

Non-Asymptotic Confidence Regions for the Least-Squares Estimate

Balázs Csanád Csáji^{*,**} Marco C. Campi^{***} Erik Weyer^{*}

^{*} *Department of Electrical and Electronic Engineering, The University of Melbourne, Parkville, Melbourne, Victoria, 3010, Australia*

^{**} *Computer and Automation Research Institute, Hungarian Academy of Sciences, Kende utca 13-17, H-1111, Budapest, Hungary*

^{***} *Department of Information Engineering, University of Brescia, Via Branze 38, 25123 Brescia, Italy*

Abstract We propose a new *finite sample* system identification method, called *Sign-Perturbed Sums* (SPS), to estimate the parameters of dynamical systems under mild statistical assumptions. The proposed method constructs *non-asymptotic* confidence regions that include the *least-squares* (LS) estimate and are guaranteed to contain the true parameters with a user-chosen *exact* probability. Our method builds on ideas imported from the “Leave-out Sign-dominant Correlation Regions” (LSCR) approach, but, unlike LSCR, also guarantees the inclusion of the LS estimate and provides confidence regions for multiple parameters with exact probabilities. This paper presents the SPS method for FIR and ARX systems together with its main theoretical properties, as well as demonstrates the approach through simple examples and experiments.

Keywords: system identification, confidence regions, finite sample results, least squares

1. INTRODUCTION

Estimating parameters of dynamical systems from experimental data is one of the fundamental problems of system identification (Söderström and Stoica [1989], Ljung [1999]). Classical solutions, such as the least-squares-, or more generally, prediction error- and correlation- methods typically provide *point estimates* and only offer *asymptotically* guaranteed confidence regions.

In practical applications, however, only a *finite number of measurements* are available, and the *noise characteristics* are *only partially known*, moreover, the noise can as well have changing intensity through time, i.e., it can be *nonstationary*. Furthermore, in many situations, for example, when the safety, stability or quality of a process has to be guaranteed, *confidence regions* are needed in addition to the standard point estimates.

A finite sample system identification method, called “Leave-out Sign-dominant Correlation Regions” (LSCR), was developed in (Campi and Weyer [2005], Dalai et al. [2007], Campi et al. [2009], Campi and Weyer [2010]). LSCR builds non-asymptotic confidence regions for parameters of various (linear and non-linear) dynamical systems under weak assumptions on the noise.

One important theoretical property of the LSCR algorithm is that it constructs regions whose probability is rigorously lower bounded, that is the user is guaranteed that the regions contain the true parameters with a minimum level of probability. However, the LSCR method is not able to provide regions with exact probabilities when more than one parameter is being estimated simultaneously.

In this paper, we introduce a new system identification method, called *Sign-Perturbed Sums* (SPS), that provides

non-asymptotic confidence regions for multiple parameters with *exact* probabilities, and which are guaranteed to include the *least-squares* (LS) estimate. The main ideas underlying the SPS method are to substitute *subsampling* in LSCR with multiplications by *random signs*, and to apply a *norm* to enforce a multi-dimensional ordering.

We begin with introducing the SPS method through simple examples. First, we consider Finite Impulse Response (FIR) systems, then, after presenting a numerical experiment, we generalize the algorithm to Autoregressive Exogenous (ARX) systems. Finally, a theorem is presented about the guarantees of the constructed regions.

2. IDEAS AND INTRODUCTORY EXAMPLES

In this section, we introduce the main ideas in an informal way through examples of increasing complexity. We postpone the more rigorous formulation of the assumptions, the pseudo-code and our main theorem to Section 3.

2.1 First-order FIR Systems

We start by considering the simplest possible FIR model, i.e., a first-order scalar system. Let

$$Y_t \triangleq b_1^* U_{t-1} + N_t, \quad (1)$$

where Y_t is the output, U_t is the input and N_t is the noise at time t . We limit the assumptions and only assume that $(N_t)_t$, which denotes the whole noise sequence, contains independent random variables which are also independent of the inputs, symmetrically distributed about zero, and have densities¹. No other assumptions are imposed, the noise can be nonstationary with unknown distributions.

¹ The density assumption is only introduced for simplicity.

