Proceedings of the
44th IEEE Conference on Decision and Control, and
the European Control Conference 2005
Seville, Spain, December 12-15, 2005

Parametric Identification of Nonlinear Systems: Guaranteed Confidence Regions

Marco Dalai, Erik Weyer, Marco C. Campi

Abstract— **In this paper we consider the problem of constructing confidence regions for the parameters of nonlinear dynamical systems. The proposed method makes use of higher order statistics and extends a previous algorithm in [3]. The obtained confidence regions are valid for any finite number of data samples and they are nonconservative, in the sense that they contain the true parameter value with an exact probability. The usefulness of the proposed approach is illustrated in simulation examples. The results presented here are preliminary results from an ongoing research on finite sample properties in nonlinear system identification.**

I. Introduction

Asymptotic theory of system identification is widely used for the construction of confidence regions for system parameters even in cases where only finitely many samples are available. However, it has recently been shown (see [1], [5], [6]) that the results obtained using asymptotic theory may be unreliable in the finite sample case. Thus, there is a need for techniques that delivers confidence regions with guaranteed probability when there are only a finite number of data points available.

The LSCR (Leave-out Sign dominant Correlation Regions) method for construction of non-asymptotic confidence regions was developed in [3]. LSCR can be used to find confidence regions for the parameters of linear models under general noise assumption. LSCR has two important features; the probability is guaranteed for any finite number of data points, and the confidence regions concentrate around the true parameter when the number of samples increases. In [3] second order statistics are used for identification of linear systems. In this paper we consider the more general problem of nonlinear system identification. It is well known (see for example [10] for a general discussion, or [13] for the particular case of bilinear systems) that second order statistics are not sufficient in this case. Here we show that it is possible to extend the LSCR framework to higher order statistics, and hence consider finite sample properties of nonlinear system identification.

In this paper we focus on parametric systems without an exogenous variable. We regard a system as an input-output relation between an inaccessible white noise w_t and the output y_t. Below we demonstrate some of the difficulties arising when second order statistics is used for the identification of nonlinear systems, and we also show how these difficulties can be overcome by the use of higher order statistics. In Section II the procedure for construction of confidence regions with guaranteed probability based on higher order statistics is introduced, and in the following section the proposed technique is applied to a simple bilinear system.

M. Dalai and M.C. Campi are with the Department of Electronic for Automation, University of Brescia, via Branze 38, 25123, Brescia, Italy. {marco.dalai,campi}@ing.unibs.it

E. Weyer is with the Department of Electrical and Electronic Engineering, The University of Melbourne, Parkville, VIC 3010, Australia. e.weyer@ee.mu.oz.au

This work is partly supported by MIUR under the project "New methods for Identification and Adaptive Control for Industrial Systems" and by the Australian Research Council under the Discovery Grant Scheme.

A. A nonlinear system example

LSCR is based on empirical evaluations of correlations of the type $E[\epsilon_t(\theta)\epsilon_{t+r}(\theta)]$ of the prediction errors using different subsets of the observed data points. For the true parameters, $\epsilon_t(\theta^0)$ is white noise, and it is therefore unlikely that nearly all estimates of the correlations will be positive or that nearly all estimates will be negative. The regions in the parameter space where the estimates take on positive or negative values too many times are excluded from the confidence set and hence the name LSCR (Leave-out Sign dominant Correlation Regions). However, second order statistics of the type $E[\epsilon_t(\theta)\epsilon_{t+r}(\theta)]$ are not sufficient for the identification of nonlinear systems in general. A simple example illustrating this is given next.

Example 1: Consider the system given by

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

where $\theta^0 = 0$ and w_t is and i.i.d sequence of random variables, symmetrically distributed around zero with unit variance. Suppose we want to use LSCR for construction of a confidence interval for the parameter θ in the model class given by

$$y_t = \theta \left(y_{t-1}^2 - 1\right) + w_t. \tag{2}$$

Following the idea of the LSCR method, we compute the optimal predictor $\hat{y}_t(\theta) = \theta(y_{t-1}^2 - 1)$ and the prediction error $\epsilon_t(\theta) = y_t - \hat{y}_t(\theta)$. For the true parameter θ^0 we have that $\epsilon_t(\theta^0) = w_t$ is white. LSCR constructs confidence regions based on many different empirical evaluations of the correlations $E[\epsilon_t(\theta)\epsilon_{t+r}(\theta)]$. In [3] it is shown that, under suitable conditions, θ^0 is the only value of θ for which these correlations are zero in the case of linear ARMA and ARMAX systems and, since the confidence sets are constructed by leaving out regions where the correlations take on positive or negative values too many times, the obtained confidence region shrinks around the true parameter value $\theta = \theta^0$ as the number of data points grows. Here we show that $E[\epsilon_t(\theta)\epsilon_{t+r}(\theta)] = 0$ does not imply $\theta = \theta^0$ for

