Available online at www.sciencedirect.com

ScienceDirect

automatica

Automatica 43 (2007) 1418–1425

www.elsevier.com/locate/automatica

Brief paper

Parameter identification for nonlinear systems: Guaranteed confidence regions through LSCR☆

Marco Dalai[a], Erik Weyer[b], Marco C. Campi[a,*]

[a]Department of Electrical Engineering and Automation, University of Brescia, Via Branze 38, 25123 Brescia, Italy
[b]Department of Electrical and Electronic Engineering, The University of Melbourne, Parkville VIC 3010, Australia

Received 6 June 2006; received in revised form 27 October 2006; accepted 19 January 2007
Available online 19 June 2007

Abstract

In this paper we consider the problem of constructing confidence regions for the parameters of nonlinear dynamical systems. The proposed method uses higher order statistics and extends the LSCR (leave-out sign-dominant correlation regions) algorithm for linear systems introduced in Campi and Weyer [2005, Guaranteed non-asymptotic confidence regions in system identification. *Automatica* 41(10), 1751–1764. Extended version available at ⟨http://www.ing.unibs.it/∼campi⟩]. The confidence regions contain the true parameter value with a guaranteed probability for any finite number of data points. Moreover, the confidence regions shrink around the true parameter value as the number of data points increases. The usefulness of the proposed approach is illustrated on some simple examples.

Keywords: Confidence sets; Finite sample results; Nonlinear system identification

1. Introduction

It is well known that a model of a dynamical system is of limited use if no quality tag which describes the accuracy of the model is attached. Confidence regions for the system parameters are commonly used as quality tags, and asymptotic theory is widely used for the construction of such regions. However, in practice one always has a finite number of samples, and—even though the asymptotic theory delivers sensible results in many cases—there are also examples (Garatti, Campi, & Bittanti, 2004) where it fails when applied to a finite number of data points. Thus, there is a need for techniques which deliver confidence regions with guaranteed probabilities when only a finite number of data points are available.

In Campi and Weyer (2005) a method called LSCR (leave-out sign-dominant correlation regions) was proposed for finding confidence regions to which the parameters of a *linear system* belong with guaranteed probability. See also Campi and Weyer (2006) for a comprehensive presentation of LSCR. LSCR extends earlier work by Hartigan (1969, 1970) to a dynamical system setting, and it has two important features: first, the probability that the confidence region contains the true parameters is guaranteed for any finite amount of data samples; second, the confidence region concentrates around the true parameter value when the number of samples increases. In Campi and Weyer (2005), second order statistics were explored for the construction of the confidence regions. In the present paper, we consider *nonlinear systems*. It is well known (see for example, Ljung, 2001 for a general discussion, or Subba Rao, 1981 for the particular case of bilinear systems) that second order statistics are insufficient for the identification of nonlinear systems. Here we show that it is possible to extend the framework of LSCR to higher order statistics, and hence to consider the problem of nonlinear system identification within this setting.

The focus of this paper is on time series, that is the system to be identified has no exogenous inputs which are measured. The outline of the paper is as follows. In the next section, we

☆ This paper was not presented at any IFAC meeting. This paper was recommended for publication in revised form by Associate Editor Antonio Vicino under the direction of Editor Torsten Söderström.

* Corresponding author. Tel.: +39 030 3715458; fax: +39 030 380014.
E-mail addresses: marco.dalai@ing.unibs.it (M. Dalai), e.weyer@ee.unimelb.edu.au (E. Weyer), marco.campi@ing.unibs.it (M.C. Campi).

0005-1098/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.automatica.2007.01.016

motivate the use of higher order statistics for nonlinear systems. Section 3 contains the procedure for the construction of the confidence region, and the properties of this procedure are also studied. In Section 4 a simulation example using a bilinear system is presented before conclusions are given in Section 5.

2. A simple nonlinear example: from second to higher order statistics

This section illustrates the problems encountered when the standard LSCR procedure of Campi and Weyer (2005) using second order statistics is applied to a nonlinear system.

Consider the system

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

where θ^0 is the parameter value to be identified and w_t is an independent sequence of Gaussian variables with zero mean and unit variance. We use the standard LSCR algorithm for construction of a confidence region for θ^0. To this end, we first rewrite the system with a generic parameter θ, $y_t=\theta(y_{t-1}^2-1)+w_t$, and then compute the associated optimal predictor: $\hat{y}_t(\theta)=\theta(y_{t-1}^2-1)$, and the prediction error: $\varepsilon_t(\theta)=y_t-\hat{y}_t(\theta)$. LSCR constructs a confidence region based on an empirical evaluation of the correlations $E[\varepsilon_t(\theta)\varepsilon_{t+r}(\theta)]$, $r\geqslant 1$. In Campi and Weyer (2005) it is shown that θ^0 is the only value of θ for which these correlations are zero in the case of linear ARMA systems and, consequently, the obtained confidence region shrinks around the true parameter value $\theta=\theta^0$ as the number of data points grows. Here we show that $E[\varepsilon_t(\theta)\varepsilon_{t+r}(\theta)]=0$ does not imply $\theta=\theta^0$ for the system in (1), i.e. second order statistics do not suffice.

Suppose that the true parameter value is $\theta^0=0$. Then $y_t=w_t$, and we have

$$\varepsilon_t(\theta) = y_t - \hat{y}_t(\theta) = w_t - \theta(w_{t-1}^2 - 1).$$

Thus,

$$\begin{aligned} &E[\varepsilon_t(\theta)\varepsilon_{t+r}(\theta)] \\ &\quad = E[(w_t - \theta(w_{t-1}^2 - 1))(w_{t+r} - \theta(w_{t+r-1}^2 - 1))]. \end{aligned} \tag{2}$$

For $r\geqslant 2$, $E[\varepsilon_t(\theta)\varepsilon_{t+r}(\theta)]=0$ for any value of θ since w_t and (w_{t-1}^2-1) are zero mean random variables, and the products in (2) only contain terms with different time indeces. For $r=1$ we have: $E[\varepsilon_t(\theta)\varepsilon_{t+1}(\theta)]=-\theta E[w_t(w_t^2-1)]=-\theta(E[w_t^3]-E[w_t])=0$. So, $E[\varepsilon_t(\theta)\varepsilon_{t+r}(\theta)]=0$ for any $r\geqslant 1$, and any value of θ. This implies that it is not possible to establish the true value of θ from the conditions $E[\varepsilon_t(\theta)\varepsilon_{t+r}(\theta)]=0$. In turn, following the analysis in Campi and Weyer (2005), we see that the confidence region obtained by using the standard LSCR algorithm does not shrink around θ^0 when the number of samples increases.