The available data are $(Y_t)_{t=1}^n$ and $(U_t)_{t=0}^{n-1}$. The goal is to construct a confidence interval around the least-squares (LS) estimate which is guaranteed to contain the true parameter b_1^* with an exact and user-chosen probability.

The system (1) can be rewritten as a linear regression

$$Y_t = \varphi_t^T \theta^* + N_t, \quad (2)$$

with $\varphi_t \triangleq [U_{t-1}]$ and $\theta^* \triangleq [b_1^*]$. The assumption that the inputs are independent of the noise ensures that the regressors, φ_t , are *exogenous*². In order to find the least-squares estimate of b_1^* , we first introduce the predictors

$$\hat{Y}_t(\theta) \triangleq b_1 U_{t-1} = \varphi_t^T \theta, \quad (3)$$

where $\theta = [b_1]$ is a generic model parameter. The prediction errors (reconstructed noises) for a given θ are

$$\hat{N}_t(\theta) \triangleq Y_t - b_1 U_{t-1} = Y_t - \varphi_t^T \theta, \quad (4)$$

and the least-squares (LS) estimate is found by minimizing the sum of the squared prediction errors, that is

$$\hat{\theta}_{\text{LS}} \triangleq \arg \min_{b_1} \sum_{t=1}^n (Y_t - b_1 U_{t-1})^2 = \arg \min_{\theta} \sum_{t=1}^n (Y_t - \varphi_t^T \theta)^2,$$

which is achieved by solving the *normal equation*, i.e.,

$$\sum_{t=1}^n U_{t-1} (Y_t - b_1 U_{t-1}) = \sum_{t=1}^n \varphi_t (Y_t - \varphi_t^T \theta) = 0. \quad (5)$$

Before we present our non-asymptotic confidence regions, we recall the construction of the standard, *approximate* confidence ellipsoids of the *asymptotic* theory.

Asymptotic Confidence Regions. It is known that, under some moment conditions, the LS estimates are *asymptotically* normal. More precisely, let $\hat{\theta}_n$ denote the LS estimate based on n data points, then $\sqrt{n}(\hat{\theta}_n - \theta^*)$ converges in distribution to the Gaussian distribution with zero mean and covariance $\Gamma \triangleq \sigma_0^2 (\mathbb{E}[\varphi_0 \varphi_0^T])^{-1}$, where σ_0^2 is the variance of the noise, assuming stationary noises and inputs.

The covariance of the error with n data points is *approximately* $\Gamma_n \triangleq \text{Cov}(\hat{\theta}_n - \theta^*) \approx 1/n \Gamma$, which allows the construction of approximate confidence ellipsoids based on the fact that $n(\hat{\theta}_n - \theta^*)^T \Gamma^{-1} (\hat{\theta}_n - \theta^*)$ converges in distribution to $\chi^2(d)$, i.e., the χ^2 distribution with $\text{dim}(\theta^*) = d = 1$ degrees of freedom (Ljung [1999]). Usually Γ is not known in practice, but it can be estimated. By using estimates for σ_0^2 and $\mathbb{E}[\varphi_0 \varphi_0^T]$, a confidence region can be built as

$$\Theta_n^\mu \triangleq \{ \theta \in \mathbb{R}^d : \|\hat{\theta}_n - \theta\|_{\Phi_n}^2 \leq \mu \hat{\sigma}_n^2 / n \}, \quad (6)$$

$$\Phi_n \triangleq \frac{1}{n} \sum_{t=1}^n \varphi_t \varphi_t^T, \quad \text{and} \quad \hat{\sigma}_n^2 \triangleq \frac{1}{n} \sum_{t=1}^n \hat{N}_t^2(\hat{\theta}_n), \quad (7)$$

where the probability that θ^* is not in Θ_n^μ can be computed as the μ -level of the standard χ^2 distribution.

The confidence regions constructed this way are *random*, since they depend on $\hat{\theta}_n$, $\hat{\sigma}_n$ and Φ_n , which are random. Moreover, they *do not have rigorous guarantees* and are usually *imprecise* for small samples (Ljung [1999]).

² Regressors that are statistically independent of the noise terms are called (strongly) exogenous, otherwise they are called endogenous.