0-7803-9568-9/05/$20.00 ©2005 IEEE

the situation in (2), i.e. second order statistics do not suffice. Since in our example $y_t = w_t$, we have

$$\epsilon_t(\theta) = y_t - \hat{y}_t(\theta) = w_t - \theta(w_{t-1}^2 - 1); \tag{3}$$

and

$$E[\epsilon_t(\theta)\epsilon_{t+r}(\theta)] = E[(w_t - \theta(w_{t-1}^2 - 1))\cdot (w_{t+r} - \theta(w_{t+r-1}^2 - 1))]. \tag{4}$$

For $r \geq 2$, $E[\epsilon_t(\theta)\epsilon_{t+r}(\theta)] = 0$ since w_t and $(w_{t-1}^2 - 1)$ are zero mean random variables, and the products in (4) only include terms with different time indices. For $r = 1$ we have $E[\epsilon_t(\theta)\epsilon_{t+1}(\theta)] = -\theta E[w_t(w_t^2 - 1)] = -\theta(E[w_t^3] - E[w_t]) = 0$ since w_t is symmetric. So $E[\epsilon_t(\theta)\epsilon_{t+r}(\theta)] = 0, r \neq 0$, for every value of θ and this implies that it is not possible to establish the true value of θ from the conditions $E[\epsilon_t(\theta)\epsilon_{t+r}(\theta)] = 0$. Thus, the confidence region obtained by using LSCR do not shrink around θ^0 in general for nonlinear systems when the number of samples increases.

Consider now the use of higher order statistics; take for example the condition $E[\epsilon_t^3(\theta)] = 0$. For $\theta = \theta^0$, $E[\epsilon_t^3(\theta)] = E[w_t^3] = 0$ because w_t is symmetric. Again using the properties of w_t we have that

$$E[\epsilon_t^3(\theta)] = E[\big(w_t - \theta\left(w_{t-1}^2 - 1\right)\big)^3] \tag{5}$$

$$= -\theta^3 E[\big(w_{t-1}^2 - 1\big)^3]. \tag{6}$$

Thus, unless $E[\big(w_{t-1}^2 - 1\big)^3] = 0$, the only solution to the equation $E[\epsilon_t^3(\theta)] = 0$ is $\theta = \theta^0 = 0$. In the same way, it is not difficult to see that $E[\epsilon_t^2(\theta)\epsilon_{t+1}(\theta)] = \theta(E[w_t^2] - E[w_t^4]) = \theta(1 - E[w_t^4])$; thus, unless $E[w_t^4] = 1$, $\theta = \theta^0 = 0$ is also the only solution to the condition $E[\epsilon_t^2(\theta)\epsilon_{t+1}(\theta)] = 0$. From above it is clear that an extension of LSCR to nonlinear systems must involve higher order statistics. In the next section we introduce these extensions and apply them to the model class (2) in a simulation example.

II. Extension of LSCR to higher order statistics

Suppose a nonlinear system S^0 belongs to a parameterized system class S_θ, that is $S^0 = S_{\theta^0}$ for some θ^0. Assume S^0 is as a mapping from an independent noise sequence w_t to an observable data sequence y_t. Assume also that w_t is zero mean and symmetrically distributed around zero. The aim is to find a confidence region for the parameter θ^0 from the observed output y_t.

In the LSCR method a sequence $\eta_t(\theta)$ is generated for every value of θ, in such a way that for the true parameter θ^0 we have that $\eta_t(\theta^0) = w_t$. Then, the confidence region for θ^0 is constructed by choosing the values of θ for which $\eta_t(\theta)$ resembles (measured in terms of empirical evaluations of the correlation functions) white noise.

In the case of linear systems, the $\eta_t(\theta)$ sequence can easily be constructed by means of optimal linear prediction, taking advantage of the fact that the prediction error $\epsilon_t(\theta^0) = y_t - \hat{y}_t(\theta^0)$ of a linear system is the input noise. Suppose for example that the model class is $A_\theta(q^{-1})y_t = C_\theta(q^{-1})w_t$. If $\hat{y}_t(\theta)$ is the optimal predictor and we let $\eta_t(\theta) = y_t - \hat{y}_t(\theta)$, then, under the assumption that C_{θ^0} is stable, we have $\eta_t(\theta^0) = w_t$.

For nonlinear systems some extra care is required in the construction of the $\eta_t(\theta)$ sequence because using the prediction error does not lead to an independent sequence in general. To see this, consider the system class $y_t = \theta y_{t-1} + y_{t-1}w_t$. The optimal predictor is $\hat{y}_t(\theta) = \theta y_{t-1}$, and hence $y_t - \hat{y}_t(\theta) = (\theta^0 - \theta)y_{t-1} + y_{t-1}w_t$. For $\theta = \theta^0$ we obtain $y_t - \hat{y}_t(\theta^0) = y_{t-1}w_t$ which is not an independent sequence. The w_t signal can instead be reconstructed by inverting the system, i.e. $\eta_t(\theta) = (y_t - \theta y_{t-1})/y_{t-1}$. In fact, as long as $y_{t-1} \neq 0$, we have $\eta_t(\theta^0) = w_t$. This simple example shows that an extension of the LSCR method to nonlinear systems must consider a more general method than optimal prediction for the construction of the $\eta_t(\theta)$ sequence. Here we use the more general idea of input-output system inversion in order to extract the noise w_t from the output y_t.