We complete this example by showing that the true value θ^0 can indeed be determined by using higher order statistics. Take for example the condition $E[\varepsilon_t^2(\theta)\varepsilon_{t+1}(\theta)]=0$. We have

$$E[\varepsilon_t^2(\theta)\varepsilon_{t+1}(\theta)] = \theta(E[w_t^2] - E[w_t^4]) = \theta(1-3) = -2\theta. \tag{3}$$

Thus, $E[\varepsilon_t^2(\theta^0)\varepsilon_{t+1}(\theta^0)]=0$ since $\theta^0=0$, while $E[\varepsilon_t^2(\theta)\varepsilon_{t+1}(\theta)]\neq 0$ for any $\theta\neq\theta^0$.

So, in order to construct confidence regions that shrink around θ^0 higher order statistics must be utilized. In the next section we generalize the LSCR method to this case.

3. Extension of LSCR to higher order statistics

Consider a nonlinear system S^0 which maps a non-measured noise process w_t into a measured signal y_t. Furthermore, assume that S^0 belongs to a parameterized system class $\{S_\theta\}$, that is $S^0=S_{\theta^0}$ for some θ^0. w_t is an independent sequence of random variables, whose distribution is symmetric around zero. Apart from this, we make no other assumptions on w_t. The distribution of w_t can as well be time-varying. We aim at finding a confidence region for the parameter vector θ^0 by observing the output y_t.

The LSCR method in Campi and Weyer (2005) constructs, for every value of θ, a sequence $w_t(\theta)$ such that for the true parameter θ^0 we have that $w_t(\theta^0)=w_t$. Then, roughly speaking, the confidence region for θ^0 is obtained by choosing the values of θ for which $w_t(\theta)$ resembles an independent process. For linear systems, one can take $w_t(\theta)=\varepsilon_t(\theta)$, the prediction error, since $\varepsilon_t(\theta^0)=w_t$, see Campi and Weyer (2005).

The case of nonlinear systems requires some extra care because $\varepsilon_t(\theta^0)\neq w_t$ and $\varepsilon_t(\theta^0)$ is not even an independent process in general. To see this, consider, e.g. the system class $y_t=\theta y_{t-1}+y_{t-1}w_t$. The optimal predictor is $\hat{y}_t(\theta)=\theta y_{t-1}$; but $y_t-\hat{y}_t(\theta^0)=y_{t-1}w_t$ is not an independent sequence!

In order to obtain a sequence $w_t(\theta)$ such that $w_t(\theta^0)=w_t$, we can proceed in a different way by resorting to system inversion instead of constructing the prediction error, see Fig. 1. For linear systems these two approaches coincide since constructing the prediction error is the same as inverting the system. In the example above we let $w_t(\theta)=(y_t-\theta y_{t-1})/y_{t-1}$, so that $w_t(\theta^0)=w_t$ as long as $y_{t-1}\neq 0$. System inversion is used as a basic building block in the algorithm presented below.

Before proceeding we formally introduce our working assumptions.

Assumptions.

(i) The observed data y_t are obtained as output of a causal system S^0 whose input is an independent noise sequence w_t symmetrically distributed around zero, i.e. $y_t=S^0(w_\tau,\tau\leqslant t)$.

(ii) The system S^0 belongs to a system model class S_θ, i.e. there exists a value θ^0 of the parameter such that $S_{\theta^0}=S^0$.

(iii) The systems in $\{S_\theta\}$ are invertible with a causal inverse, i.e. for every θ there exists an inverse system S_θ^{-1} such that $S_\theta^{-1}(y_\tau(\theta),\tau\leqslant t)=w_t$, where $y_t(\theta)=S_\theta(w_\tau,\tau\leqslant t)$.

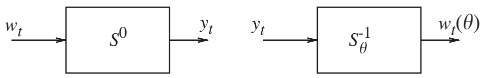

Fig. 1. Scheme for the extraction of $w_t(\theta)$.

The assumptions state that the model class consists of causal systems which are also causally invertible and that the true data generating system belongs to the model class.

3.1. Construction of the confidence region

We next describe the algorithm for the construction of the confidence region.

Algorithm.

(A.1). Compute $w_t(\theta) = S_\theta^{-1}(y_\tau, \tau \leqslant t)$ for $t = 1, 2, \ldots, K$.

(A.2). Choose an integer $s \geqslant 0$ and let $\mathbf{e} = (e_0, e_1, \ldots, e_s)$ be a vector of nonnegative integers such that at least one of the e_j, $0 \leqslant j \leqslant s$, is odd (the way $\mathbf{e}$ should be chosen is discussed later). For every $t = 1, 2, \ldots, K - s = N$, compute

$$f_{t,\mathbf{e}}(\theta) = \prod_{j=0}^{s} w_{t+j}(\theta)^{e_j}.$$

(A.3). Let $I^N = \{1, \ldots, N\}$ and consider a collection G^N of different subsets $I_i^N \subseteq I^N$, $i=1, \ldots, M$, forming a group under the symmetric difference operation (i.e. $(I_i^N \cup I_j^N) - (I_i^N \cap I_j^N) \in G^N$ if $I_i^N, I_j^N \in G^N$). Suppose, without loss of generality, that I_M^N is the zero element of the group G^N: $I_M^N = \emptyset$, the empty set. Compute

$$g_{i,\mathbf{e}}^N(\theta) = \frac{1}{\# I_i^N} \sum_{k \in I_i^N} f_{k,\mathbf{e}}(\theta), \quad i = 1, \ldots, M - 1$$

(# stands for "number of elements in the set").