Non-Asymptotic Confidence Regions. Now, we introduce our exact, finite sample confidence regions. As we will see, their construction exploits the symmetry of the noise. Let

$$S_0(\theta) \triangleq \sum_{t=1}^n \varphi_t (Y_t - \varphi_t^T \theta), \quad (8)$$

and further introduce sign-perturbed normal equations

$$S_i(\theta) \triangleq \sum_{t=1}^n \varphi_t \alpha_{i,t} (Y_t - \varphi_t^T \theta), \quad (9)$$

for $i = 1, \dots, m-1$, where $\alpha_{i,t}$ ($t = 1, \dots, n$) are i.i.d. *random signs*, that is $\alpha_{i,t} = \pm 1$ with probability 1/2 each.

Observe that corresponding to the true value, θ^* , we have

$$S_0(\theta^*) = \sum_{t=1}^n \varphi_t N_t, \quad (10)$$

$$S_i(\theta^*) = \sum_{t=1}^n \varphi_t \alpha_{i,t} N_t, \quad (11)$$

where $i = 1, \dots, m-1$. Note that since N_t is symmetrically distributed about 0, N_t and $\alpha_{i,t} N_t$ have the same distribution. Now, consider the order of the $(S_i(\theta^*))_{i=0}^{m-1}$ from the smallest to the largest³. Since $(N_t)_{t=1}^n$ and $(\alpha_{i,t} N_t)_{t=1}^n$ have the same symmetric distribution, all orderings are equally probable. I.e. there is no reason a particular $S_i(\theta^*)$ should be bigger or smaller than another $S_j(\theta^*)$ and the probability that the sum $S_i(\theta^*)$ is the l th largest one is $1/m$ independently of i and l . Moreover, this ordering property still holds if we take squares ($i = 0, \dots, m-1$),

$$Z_i(\theta^*) \triangleq S_i^2(\theta^*). \quad (12)$$

A confidence interval for θ^* can now be obtained as follows. Compute $Z_i(\theta) = S_i^2(\theta)$, $i = 0, \dots, m-1$, for a generic parameter value θ . Exclude those values of θ for which $Z_0(\theta)$ are among the q largest values of $(Z_i(\theta))_{i=0}^{m-1}$. It can be shown that the true parameters, θ^* , belong to the constructed region with exact probability $1 - q/m$.

Remarks. The least-squares estimate has the property that $S_0(\hat{\theta}_{\text{LS}}) = 0$ and hence it is always in the confidence region (which is in fact an interval). When $\theta \neq \theta^*$ we have that

$$Z_0(\theta) = \left[\sum_{t=1}^n (\theta^* - \theta) U_{t-1}^2 + U_{t-1} N_t \right]^2, \quad (13)$$

$$Z_i(\theta) = \left[\sum_{t=1}^n \alpha_{i,t} (\theta^* - \theta) U_{t-1}^2 + \alpha_{i,t} U_{t-1} N_t \right]^2. \quad (14)$$

Due to the sign-perturbations, $\sum_t \alpha_{i,t} (\theta^* - \theta) U_{t-1}^2$ grows slower than the term $\sum_t (\theta^* - \theta) U_{t-1}^2$ in $Z_0(\theta)$. Hence, for θ sufficiently far away from θ^* , θ will be excluded from the confidence interval since $Z_0(\theta)$ will be the largest.

2.2 General FIR Systems

Now, consider general FIR systems of the form

$$Y_t = \sum_{k=1}^d b_k^* U_{t-k} + N_t = \varphi_t^T \theta^* + N_t, \quad (15)$$

³ Due to the density assumption, the probability that two sums are equal is zero, assuming that all sign-sequences are different. Note that we could also add small tie-breaking variables to ensure this.

for $t = 1, \dots, n$, where $\varphi_t \triangleq [U_{t-1}, \dots, U_{t-d}]^T$, $\theta^* \triangleq [b_1^*, \dots, b_d^*]^T$ and the assumptions are as before.

Our aim is again to construct a non-asymptotic confidence region around the LS estimate that contains the true parameters with a user-chosen *exact* probability.