The ideas presented above are reflected in the following set of assumptions.

- a.1 The observed data y_t is the output of the system S^0 driven by an independent noise w_t symmetrically distributed around zero, i.e. $y_t = S^0(w_t)$;
- a.2 The system S^0 can be represented within the model class S_θ. This means that there exists a value θ^0 of the parameter such that $S_{\theta^0} = S^0$;
- a.3 The systems in the considered class S_θ are invertible, i.e. there exist inverse systems S_θ^{-1} such that $S_\theta^{-1}(S_\theta(w_t)) = w_t$.

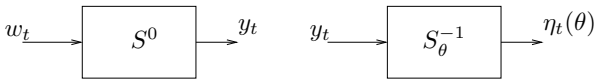

Fig. 1. Scheme for generation of $\eta_t(\theta)$.

A. Construction of the confidence region Θ^N

We first describe the procedure for constructing the region, and then state the main theorem concerning the reliability of the constructed region.

Algorithm

- I.1 Compute $\eta_t(\theta) = S_\theta^{-1}(y_t)$ for $t = 1, 2, \ldots, K$;
- I.2 Choose an integer $s \geq 0$ and let $\mathbf{e} = (e_0, e_1, \ldots, e_s)$ be a vector of nonnegative integers such that e_0 and e_s are greater than zero and at least one of the e_j, $0 \leq j \leq s$, is odd. For every $t = 1, 2, \ldots, K - s = N$, compute
$$f_{t,\mathbf{e}}(\theta) = \prod_{j=0}^{s} \eta_{t+j}(\theta)^{e_j}; \tag{7}$$
- I.3 Let $I = \{1, \ldots, N\}$ and consider a collection G of subsets $I_i \subseteq I, i = 1, \ldots, M$, forming a group under the symmetric difference operation (i.e. $(I_i \cup I_j) - (I_i \cap I_j) \in G$ if $I_i, I_j \in G$). Compute
$$g_{i,\mathbf{e}}^N(\theta) = \frac{1}{\#I_i} \sum_{k \in I_i} f_{k,\mathbf{e}}(\theta), \quad i = 1, \ldots, M-1; \tag{8}$$

where $\#I_i$ is the number of elements in I_i.

I.4 Select an integer q in the interval $[1,(M+1)/2)$ and find the 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 this construction is that, for $\theta = \theta^0$, the functions $g_{i,\mathbf{e}}^N(\theta)$ assume positive or negative value at random ($\eta_t(\theta^0)$ is a symmetrically distributed zero mean independent sequence, and at least one of the powers is odd), so that it is unlikely that almost all of them are positive or that almost all of them are negative. Since point I.4 in the construction of $\Theta_{\mathbf{e}}^N$ discards regions where all $g_{i,\mathbf{e}}^N(\theta)$)'s but a small fraction (q should be taken to be small compared to M) are of the same sign, we expect that $\theta^0 \in \Theta_{\mathbf{e}}^N$ with high probability. This is put on solid mathematical grounds in Theorem 1 below, showing that the probability that $\theta^0 \in \Theta_{\mathbf{e}}^N$ is actually $1-2q/M$. Thus, q is a tuning parameter that has to be selected such that a desired probability of the confidence region is obtained.

Theorem 1: Assume that w_t admits a density and is symmetrically distributed around 0. Then the region $\Theta_{\mathbf{e}}^N$ constructed above is such that

$$P[\theta^0 \in \Theta_{\mathbf{e}}^N] = 1 - 2q/M. \tag{9}$$

Proof: See Appendix A. ■

In general, we will be interested in using more than one $f_{t,\mathbf{e}}(\theta)$ function; thus we will have different vectors $\mathbf{e}_l$ and the final confidence set is constructed as the intersection of the sets $\Theta_{\mathbf{e}_l}^N$. If we have h vectors $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_h$ then the confidence region equals $\Theta^N = \cap_{l=1}^h \Theta_{\mathbf{e}_l}^N$.

Theorem 2: The set Θ^N constructed above has the property that

$$P[\theta^0 \in \Theta^N] \geq 1 - 2hq/M. \tag{10}$$

Proof: The proof follows from Theorem 1. The inequality in (10) is due to that the events $\theta^0 \notin \Theta_{\mathbf{e}_l}^N$ may overlap. ■

Example 2 (Continuation of Example 1): Here we investigate the differences between using second and third order statistics. In this case, since the sequence $\eta_t(\theta)$ is the prediction error $y_t - \hat{y}_t(\theta)$ we continue to use the notation from Example 1 and let $\eta_t(\theta) = \epsilon_t(\theta)$. The input w_t is an i.i.d. sequence of zero mean Gaussian random variables with variance 1, and we seek a 90% confidence interval for θ.