(A.4). Select an integer q in the interval $[1, (M+1)/2)$ and find the confidence region $\Theta_{\mathbf{e}}^N$ where at least q of the $g_{i,\mathbf{e}}^N(\theta)$ functions are bigger than zero and at least q are smaller than zero.

The intuitive idea behind the algorithm is as follows. For the true parameter vector θ^0, $w_t(\theta^0) = w_t$ is an independent sequence symmetrically distributed around zero. Since at least one e_j is odd, $f_{t,\mathbf{e}}(\theta^0)$ is a zero mean random variable. Moreover, when $\theta = \theta^0$, the functions $g_{i,\mathbf{e}}^N(\theta)$,$i = 1, \ldots, M - 1$, are sums of zero mean random variables. It is therefore unlikely that nearly all of them are positive or that nearly all of them are negative. Based on this observation we exclude the regions in parameter space where the $g_{i,\mathbf{e}}^N(\theta)$ functions take on positive or negative values too many times.

Note that the construction of $\Theta_{\mathbf{e}}^N$ does not require any knowledge of the characteristics of the noise w_t. The Algorithm let the data speak for themselves and constructs the region $\Theta_{\mathbf{e}}^N$ correspondingly: $\Theta_{\mathbf{e}}^N$ does depend on the noise level, but this is through data only, not through a priori assumptions.

The next theorem says that the Algorithm always produces a region that contains θ^0 with a probability chosen by the user.

Theorem 1. *The region $\Theta_{\mathbf{e}}^N$ constructed above has the property that*

$$P[\theta^0 \in \Theta_{\mathbf{e}}^N] = 1 - 2q/M.$$

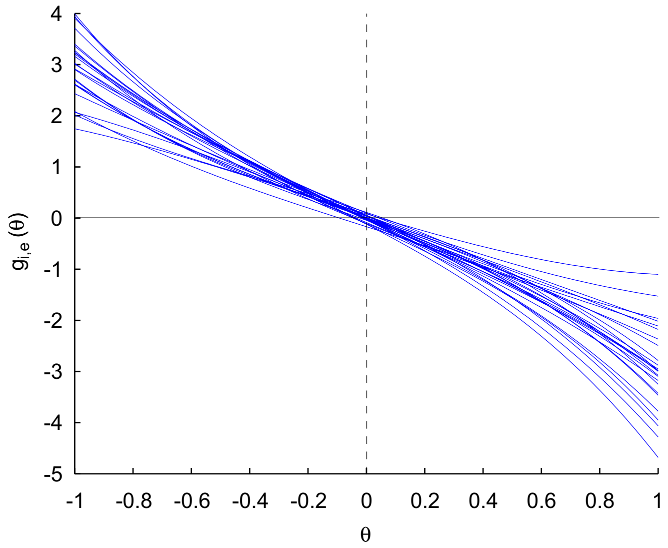

Fig. 2. Some of the $g_{i,\mathbf{e}}^N(\theta)$ functions obtained with $\mathbf{e} = (2, 1)$, $N = 1000$.

Proof. See Appendix A.1. □

Thus, the user controls the probability that $\theta^0 \in \Theta_{\mathbf{e}}^N$ via the choice of q.

In general, in order to determine a confidence region of suitable shape we may want to intersect several regions $\Theta_{\mathbf{e}}^N$ obtained with different $\mathbf{e}$ vectors. If we have h vectors $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_h$, then the confidence region is given by

$$\Theta^N = \bigcap_{l=1}^{h} \Theta_{\mathbf{e}_l}^N. \tag{4}$$

Theorem 2. *The region Θ^N constructed above has the property that*

$$P[\theta^0 \in \Theta^N] \geqslant 1 - 2hq/M. \tag{5}$$

Proof. The proof follows from Theorem 1. The inequality in (5) is due to possible overlaps between the events $\theta^0 \notin \Theta_{\mathbf{e}_l}^N$, $l = 1, \ldots, h$. □

To make the procedure more concrete, we next apply it to the example in Section 2.

Example 3. Suppose we want to find a 90% confidence region for θ^0. Since $y_t = \theta(y_{t-1}^2 - 1) + w_t$, let $w_t(\theta) = y_t - \theta(y_{t-1}^2 - 1)$. Note that, in this example, $w_t(\theta) = \varepsilon_t(\theta)$, the prediction error. In Section 2, we established that $E[\varepsilon_t(\theta)^2 \varepsilon_{t+1}(\theta)] = 0$ only for $\theta = \theta^0$. Motivated by this observation we take $\mathbf{e} = (2, 1)$.

We simulated the system with $N = 1000$ and constructed the group G^N as explained in Appendix A.3 with $M = 256$. We discarded the parameter values where less than $q = 12$ functions out of the $M = 256$ functions were positive or less than 12 functions were negative. Fig. 2 shows some of the obtained $g_{i,\mathbf{e}}^N(\theta)$ functions. The confidence interval for θ^0 turned out to be $[-0.05, 0.03]$.