In order to construct an *exact* non-asymptotic confidence region, we can follow the construction of Section 2.1. However, $S_i(\theta)$, $i = 0, \dots, m-1$, are now *vector-valued*. We could define the sign-perturbed sums similarly as before, but since the φ_t 's are now vectors, we take their internal dependencies into account and define the sums as

$$S_0(\theta) \triangleq \Phi_n^{-\frac{1}{2}} \sum_{t=1}^n \varphi_t (Y_t - \varphi_t^T \theta), \quad (16)$$

$$S_i(\theta) \triangleq \Phi_n^{-\frac{1}{2}} \sum_{t=1}^n \varphi_t \alpha_{i,t} (Y_t - \varphi_t^T \theta), \quad (17)$$

where matrix Φ_n is defined as in (7). We can construct a confidence region even without using *weights* $\Phi_n^{-1/2}$, but, as we will see (cf. Figure 1), they help to shape the region.

To proceed with the construction, we need to *order* the vector-valued sign-perturbed sums above. We achieve this by mapping them to real-numbers by using a norm:

$$Z_i(\theta) \triangleq \|S_i(\theta)\|_2^2. \quad (18)$$

Then, the confidence region containing the least-squares estimate, and the true parameter θ^* with *exact* probability $1 - q/m$, is again obtained by excluding those θ for which $Z_0(\theta)$ are among the q largest values of $(Z_i(\theta))_{i=0}^{m-1}$.

Numerical Example. A natural question to ask is what these confidence regions look like. Figure 1 demonstrates this through a simulation example. It presents 99% confidence regions for parameters of a 2nd order FIR system. The true parameters were $\theta^* = [0.2, 0.3]^T$, and we used an i.i.d. noise sequence of zero mean Gaussian variables with variance one, i.e., standard normal. The inputs were

$$U_{t+1} = \kappa U_t + V_t, \quad (19)$$

where $\kappa = 1/2$, $U_{-1} = 0$ and $(V_t)_t$ was an i.i.d. standard normal process, independent of the noise sequence $(N_t)_t$. The construction was based on $n = 200$ observations, and the design parameters were $m = 100$ and $q = 1$. The constructed region with and without the weights, $\Phi_n^{-1/2}$, is demonstrated. For comparison, the 99% confidence ellipsoid based on the asymptotic theory is also shown.

Figure 1 demonstrates that our non-asymptotic confidence regions are comparable in size with the standard, approximate confidence ellipsoids; and the weighted approach produces confidence regions with similar shape as the ellipsoids obtained with the asymptotic theory.

2.3 First-order ARX Systems

Now, we move onto ARX systems, which require some extra care due to the presence of past outputs in the regression vectors, which makes them *endogenous*.

First, consider the following first order system

$$Y_t \triangleq -a_1^* Y_{t-1} + b_1^* U_{t-1} + N_t, \quad (20)$$

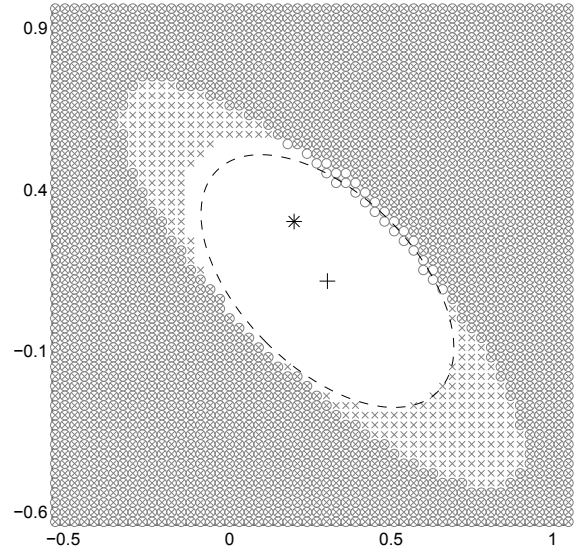


Figure 1. 99% confidence regions for the parameters of a 2nd order FIR system. The (complement of the) constructed region obtained by using weights in (16) and (17) is denoted by “x”. The (complement of the) region built without weights is denoted by “o”. The “*” shows the true parameters, while the “+” is the LS estimate. The dashed ellipsoid is the 99% confidence set obtained with the asymptotic theory.

with $|a_1^*| < 1$. In linear regression form it reads

$$Y_t = \varphi_t^T \theta^* + N_t, \quad (21)$$

where $\varphi_t \triangleq [-Y_{t-1}, U_{t-1}]^T$ and $\theta^* \triangleq [a_1^*, b_1^*]^T$.