Fig. 2 shows some of the $g_{i,\mathbf{e}}^N(\theta)$ functions obtained for $N = 11500$. These functions are estimates of the second order statistic $E[\epsilon_t(\theta)\epsilon_{t+1}(\theta)]$ and the third order statistic $E[\epsilon_t^3(\theta)]$. Note that these two statistics correspond to the choices $s = 1$, $\mathbf{e} = (1,1)$ and $s = 0$, $\mathbf{e} = (3)$ respectively. It is interesting to note the different behavior of the functions; the second order statistic gives very flat $g_{i,\mathbf{e}}^N(\theta)$ functions which lead to a large confidence interval while the third order statistic gives $g_{i,\mathbf{e}}^N(\theta)$ functions that depart from zero for increasing values of $|\theta|$, leading to a small confidence interval. Even better confidence intervals are obtained with the third order statistic $E[\epsilon_t^2(\theta)\epsilon_{t+1}(\theta)]$ (which corresponds to $s = 1$ and $\mathbf{e} = (2,1)$). Fig. 3 shows the $g_{i,\mathbf{e}}^N(\theta)$ functions

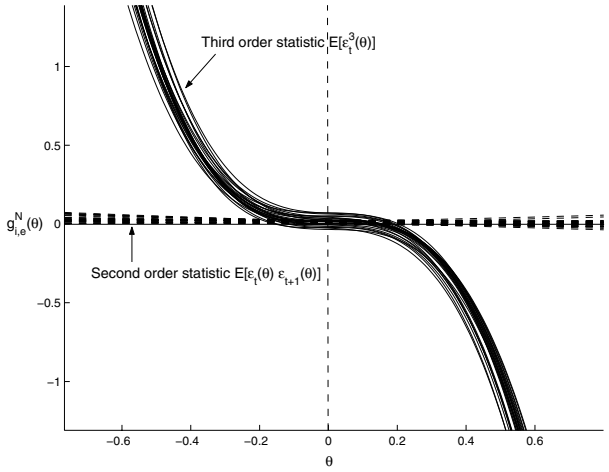

Fig. 2. $g_{i,\mathbf{e}}^N(\theta)$ functions for Example 2, with $N = 11500$, for $\mathbf{e} = (1,1)$ and $\mathbf{e} = (3)$.

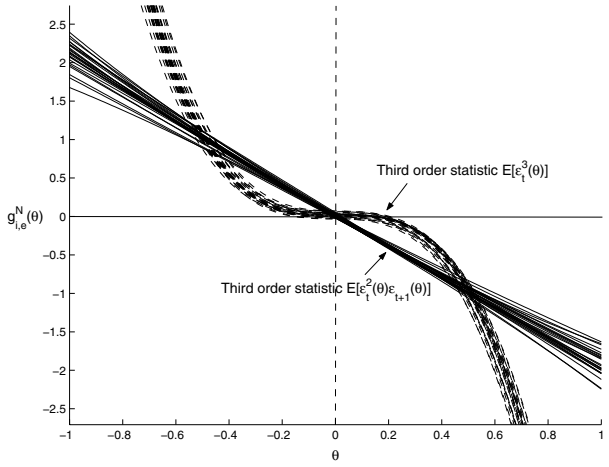

Fig. 3. $g_{i,\mathbf{e}}^N(\theta)$ functions for Example 2, with $N = 11500$, for $\mathbf{e} = (3)$ and $\mathbf{e} = (2,1)$.

obtained with this statistic and the corresponding functions for $E[\epsilon_t^3(\theta)]$. This different behavior is due to the fact that, for $\theta = 0$, $E[\epsilon_t^2(\theta)\epsilon_{t+1}(\theta)] = \theta(1 - E[w_t^4])$ has a zero of order 1 while $E[\epsilon_t^3(\theta)] = -\theta^3 E[\left(w_{t-1}^2 - 1\right)^3]$ has a zero of order 3. Figs. 4 and 5 show the confidence regions obtained for these two third order statistics when N takes different values in the range $[1000, 12000]$.