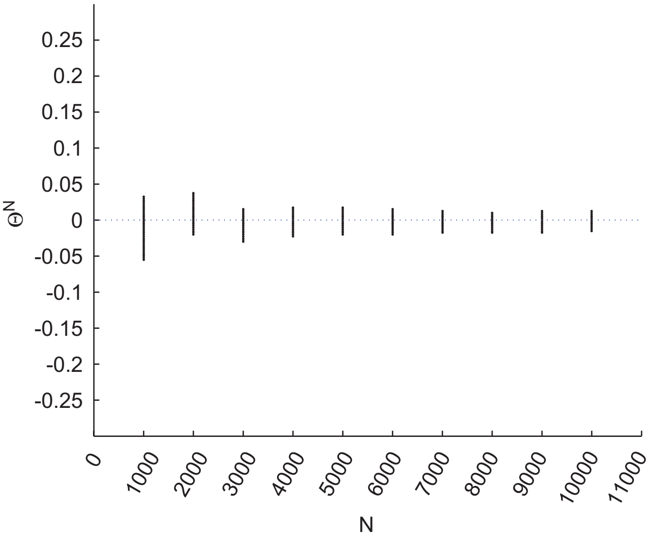

Fig. 3. Ninety percent confidence regions with $\mathbf{e} = (2, 1)$ for increasing N.

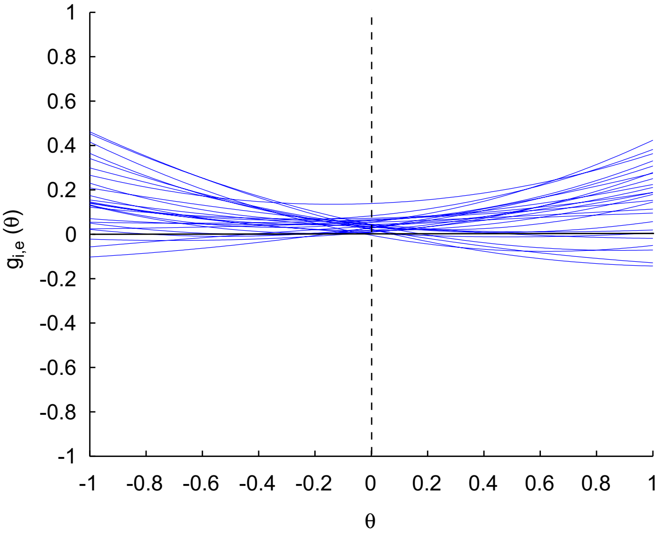

Fig. 4. Some of the $g^N_{i,\mathbf{e}}(\theta)$ functions obtained with $\mathbf{e} = (1, 1)$, $N = 1000$.

The $g^N_{i,\mathbf{e}}(\theta)$ functions are estimates of the third order statistic $E[w_t^2(\theta)w_{t+1}(\theta)]$. From Eq. (3), $E[w_t^2(\theta)w_{t+1}(\theta)] = -2\theta$, so the $g^N_{i,\mathbf{e}}(\theta)$ functions cut the θ axis near $\theta=0$ and the confidence region is a neighborhood of $0 = \theta^0$.

As N increases one expects that the $g^N_{i,\mathbf{e}}(\theta)$ functions become better and better approximations of the function -2θ. Correspondingly, the confidence interval is expected to shrink around $\theta = 0$. In Fig. 3 the confidence intervals obtained for increasing values of N are shown, and the trend is that the length of the intervals decreases as N increases. Section 3.2 provides a general study of the convergence properties of the algorithm.

As a comparison, Fig. 4 shows some of the $g^N_{i,\mathbf{e}}(\theta)$ functions obtained using the second order statistic $E[w_t(\theta)w_{t+1}(\theta)]$, i.e. by choosing $\mathbf{e} = (1, 1)$. As we noticed in Section 2, $E[w_t(\theta)w_{t+1}(\theta)] = 0$ for all values of θ; hence the $g^N_{i,\mathbf{e}}(\theta)$ functions are all flat along the θ axis, and the confidence interval does not shrink around $\theta = 0$. □

Theorems 1 and 2 are very general and apply to any statistic as described in point A.2 of the algorithm. Consequently, the probability of the obtained region is always guaranteed. On the other hand, the effect of the used statistics shows up in the shape of the obtained region. Determining suitable statistics is a problem for which no general guidelines can be given, and the user should choose the statistics based on an analysis of the system class at hand. See also Section 4 for an example.

3.2. Asymptotic behavior

When $N \to \infty$, we would like the confidence region Θ^N given by (4) to shrink around the true value θ^0. In this section, we discuss general conditions for this to happen.

We need the following additional assumptions (while some of these assumptions can be relaxed, we have preferred to maintain them to avoid very technical mathematical derivations).

Assumptions.

(iv) The input noise w_t is independent and identically distributed (i.i.d.).

(v) For every θ the considered statistics are in L^1, i.e. $E[|f_{t,\mathbf{e}_l}(\theta)|] < \infty, l = 1, 2, \ldots, h$, and θ^0 is the only solution to the set of conditions $E[f_{t,\mathbf{e}_l}(\theta)] = 0$, $l = 1, 2, \ldots, h$.

(vi) The groups G^N are constructed as explained in Appendix A.3, the value of M is fixed and the value of N is increasing.

Theorem 4. *Under the hypotheses above, for every fixed* $\theta \neq \theta^0$,

$$P[\exists \bar{N} \mid \theta \notin \Theta^N, \forall N > \bar{N}] = 1.$$

Proof. See Appendix A.2.

In other words, Theorem 4 says that any $\theta \neq \theta^0$ is eliminated from Θ^N starting at some $\bar{N}$ with probability 1.

Remark 5. The Algorithm in Section 3.1 can be generalized so that the assumption of Theorem 4 that $E[|f_{t,\mathbf{e}_l}(\theta)|] < \infty, l = 1, 2, \ldots, h$, is certainly satisfied. In points A.2 and A.3 of the Algorithm, the $f_{t,\mathbf{e}}(\theta)$ functions can be replaced by more general expressions. By inspection of the proof of Theorem 1, we see that the only property of $f_{t,\mathbf{e}}(\theta)$ used is that $f_{t,\mathbf{e}}(\theta)$ is a function of $w_t(\theta), w_{t+1}(\theta), \ldots, w_{t+s}(\theta)$ which is even or odd in all arguments and odd in at least one argument. For example, suppose $s = 2$ and $\mathbf{e} = (2, 1, 2)$, then $f_{t,\mathbf{e}}(\theta) = w_t(\theta)^2 w_{t+1}(\theta) w_{t+2}(\theta)^2$. This function is even in $w_t(\theta)$ and $w_{t+2}(\theta)$ and odd in $w_{t+1}(\theta)$. However, other functions than monomials exhibit the same odd–even structure. For example, the function $\tanh^2(w_t(\theta)) \tanh(w_{t+1}(\theta)) \tanh^2(w_{t+2}(\theta))$, where tanh is the hyperbolic tangent, can be used and Theorem 1 still holds. This observation makes it easier to satisfy the first part of Assumption (v) where it is required that $E[|f_{t,\mathbf{e}}(\theta)|] < \infty$ since such a condition is automatically satisfied by considering bounded functions such as $\tanh^2(w_t(\theta))\tanh(w_{t+1}(\theta))\tanh^2(w_{t+2}(\theta))$.