The available data are $(Y_t)_{t=0}^n$, $(U_t)_{t=0}^{n-1}$ and we aim at an exact confidence region for θ^* around the LS estimate.

As before, the least-squares estimate is found by solving

$$\sum_{t=1}^n \varphi_t (Y_t - \varphi_t^T \theta) = 0. \quad (22)$$

Corresponding to $\theta = \theta^*$ we have (without the weights)

$$S_0(\theta^*) = \sum_{t=1}^n \varphi_t (Y_t - \varphi_t^T \theta^*) = \sum_{t=1}^n \begin{bmatrix} -Y_{t-1} \\ U_{t-1} \end{bmatrix} N_t. \quad (23)$$

In this case, the direct application of the previous approach will not work. The reason is that $\sum_{t=1}^n -Y_{t-1} N_t$ and $\sum_{t=1}^n -Y_{t-1} \alpha_{i,t} N_t$ will not have the same distribution since Y_{t-1} depends on N_{t-1}, N_{t-2}, \dots . The idea is to make everything “symmetric” again by “constructing” alternative output terms \bar{Y}_{t-1} using sign-perturbed prediction errors such that corresponding to $\theta = \theta^*$, $\sum_{t=1}^n -Y_{t-1} N_t$ and $\sum_{t=1}^n -\bar{Y}_{t-1} \alpha_{i,t} N_t$ have the same distribution.

For a given generic value $\theta = [a_1, b_1]^T$, we can compute the prediction errors (reconstructed noise terms) as

$$\hat{N}_t(\theta) \triangleq Y_t - \varphi_t^T \theta, \quad t = 1, \dots, n. \quad (24)$$

As before, we can sign-perturb these prediction errors obtaining $\alpha_{i,t} \hat{N}_t(\theta)$; \bar{Y}_t can now be “constructed” as

$$\bar{Y}_t(\theta, \alpha_i) \triangleq -a_1 \bar{Y}_{t-1}(\theta, \alpha_i) + b_1 U_{t-1} + \alpha_{i,t} \hat{N}_t(\theta), \quad (25)$$

with initial condition $Y_0(\theta, \alpha_i) = Y_0$. If $\theta = \theta^*$, we have $\hat{N}_t(\theta) = N_t$, and $\bar{Y}_t(\theta^*, \alpha_i)$ can be expressed as

$$\bar{Y}_t(\theta^*, \alpha_i) = (-a_1^*)^t Y_0 + \sum_{k=0}^{t-1} b_1^* (-a_1^*)^k U_{t-1-k} + (26)$$

$$+ \sum_{k=0}^{t-1} (-a_1^*)^k (\alpha_{i,t-k} N_{t-k}). \quad (27)$$

We can now introduce the *sign-perturbed sums*

$$S_0(\theta) \triangleq \Phi_n^{-\frac{1}{2}} \sum_{t=1}^n \varphi_t (Y_t - \varphi_t^T \theta), \quad (28)$$

$$S_i(\theta) \triangleq \bar{\Phi}_n^{-\frac{1}{2}}(\theta, \alpha_i) \sum_{t=1}^n \bar{\varphi}_t(\theta, \alpha_i) \alpha_{i,t} (Y_t - \varphi_t^T \theta), \quad (29)$$

$i = 1, \dots, m-1$, where $\bar{\varphi}_t(\theta, \alpha_i) \triangleq [-\bar{Y}_{t-1}(\theta, \alpha_i), U_{t-1}]^T$ and $\bar{\Phi}_n(\theta, \alpha_i)$ are the (perturbed) covariance estimates

$$\bar{\Phi}_n(\theta, \alpha_i) \triangleq \frac{1}{n} \sum_{t=1}^n \bar{\varphi}_t(\theta, \alpha_i) \bar{\varphi}_t^T(\theta, \alpha_i). \quad (30)$$