Remark 2.1: We note here that in point I.2 and I.3 of the construction procedure it is possible to replace $f_{t,\mathbf{e}}$ with other functions. The only property of $f_{t,\mathbf{e}}$ used in the proof of Theorem 1 is that $f_{t,\mathbf{e}}$ is a function of $\eta_t, \eta_{t+1}, \ldots, \eta_{t+s}$ (the argument θ has been omitted) which is even or odd in all arguments and odd in at least one. For example, suppose $s = 2$; a possible candidate for $f_{t,\mathbf{e}}$, obtained choosing $\mathbf{e} = (2,1,3)$, is $\eta_t^2\eta_{t+1}\eta_{t+2}^3$. This function is odd in η_{t+1} and even in η_t and η_{t+2}. On the other hand, other functions than monomials can exhibit the same odd-even structure. For example the function $\cos(\eta_t)\sin(\eta_{t+1})\arctan(\eta_{t+2})$ can be used and Theorem 1 still holds.

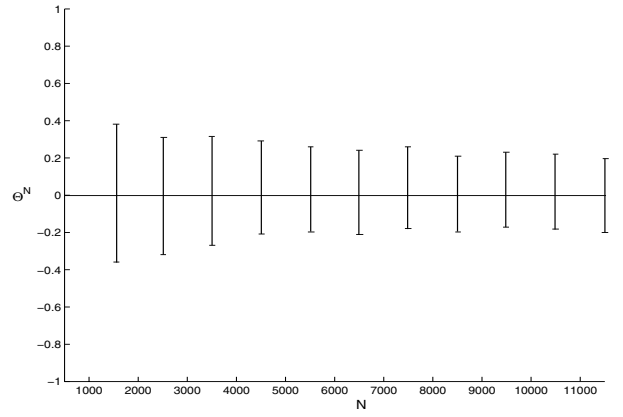

Fig. 4. 90% confidence intervals for an increasing number of data points using the third order statistic $E[\epsilon_t^3(\theta)]$.

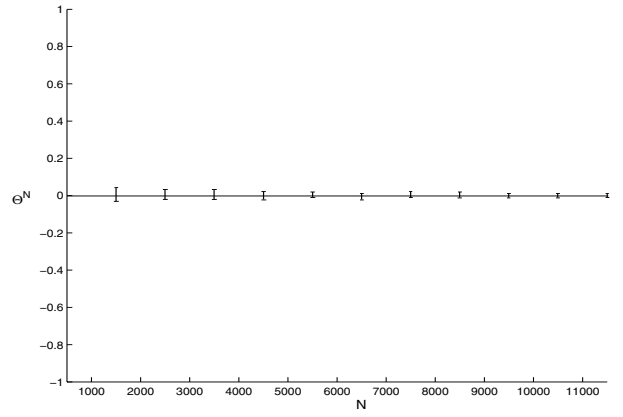

Fig. 5. 90% confidence intervals for an increasing number of data points using the third order statistic $E[\epsilon_t^2(\theta)\epsilon_{t+1}(\theta)]$.

III. Simulation example: a simple bilinear system

In this simulation example we consider a well studied class of nonlinear systems, namely bilinear systems (see [11], [13], [12], [4], [2]). A detailed study of the system used here can be found in [14]. Consider the system

$$y_t = \theta^0 y_{t-2} w_{t-1} + w_t; \tag{11}$$

where w_t is independent and symmetrically distributed around zero with unit variance $\sigma^2 = 1$. By iterating (11) the output y_t can be written as

$$y_t = \sum_{k=0}^{q-1} {\theta^0}^k w_{t-2k} \prod_{j=1}^{k} w_{t-2j+1} + {\theta^0}^q y_{t-2q} \prod_{j=1}^{p} w_{t-2j+1}. \tag{12}$$

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

$$E\left[\left(\prod_{j=1}^{k} w_{t-2j+1}\right)^2\right] = \prod_{j=1}^{k} E[w_{t-2j+1}^2] \tag{13}$$

$$= \sigma^{2k} = 1. \tag{14}$$

Thus, if $|\theta^0| < 1$, by letting $q \to \infty$ in (12) we can take as a candidate stationary solution

$$y_t = \sum_{k=0}^{\infty} {\theta^0}^k w_{t-2k} \prod_{j=1}^{k} w_{t-2j+1}. \tag{15}$$

It can be shown that (15) is the L^2 as well as the almost sure limit of the sum in the first term of (12). Moreover, it is stationary and it satisfies (11).

In the simulations we used $\theta^0 = 0.2$ and w_t was an i.i.d. sequence normally distributed with zero mean and unit variance. $\eta_t(\theta)$ was calculated according to