4. Application example: a simple bilinear system

Here we illustrate the proposed approach on a bilinear system, see Bruni, Di Pillo, and Koch (1974), Fnaiech and Ljung (1987), Mohler and Kolodziej (1980), Priestley (1991), and Subba Rao (1981).

Consider the system

$$y_t = \theta^0 y_{t-2} w_{t-1} + w_t, \tag{6}$$

where w_t is i.i.d. with symmetric distribution around zero and with unit variance. This system has been studied in detail in Terdik and Máth (1998). By iterating (6), it is easy to see that the output y_t can, for any $q \geqslant 1$, be written as

$$y_t = \sum_{k=0}^{q-1} \theta^{0^k} w_{t-2k} \prod_{j=1}^{k} w_{t-2j+1} + \theta^{0^q} y_{t-2q} \prod_{j=1}^{q} w_{t-2j+1}. \tag{7}$$

Note that the product $\prod_{j=1}^{q} w_{t-2j+1}$ has second order moment equal to 1 for any q:

$$E\left[\left(\prod_{j=1}^{q} w_{t-2j+1}\right)^2\right] = \prod_{j=1}^{q} E[w_{t-2j+1}^2] = 1.$$

Thus, if $|\theta^0| < 1$, by letting $q \to \infty$ in (7) we can take

$$y_t = \sum_{k=0}^{\infty} \theta^{0^k} w_{t-2k} \prod_{j=1}^{k} w_{t-2j+1} \tag{8}$$

as a candidate stationary solution. A calculation omitted here shows that the series on the right-hand side of (8) is indeed convergent in the L^2-sense as well as almost surely, the limit is stationary and it satisfies the system Eq. (6). We will refer to this stationary solution in what follows.

A simulation with $\theta^0 = 0.2$ and w_t normally distributed with zero mean and unit variance was carried out. A confidence region was then constructed as explained next.

Following the procedure in the previous section, $w_t(\theta)$ was obtained by applying the inverse system S_θ^{-1} ($|\theta| < 1$) to the output y_t, which can be done by solving the recursive relation

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

Note that for $\theta = 0$ we have $w_t(0) = y_t$. This has important consequences: after some cumbersome calculations it is possible to show that y_t satisfies $E[y_t y_{t+r}] = 0$ for every $r > 0$ and $E[y_t y_{t+r} y_{t+l}] = 0$ for every $l \geqslant r \geqslant 0$ except for $(r, l) = (1, 2)$. Since $w_t(0) = y_t$, this implies that—independently of the true parameter θ^0—the value $\theta = 0$ is a solution of the equations $E[w_t(\theta) w_{t+r}(\theta)] = 0$ for every $r > 0$ and $E[w_t(\theta) w_{t+r}(\theta) w_{t+l}(\theta)] = 0$ for every $l \geqslant r \geqslant 0$ with $(r, l) \neq (1, 2)$. So it is clear that the only possible statistic (up to third order) is $E[w_t(\theta) w_{t+1}(\theta) w_{t+2}(\theta)]$. Indeed, this choice turns out to be an effective one since it can be shown that the only solution to $E[w_t(\theta) w_{t+1}(\theta) w_{t+2}(\theta)] = 0$ is the true parameter $\theta = \theta^0$.

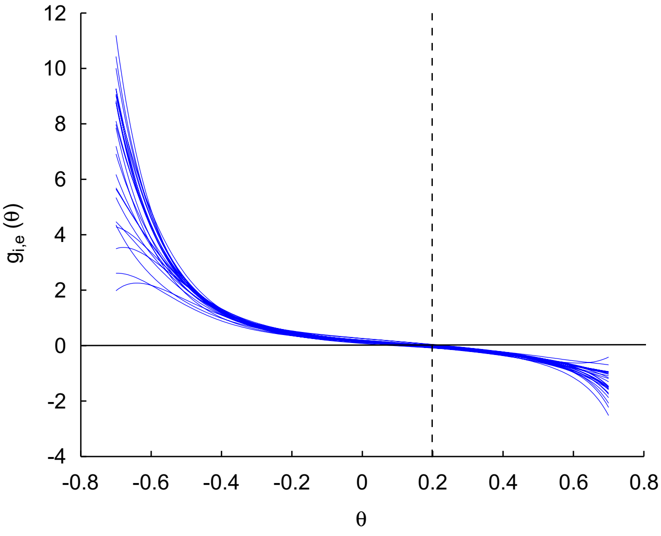

Fig. 5. Some of the $g_{i,\mathbf{e}}^N(\theta)$ functions obtained with $\mathbf{e} = (1, 1, 1)$, $N = 1000$.

