (\theta_0 - \theta) u_{i-k} + \left(1/\theta_0 \right) \eta_i.
\]

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 \( u_{i-s}^t \) arriving at a contradiction.

From the assumption (A1), \( X_i^t \) is independent of \( u_{i,s} \). Hence,

\[
\Pr(u_{i,s} \cdot e_i(\theta) < 0) = \Pr(u_{i,s} < 0, X_i^t < 0) \quad \text{for } i = 1, 2, \ldots, M - 1.
\]

Similarly, we can prove that

\[
\Pr(u_{i,s} \cdot e_i(\theta) > 0) \quad \text{for } i = 1, 2, \ldots, M - 1.
\]

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 \( \hat{g}_i(\theta) / N, i = 1, 2, \ldots, M - 1 \), tend to 0.5 · \( E[f(u_{i-s}(\theta))] \) as \( N \) goes to infinity. Then, for \( \theta \not= \theta^0 \), it is known from Proposition 5 that \( E[f(u_{i-s}(\theta))] = E[\text{sign}(u_{i,s} \cdot e_i(\theta))] \not= 0 \) for some \( s \in \{1, 2, \ldots, L\} \), and when \( N \to \infty \) all the \( \hat{g}_i(\theta) \) and \( g_i(\theta), i = 1, 2, \ldots, M - 1 \), will have the
same sign as \(E[f_{\cdot;\cdot;\cdot;\cdot}(\theta)]\). Consequently, \(\theta\) will be discarded from \(\Theta_N\) for \(N\) large enough, as stated in the theorem.

Pick any \(\theta \neq \theta^0\) and \(N \in \{1, 2, \ldots, L\}\). The process \(f_{\cdot;\cdot;\cdot;\cdot}(\theta) = \sum_{i=1}^{N} u_i e_i(\theta)\) inherits the properties of being strictly stationary and strictly ergodic from \(u_i\) and \(e_i(\theta)\) (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 \(\sum_{i=1}^{N} f_{\cdot;\cdot;\cdot;\cdot}(\theta)/N\) converges to \(E[f_{\cdot;\cdot;\cdot;\cdot}(\theta)]\) with probability one. Then,

\[
\frac{\tilde{g}_{\cdot;\cdot;\cdot;\cdot}(\theta)}{N} = \frac{1}{N} \sum_{i=1}^{N} h_{i;\cdot;\cdot;\cdot;\cdot}(\theta)
\Rightarrow 0.5 \cdot E[f_{\cdot;\cdot;\cdot;\cdot}(\theta)] \neq 0, \quad i = 1, 2, \ldots, M - 1,
\]

with probability one. It follows that all of the functions \(g_{\cdot;\cdot;\cdot;\cdot;\cdot}(\theta) - \tilde{g}_0(\theta) = \tilde{g}_1(\theta) + \tilde{g}_2(\theta) - \tilde{g}_0(\theta)\), \(i = 1, 2, \ldots, M - 1\), will have the same sign as \(E[f_{\cdot;\cdot;\cdot;\cdot}(\theta)]\) for all \(N > N_0\) with a sufficiently large value of \(N\), and the \(\theta\) will be excluded from the confidence set \(\Theta_N\) (and hence from \(\Theta_{N_0}\)). Therefore (5) holds.

**Appendix C. Proof of Proposition 1**

From (D2), \(u_i\) is white with spectral density \(\Phi_0 > 0\) and hence

\[
E[\langle u_i u_j^T \rangle] = E \left[ \left\{ \mathbf{B}_{N-1}(z^{-1})[u_i] \right\}_{i=N} \cdot \left\{ \mathbf{B}_N^T(z^{-1})[u_i] \right\}_{i=N} \right] = \frac{\Phi_0}{2\pi} \int_{-\pi}^{\pi} \mathbf{B}_{N-1}(e^{-j\omega})^T \mathbf{B}_N(e^{j\omega}) d\omega = \begin{cases} \Phi_0 \cdot I, & \text{for } j = i, (1 \text{ identity matrix}) \\ 0, & \text{for } i \neq j \end{cases}
\]

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

\[
\frac{1}{2\pi} \int_{-\pi}^{\pi} \mathbf{B}_n \mathbf{B}_m^T d\omega = \begin{cases} 1, & n = m \\ 0, & n \neq m \end{cases}
\]

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

**References**