$$\eta_t(\theta) = y_t - \theta y_{t-2} \eta_{t-1}(\theta). \tag{16}$$

For the value $\theta = 0$ we have $\eta_t(0) = y_t$. After some cumbersome calculations, it is possible to show that the stationary solution 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 $l \geq r \geq 0$ except $(r, l) = (1, 2)$. As $\eta_t(0) = y_t$, this implies that the value $\theta = 0$, regardless of the true parameter θ^0, is a solution to the equations $E[\eta_t(\theta)\eta_{t+r}(\theta)] = 0$ for every $r > 0$ and $E[\eta_t(\theta)\eta_{t+r}(\theta)\eta_{t+l}(\theta)] = 0$ for every $l \geq r \geq 0$ except $(r, l) = (1, 2)$. So it is clear that the only possible statistic (up to the third order) is $E[\eta_t(\theta)\eta_{t+1}(\theta)\eta_{t+2}(\theta)]$, and it can be shown that the only solution to $E[\eta_t(\theta)\eta_{t+1}(\theta)\eta_{t+2}(\theta)] = 0$ is indeed the true parameter $\theta = \theta^0$. In our simulation we have therefore used this third order statistic in order to construct the confidence region for θ^0. We have therefore chosen $s = 2$ and $\mathbf{e} = (1, 1, 1)$ in point I.2 in section II-A. The group G is constructed as in Appendix B, and the functions $g_{i,\mathbf{e}}^N(\theta)$ are given by

$$g_{i,\mathbf{e}}^N(\theta) = \frac{1}{\#I_i} \sum_{k \in I_i} \eta_k(\theta)\eta_{k+1}(\theta)\eta_{k+2}(\theta) \tag{17}$$

for $i = 1, 2, \ldots, M-1$. The 90% confidence intervals obtained for increasing number of data points are shown in Fig. 6. The $g_{i,\mathbf{e}}^N(\theta)$ functions obtained for $N = 12000$ are shown in Fig. 7. In Fig. 8 the obtained $g_{i,\mathbf{e}}^N(\theta)$ functions of the second order statistic with $\mathbf{e} = (1, 1)$ are shown. It is clear that the use of the second order statistic gives useless (even if guaranteed) confidence regions while the use of third order statistic gives good results.

IV. Conclusion

In this paper we have derived a method for the construction of confidence regions for the parameters of nonlinear systems. The proposed technique is based on the use of higher order statistics, and it extends the LSCR method recently introduced in [3] for linear systems. The obtained confidence regions have guaranteed probability for any finite number of data samples. The simulation examples show that the method works well and that it is important to use higher order statistics.

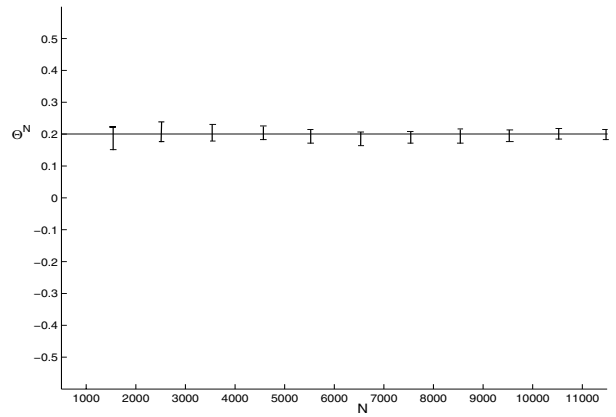

Fig. 6. Confidence regions at 90% with third order statistic $E[\eta_t(\theta)\eta_{t+1}(\theta)\eta_{t+2}(\theta)]$ for increasing number of data points

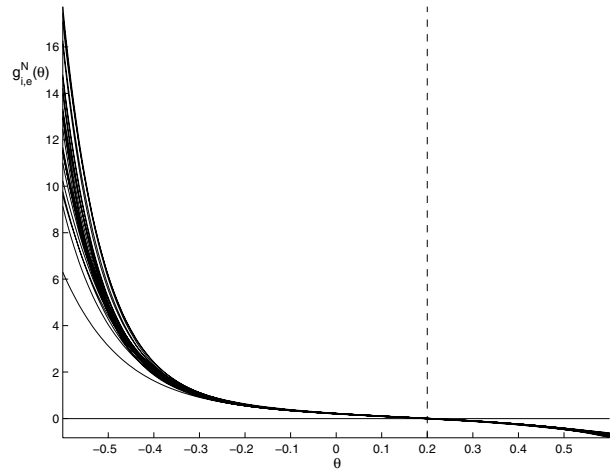

Fig. 7. $g^N_{i,\mathbf{e}}(\theta)$ functions (N=11500) for the bilinear system (11) with $\theta^0 = 0.2$, for $\mathbf{e} = (1,1,1)$ (third order statistic $E[\eta_t(\theta)\eta_{t+1}(\theta)\eta_{t+2}(\theta)]$).