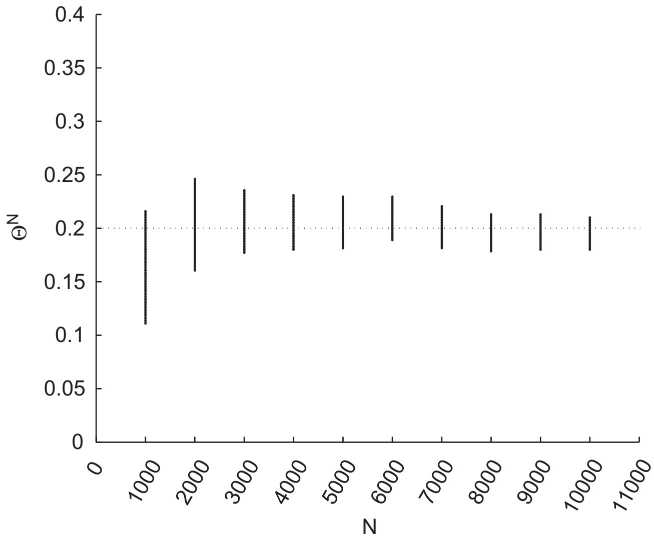

Fig. 6. Ninety percent confidence regions with $\mathbf{e} = (1, 1, 1)$ for increasing N.

Following the above reasoning, we selected $\mathbf{e} = (1, 1, 1)$ in point A.1 in Section 3.1. The group G^N was constructed as in Appendix A.3 with $M = 256$, and the functions $g_{i,\mathbf{e}}^N(\theta)$, for $i = 1, 2, \ldots, M-1$, are given by

$$g_{i,\mathbf{e}}^N(\theta) = \frac{1}{\# I_i^N} \sum_{k \in I_i^N} w_k(\theta) w_{k+1}(\theta) w_{k+2}(\theta).$$

Some of the $g_{i,\mathbf{e}}^N(\theta)$ functions obtained with $N = 1000$ are shown in Fig. 5. The corresponding 90% confidence region for θ^0 turned out to be [0.11, 0.21]. In Fig. 6 the confidence regions for different values of N are plotted.

5. Conclusion

In this paper we have derived a method for construction of confidence regions for the parameters of nonlinear systems. The obtained confidence regions have guaranteed probability to contain the true parameter value for any finite number of data points. Moreover, the confidence regions shrink around the true θ^0 under natural assumptions on the data generating system and the model class provided the higher order statistics are suitably chosen.

Acknowledgments

Research partly supported by MIUR under the project “New Methods for Identification and Adaptive Control for Industrial Systems”, by EC contract IST-2000-28304 “SPATION” and by the Australian Research Council under the Discovery Grant DP 0558579.

Appendix A. Proofs

A.1. Proof of Theorem 1

The proof is similar to the proof of Theorem 2.1 in Campi and Weyer (2005), the only difference being that Proposition A.1 in appendix A of Campi and Weyer (2005) is replaced by Proposition 6 below. Throughout, we omit to indicate explicitly the dependence on N and we, e.g. write G for G^N and I_i for I_i^N.

Proposition 6. *Let w_t be a sequence of independent random variables with symmetric distribution around zero. Let $I = \{1, \dots, N\}$, and let G be a collection of subsets $I_i \subseteq I$, $i = 1, \dots, M$, forming a group under the symmetric difference operation (i.e. $I_i \Delta I_j := (I_i \cup I_j) - (I_i \cap I_j) \in G$ if $I_i, I_j \in G$). Choose an integer $s \geqslant 0$ and let $\mathbf{e} = (e_0, e_1, \dots, e_s)$ be a vector of nonnegative integers such that at least one e_j is odd. For every $t \in I$, let $W_t = \prod_{j=0}^{s} w_{t+j}^{e_j}$. Pick any $\bar I \in G$; then, the set of variables*

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

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

$$\left\{ \sum_{k \in I_i} W_k - \sum_{k \in \bar I} W_k, \quad i = 1, \dots, M \right\}, \tag{A.2}$$

provided that the order of the variables is suitably rearranged.

Proof. The idea of the proof is to introduce new variables $\tilde w_t = -w_t$ for some of the w_t and to rewrite these w_t as $-\tilde w_t$ in (A.2) in such a way that the set (A.2) is written as (A.1) with some of the w_t replaced with $\tilde w_t$. As w_t is symmetrically distributed around 0, w_t and $\tilde w_t$ will have the same distribution and (A.2) and (A.1) will have the same joint M-dimensional distribution.

Consider the whole set of elements

$$W_1, \ W_2, \ W_3, \ \dots, \ W_N. \tag{A.3}$$

We scan these elements from left to right and we rewrite some of them in the new notation. Starting from W_1, we do not change anything until we find an element—say $W_{\bar k}$—in the set $\{W_k, k \in \bar I\}$. Recall that

$$W_{\bar k} = w_{\bar k}^{e_0} w_{\bar k+1}^{e_1} \cdots w_{\bar k+s}^{e_s}.$$

Let p be the maximum integer such that e_p is odd and define $\tilde w_{\bar k+p} = -w_{\bar k+p}$. Then rewrite $W_{\bar k}$ as

$$W_{\bar k} = -w_{\bar k}^{e_0} w_{\bar k+1}^{e_1} \cdots \tilde w_{\bar k+p}^{e_p} \cdots w_{\bar k+s}^{e_s}.$$

We next substitute the old variable $w_{\bar k+p}$ with the new one $-\tilde w_{\bar k+p}$ in all other elements W_k of the sequence (A.3) where the variable $w_{\bar k+p}$ shows up. The important thing to note is that the substitution of $w_{\bar k+p}$ with $-\tilde w_{\bar k+p}$ does not introduce any “minus” sign in front of the elements W_k with $k < \bar k$. In fact, if $w_{\bar k+p}$ is contained in an element $W_{k'}$ with $k' < \bar k$ then, by construction, this $w_{\bar k+p}$ is raised to an even exponent. Thus, with this substitution only the signs of the W_k for $k > \bar k$ can be affected. We continue with our procedure and check the sign of $W_{\bar k+1}$, $W_{\bar k+2}$and so on. If the generic element W_k has sign “+” and $k \in \bar I$, or if W_k has sign “−” and $k \notin \bar I$, we substitute the variable w_{k+p} with $-\tilde w_{k+p}$, stopping the procedure when all the W_k have been scanned. (See Example 7 at the end of the proof for an example of this procedure.)