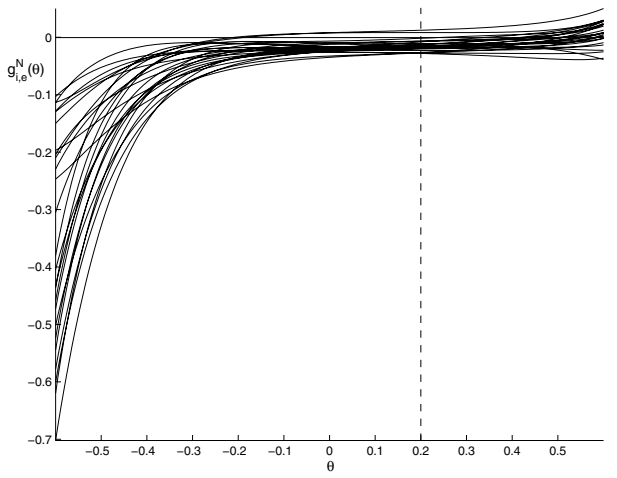

Fig. 8. $g^N_{i,\mathbf{e}}(\theta)$ functions (N=11500) for the bilinear system (11), with $\theta^0 = 0.2$, for $\mathbf{e} = (1,1)$ (second order statistic $E[\eta_t(\theta)\eta_{t+1}(\theta)]$).

REFERENCES

[1] Bittanti S., M.C. Campi and S. Garatti (2002), "New results on the asymptotic theory of system identification for the assessment of the quality of estimated models", *In Proc. 41st Conf. on Decision and Control*, Las Vegas, USA, 2002.

[2] Bruni C., G. DiPillo and G. Koch, "Bilinear systems: an appealing class of nearly linear systems in theory and applications," *IEEE Trans. Automat. Contr.*, vol. AC-19, pp. 334-348, 1974.

[3] Campi M.C. and E. Weyer (2004), "Guaranteed Non-asymptotic Confidence Regions in System Identification", *Automatica* Vol. 41, no. 10, pp. 1751-1764.

[4] Fnaiech F. and L. Ljung, "Recursive identification of bilinear systems," *Int. J. Contr.*, vol. 45, no. 2, pp. 453470, 1987.

[5] Garatti S., M.C. Campi and S. Bittanti (2003), "Model quality assessment for Instrumental Variable methods", *In Proc. 42st Conf. on Decision and Control*, Maui, Hawaii, USA , 2003.

[6] Garatti S., M.C. Campi and S. Bittanti (2004), "Assessing the quality of identified models through the asymptotic theory - When is the result reliable?", *Automatica*, vol. 40, no. 8, pp. 1319-1332.

[7] Gordon, L. (1974), "Completely Separating Groups in Subsampling", *Annals of Statistics*, Vol. 2, pp. 572-578.

[8] Hartigan, J.A. (1969), "Using Subsample Values as Typical Values", *Journal of American Statistical Association*, Vol. 64, pp. 1303-1317.

[9] Hartigan, J.A. (1970), "Exact Confidence Intervals in Regression Problems with Independent Symmetric Errors", *Annals of Mathematical Statistics*, Vol. 41, pp. 1992-1998.

[10] Ljung L. (2001) "Estimating Linear Time-invariant Models of Nonlinear Time-varying Systems". *European Journal of Control*, Vol 7, pp 203-219, 2001.

[11] Mohler, R. R. and W. J. Kolodziej, "An overview of bilinear system theory and applications", *IEEE Trans. Syst., Man., Cybern.*, vol. SMC-10, pp. 683-688, 1980.

[12] Priestley M., "Non-linear and Non-stationary Time Series Analysis". New York: Academic, 1991.

[13] Subba Rao T. (1981),"On the theory of bilinear time series models." *J. Roy. Statist. Soc. Ser. B*, Vol 43, pp 244-255, 1981.

[14] Terdik GY. and J. Máth. (1998),"A new test of linearity of time series based on the bispectrum." *Journal of Time Series Analysis*, Vol 19, no. 6 pp 737-753, 1998.

APPENDIX

A. Proof of Theorem 1

The proof of Theorem 1 follows along the same line as the proof of Theorem 2.1 in [3]. The only difference is in the formulation and proof of Proposition A.1 in appendix A of [3]. Here we state and prove the equivalent proposition for higher order statistics.

Proposition 1.1: Let $\{w_t\}$ be a sequence of independent random variables with symmetric distribution around zero. Let $I = \{1, \dots, N\}$, and let G be a collection of subsets $I_i \subseteq I$, $i = 1 \dots M$, forming a group under the symmetric difference operation (i.e. $I_i \Delta I_j := (I_i \cup I_j) - (I_i \cap I_j) \in G$ if $I_i, I_j \in G$). Take an integer $s \geq 0$ and let $\mathbf{e} = (e_0, e_1, \dots, e_s)$ be a vector of nonnegative integers such that e_0 and e_s are greater than zero and at least one of the e_i is odd. For every $k \in I$, set $W_k = \prod_{i=0}^{s} w_{k+i}^{e_i}$. Pick any $\bar{I} \in G$; then, the set of variables

$$\left\{ \sum_{k \in I_i} W_k, \quad i = 1, \dots, M \right\} \tag{18}$$

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

$$\left\{ \sum_{k \in I_i} W_k - \sum_{k \in \bar{I}} W_k, \quad i = 1, \dots, M \right\}, \tag{19}$$

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 (19) in such a way that the set (19) is written as (18) 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 (19) and (18) will have the same joint M-dimensional distribution. More specifically, we want to substitute some w_t with $-\tilde{w}_t$ in such a way that the number of substitutions in every element W_k is odd if $k \in \bar{I}$ and even if $k \notin \bar{I}$. In other words, we want that W_k is written, in the new notation (i.e. replacing some of the w_t with $-\tilde{w}_t$), with a "minus" in front if and only if $k \in \bar{I}$. We show how this can be done. Consider the whole set of elements

$$W_1\,, \quad W_2\,, \quad W_3\,, \quad \dots\,, \quad W_N \tag{20}$$

We scan these elements from left to right adjusting the signs one by one. Starting from W_1, we do not change sign until we find an element - say $W_{\bar{k}}$ - in the set $\{W_k, k \in \bar{I}\}$. When this happens we change the sign of $W_{\bar{k}}$. Let p be the maximum integer such that e_p is odd; then, we define $\tilde{w}_{\bar{k}+p} = -w_{\bar{k}+p}$ and change the sign of $W_{\bar{k}}$ by substituting $w_{\bar{k}+p}$ with $-\tilde{w}_{\bar{k}+p}$. This means that $W_{\bar{k}}$ is now written as

$$W_{\bar{k}} = -w_{\bar{k}}^{e_0} w_{\bar{k}+1}^{e_1} \cdots \tilde{w}_{\bar{k}+p}^{e_p} \cdots w_{\bar{k}+s}^{e_s} \tag{21}$$

Then, we 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 (20) where the variable $w_{\bar{k}+p}$ shows up. The important thing to note is that the substitution of $w_{\bar{k}+p}$ with $-\tilde{w}_{\bar{k}+p}$ cannot change the sign of any of the elements W_k with $k < \bar{k}$. In fact, suppose $w_{\bar{k}+p}$ is contained in an element $W_{k'}$ with $k' < \bar{k}$; then clearly $w_{\bar{k}+p} = w_{k'+i'}$ with $i' > p$. This means, by the definition of p, that $w_{\bar{k}+p}$ appears in $W_{k'}$ with an even exponent and, thus, substituting it with $-\tilde{w}_{\bar{k}+p}$ does not cause any sign change. Thus, with this substitution we may only affect the signs of the W_k for $k > \bar{k}$, but this is not a problem, since we can readjust them, where necessary, in successive steps. Now that $W_{\bar{k}}$ is processed, 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 *viceversa*) we change the variable w_{k+p} with $-\tilde{w}_{k+p}$, stopping the procedure when all the W_k have been scanned.

Now, 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 $V_k = \prod_{i=0}^{s} v_{k+i}^{e_i}$; then, if $k \in \bar{I}$ we have $W_k = -V_k$, while if $k \notin \bar{I}$ $W_k = V_k$. Now, note that the i-th element of (19) is given by

$$\begin{aligned} \sum_{k \in I_i \setminus \bar{I}} W_k - \sum_{k \in \bar{I} \setminus I_i} W_k &= \sum_{k \in I_i \setminus \bar{I}} V_k + \sum_{k \in \bar{I} \setminus I_i} V_k \\ &= \sum_{k \in I_i \Delta \bar{I}} V_k \end{aligned} \tag{22}$$

Furthermore, as G is a group under the symmetric difference, $I_i \Delta \bar{I} \in G$, $\forall i$, and it is easy to see that $I_i \Delta \bar{I} = I_j$ if and only if $I_j \Delta \bar{I} = I_i$. Thus, the set $\{I_i,\ i = 1, \dots, M\}$ coincides with the set $\{I_i \Delta \bar{I},\ i = 1, \dots, M\}$. This means that (19) can be written, by reordering the elements and using (22), as

$$\left\{ \sum_{k \in I_i} V_k, \quad i = 1, \dots, M \right\}. \tag{23}$$

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 (23) has the same M-dimensional joint distribution as the set of variables in (18). ■

B. Group construction

Given a set $I = \{1, 2, \dots, N\}$ and an integer $M = 2^m$ we use the following extension of Gordon's method [7] for constructing a set G of M subsets $I_i, i = 1, \dots, M$ which is a group under symmetric differences.

1) Generate an $M \times M-1$ matrix Q^{M-1} in the following way. Let $R(1) = [1]$, and recursively compute ($k = 2, 3, \dots, m$)

$$R(k) = \begin{bmatrix} R(k-1) & R(k-1) & 0 \\ R(k-1) & J - R(k-1) & e \\ 0^T & e^T & 1 \end{bmatrix}, \tag{24}$$

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

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

2) Set $n = \lceil N/(M-1) \rceil$ and construct the matrix

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

where Q^{M-1} is repeated n times. The incidence matrix Q^N of the group G (i.e. the matrix with generic element $Q^N(i,j) = 1$ if $j \in I_i^N$ and zero otherwise) is obtained by taking the first N columns of the matrix $Q^{n(M-1)}$.