Set $v_k = w_k$ if w_k has not been substituted and $v_k = \tilde w_k$ if w_k has been substituted. Define the new elements $V_k = \prod_{j=0}^{s} v_{k+j}^{e_j}$. If $k \in \bar I$ we have $W_k = -V_k$, while if $k \notin \bar I$ $W_k = V_k$. Now, the ith element of (A.2) is given by

$$\begin{aligned} \sum_{k \in I_i - \bar I} W_k - \sum_{k \in \bar I - I_i} W_k &= \sum_{k \in I_i - \bar I} V_k + \sum_{k \in \bar I - I_i} V_k \\ &= \sum_{k \in I_i \Delta \bar I} V_k. \end{aligned} \tag{A.4}$$

As G is a group under the symmetric difference, the set $\{I_i \Delta \bar I, \ i = 1, \dots, M\}$ coincides with the set $\{I_i, \ i = 1, \dots, M\}$. This means that (A.2) can be written, by reordering the elements and using (A.4), as

$$\left\{ \sum_{k \in I_i} V_k, \quad i = 1, \dots, M \right\}. \tag{A.5}$$

But, for every k, v_k and w_k have the same distribution and, as the w_k are independent, so are the v_k. Thus, for every k, W_k and V_k have the same distribution and, more generally, the set of variables in (A.5) has the same joint M-dimensional distribution as the set of variables in (A.1). □

Example 7. For the sake of clarity we give a simple example illustrating the procedure explained in the proof for the substitutions of the w_k with $-\tilde{w}_k$. Set $I=\{1,\dots,7\}$, $s=4$, $\mathbf{e}=(1,0,3,2)$ and $\bar{I}=\{2,4,5\}$. The sequence of elements W_k is

$$w_1 w_3^3 w_4^2,\quad w_2 w_4^3 w_5^2,\quad w_3 w_5^3 w_6^2,\quad w_4 w_6^3 w_7^2,$$
$$w_5 w_7^3 w_8^2,\quad w_6 w_8^3 w_9^2,\quad w_7 w_9^3 w_{10}^2.$$

We consider these elements from left to right. As $1 \notin \bar{I}$, we skip W_1. Then we find that $2 \in \bar{I}$. Here, $p=2$, so that we substitute w_4 with $-\tilde{w}_4$ obtaining

$$w_1 w_3^3 \tilde{w}_4^2,\quad -w_2 \tilde{w}_4^3 w_5^2,\quad w_3 w_5^3 w_6^2,\quad -\tilde{w}_4 w_6^3 w_7^2,$$
$$w_5 w_7^3 w_8^2,\quad w_6 w_8^3 w_9^2,\quad w_7 w_9^3 w_{10}^2.$$

Note that the substitution of w_4 has not changed the sign of W_1. Continuing, we skip W_3 and W_4 because their signs are already correct (there is a "+" in front of W_3 and $3 \notin \bar{I}$, and there is a "−" in front of W_4 and $4 \in \bar{I}$). We stop again at W_5, which is written without a "−" while $5 \in \bar{I}$. Thus, we substitute w_7 with $-\tilde{w}_7$ obtaining

$$w_1 w_3^3 \tilde{w}_4^2,\quad -w_2 \tilde{w}_4^3 w_5^2,\quad w_3 w_5^3 w_6^2,\quad -\tilde{w}_4 w_6^3 \tilde{w}_7^2,$$
$$-\, w_5 \tilde{w}_7^3 w_8^2,\quad w_6 w_8^3 w_9^2,\quad -\tilde{w}_7 w_9^3 w_{10}^2.$$

Finally, we skip W_6 and we stop at W_7 because there is a "−", but $7 \notin \bar{I}$. Thus we change w_9 with $-\tilde{w}_9$ obtaining

$$w_1 w_3^3 \tilde{w}_4^2,\quad -w_2 \tilde{w}_4^3 w_5^2,\quad w_3 w_5^3 w_6^2,\quad -\tilde{w}_4 w_6^3 \tilde{w}_7^2,$$
$$-\, w_5 \tilde{w}_7^3 w_8^2,\quad w_6 w_8^3 \tilde{w}_9^2,\quad \tilde{w}_7 \tilde{w}_9^3 w_{10}^2,$$

and the procedure is completed.

A.2. Proof of Theorem 4

We will prove that with probability 1 the functions $g_{i,\mathbf{e}_l}^N(\theta)$, $i=1,\dots,M-1$, tend to $E[f_{t,\mathbf{e}_l}(\theta)]$ when N goes to infinity. For $\theta \neq \theta^0$ there is an l such that $E[f_{t,\mathbf{e}_l}(\theta)] \neq 0$ (see assumption (v)), and for that value of l, when $N \to \infty$ all the $g_{i,\mathbf{e}_l}^N(\theta)$, $i=1,\dots,M-1$, will have the same sign as $E[f_{t,\mathbf{e}_l}(\theta)]$. Consequently, θ will be discarded from Θ^N for N large enough, as stated in the theorem.

Take an element I_i^N in the group G^N. For a fixed i, we consider the elements in I_i^N for increasing N. Note first that I_i^N is a set increasing with N, i.e. $I_i^{N_1} \subseteq I_i^{N_2}$ if $N_1 \leqslant N_2$. Let $N=n(M-1)$, for $n=1,2,\dots$, that is we restrict attention to N that are multiples of $(M-1)$ (the case of generic N's easily follows). The set $I_i^{n(M-1)}$ can be decomposed as

$$I_i^{n(M-1)} = \bigcup_{j \in I_i^{(M-1)}} \{j, j+(M-1), \dots, j+(n-1)(M-1)\},$$

so focusing on subsets of regularly spaced indices. We now have

$$\begin{aligned} g_{i,\mathbf{e}_l}^{n(M-1)}(\theta) &= \frac{1}{\# I_i^{n(M-1)}} \sum_{k \in I_i^{n(M-1)}} f_{k,\mathbf{e}_l}(\theta) \\ &= \frac{1}{n \cdot \# I_i^{M-1}} \sum_{j \in I_i^{M-1}} \sum_{r=0}^{n-1} f_{j+r(M-1),\mathbf{e}_l}(\theta) \\ &= \frac{1}{\# I_i^{M-1}} \sum_{j \in I_i^{M-1}} \frac{1}{n} \sum_{r=0}^{n-1} f_{j+r(M-1),\mathbf{e}_l}(\theta). \end{aligned} \tag{A.6}$$

We want to show that

$$\frac{1}{n} \sum_{r=0}^{n-1} f_{j+r(M-1),\mathbf{e}_l}(\theta) \to E[f_{t,\mathbf{e}_l}(\theta)] \quad \text{a.s.,} \tag{A.7}$$

for any j, so concluding the proof.

w_t is an i.i.d. process and hence it is strict sense stationary and ergodic. Since $w_t(\theta) = S_\theta^{-1}(y_\tau, \tau \leqslant t)$ and $y_t = S^0(w_\tau, \tau \leqslant t)$ we have that $w_t(\theta)$ is a function of $w_t, w_{t-1}, \dots,$ etc. and $f_{j+r(M-1),\mathbf{e}_l}(\theta)$ is a function of $w_{j+r(M-1)+s}, w_{j+r(M-1)+s-1}, \dots,$ etc. Thus $f_{j+r(M-1),\mathbf{e}_l}(\theta)$ inherits from w_t the property of being strict sense stationary and ergodic, from which (A.7) follows from Birchoff–Khinchin theorem (see Shiryaev, 1991, Theorem 3 in Section 3, Chapter 5).

The reader may be interested in noting that the splitting of (A.6) in a double summation formula is necessary because, even though $f_{t,\mathbf{e}_l}(\theta)$ is stationary, $f_{k,\mathbf{e}_l}(\theta)$, $k \in I_i^{n(M-1)}$, is in general not a stationary sequence due to the irregular sampling.

A.3. Group construction

Given a set $I^N=\{1,2,\dots,N\}$ and an integer $M=2^m$, we use the following extension of Gordon's method, (Gordon, 1974), for constructing a collection G^N of M subsets I_i^N, $i=1,\dots,M$, which is a group under the symmetric difference.

(1) Generate an $M \times (M-1)$ matrix Q^{M-1} using Gordon's construction (Gordon, 1974). That is, let $R(1)=[1]$, and recursively compute ($k=2,3,\dots,m$)

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

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

$$Q^{M-1} = \begin{bmatrix} R(m) \\ 0^{\mathrm{T}} \end{bmatrix}.$$

(2) Construct the matrix

$$Q = [Q^{M-1} \quad Q^{M-1} \cdots Q^{M-1}]$$

by listing enough Q^{M-1} matrices so that Q has at least N columns and then extract the submatrix Q^N of Q containing the first N columns of Q. The so obtained Q^N is the incidence matrix of G^N, i.e. the matrix with generic element $Q^N(i, j) = 1$ if $j \in I_i^N$ and zero otherwise.

References

Bruni, C., Di Pillo, G., & Koch, G. (1974). Bilinear systems: An appealing class of nearly linear systems in theory and applications. *IEEE Transactions on Automatic Control*, *AC-19*, 334–348.

Campi, M. C., & Weyer, E. (2005). Guaranteed non-asymptotic confidence regions in system identification. *Automatica*, *41*(10), 1751–1764 Extended version available at ⟨http://www.ing.unibs.it/∼campi⟩.

Campi, M. C., & Weyer, E. (2006). Identification with finitely many data points: the LSCR approach, *semi-plenary presentation. In Proceedings of the 14th IFAC Symposium on system identification*, (pp. 46–64) *SYSID 2006*, Newcastle, Australia.

Fnaiech, F., & Ljung, L. (1987). Recursive identification of bilinear systems. *International Journal of Control*, *45*(2), 453–470.

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.

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

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

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

Ljung, L. (2001). Estimating linear time-invariant models of nonlinear time-varying systems. *European Journal of Control*, *7*, 203–219.

Mohler, R. R., & Kolodziej, W. J. (1980). An overview of bilinear system theory and applications. *IEEE Transactions Systems Man and Cybernetics*, *SMC-10*, 683–688.

Priestley, M. (1991). *Non-linear and non-stationary time series analysis*. New York: Academic press.

Shiryaev, A. N. (1991). *Probability*. (2nd ed.), New York: Springer.

Subba Rao, T. (1981). On the theory of bilinear time series models. *Journal of Royal Statistical Society Series B*, *43*, 244–255.

Terdik, G. Y., & Máth, J. (1998). A new test of linearity of time series based on the bispectrum. *Journal of Time Series Analysis*, *19*(6), 737–753.

Marco Dalai was born in 1979 in Manerbio, Italy. He obtained his Dr. Eng. degree in Electronic Engineering in 2003 from the University of Brescia, Italy, and, since 2004, he has been a Ph.D. student in Information Engineering at the Department of Electronics for Automation of this same university.

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 a Senior Lecturer. His research interests are in the area of system identification and control.

Marco Claudio Campi is Professor of Automatic Control at the University of Brescia, Italy. He was born in Tradate, Italy, on December 7, 1963. 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. Since 1992, he has been with the University of Brescia, Italy.

Marco Campi is an Associate Editor of Systems and Control Letters, and a past Associate Editor of Automatica and the European Journal of Control. Serves as Chair of the Technical Committee IFAC on Stochastic Systems (SS) and is a member of the Technical Committee IFAC on Modeling, Identification and Signal Processing (MISP) and of the Technical Committee IFAC on Cost Oriented Automation. Moreover, he is a Distinguished Lecturer under the IEEE Control Systems Society (CSS) Program. His doctoral thesis was awarded the "Giorgio Quazza" prize as the best original thesis for year 1988. 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.

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.