Corresponding to θ^* everything is “symmetric”. Indeed

$$S_0(\theta^*) = \bar{\Phi}_n^{-\frac{1}{2}}(\theta^*, \mathbb{1}) \sum_{t=1}^n \begin{bmatrix} -\bar{Y}_{t-1}(\theta^*, \mathbb{1}) \\ U_{t-1} \end{bmatrix} N_t, \quad (31)$$

where $\mathbb{1}$ is the all-one vector. Note that $\bar{\Phi}_n^{-\frac{1}{2}} = \bar{\Phi}_n^{-\frac{1}{2}}(\theta^*, \mathbb{1})$ and $Y_{t-1} = \bar{Y}_{t-1}(\theta^*, \mathbb{1})$, since $\hat{N}_t(\theta^*) = N_t$. We also have

$$S_i(\theta^*) = \bar{\Phi}_n^{-\frac{1}{2}}(\theta^*, \alpha_i) \sum_{t=1}^n \begin{bmatrix} -\bar{Y}_{t-1}(\theta^*, \alpha_i) \\ U_{t-1} \end{bmatrix} (\alpha_{i,t} N_t), \quad (32)$$

in both of the cases we can express Y_t and $\bar{Y}_{t-1}(\theta^*, \alpha_i)$ as

$$Y_{t-1} = (-a_1^*)^t Y_0 + \sum_{k=0}^{t-2} b_1^* (-a_1^*)^k U_{t-1-k} + (33)$$

$$+ \sum_{k=0}^{t-2} (-a_1^*)^k N_{t-k}, \quad (34)$$

$$\bar{Y}_{t-1}(\theta^*, \alpha_i) = (-a_1^*)^t Y_0 + \sum_{k=0}^{t-2} b_1^* (-a_1^*)^k U_{t-1-k} + (35)$$

$$+ \sum_{k=0}^{t-2} (-a_1^*)^k (\alpha_{i,t-k} N_{t-k}). \quad (36)$$

We can observe that in the equation for $S_i(\theta^*)$, $i \neq 0$, every occurrence of N_t has been replaced by $\alpha_{i,t} N_t$, and therefore the “symmetry” has been restored.

We can now proceed as for FIR systems and compute

$$Z_i(\theta) = \|S_i(\theta)\|_2^2. \quad (37)$$

A confidence set can again be constructed by excluding those θ for which $Z_0(\theta)$ are among the q largest values of $(Z_i(\theta))_{i=0}^{m-1}$. The inclusion of the LS estimate is guaranteed and θ^* is in the confidence set with probability $1 - q/m$.

3. GENERAL ARX SYSTEMS

In this section, we provide a more rigorous formulation of the method for general ARX systems. As we will see,

the construction includes additional tie-breaking variables, denoted by $(\varepsilon_i)_i$. We also present a pseudo-code as well as a theorem about the guarantees of the constructed regions.

The data generating system is an ARX system

$$Y_t + a_1^* Y_{t-1} + \dots + a_{n_a}^* Y_{t-n_a} \triangleq b_1^* U_{t-1} + \dots + b_{n_b}^* U_{t-n_b} + N_t,$$

which can be written in linear regression form as

$$Y_t = \varphi_t^T \theta^* + N_t, \quad (38)$$

$$\varphi_t \triangleq [-Y_{t-1}, \dots, -Y_{t-n_a}, U_{t-1}, \dots, U_{t-n_b}]^T, \quad (39)$$

$$\theta^* \triangleq [a_1^*, \dots, a_{n_a}^*, b_1^*, \dots, b_{n_b}^*]^T. \quad (40)$$

Let θ be a generic parameter

$$\theta = [a_1, \dots, a_{n_a}, b_1, \dots, b_{n_b}]^T. \quad (41)$$

We make the following three assumptions:

Assumption 1. The orders n_a and n_b are known.

Assumption 2. $(N_t)_t$ is an *independent* (but not necessarily identically distributed) noise sequence (not observed), where each N_t is *symmetrically* distributed about zero.

Assumption 3. $(U_t)_t$ is an *observed* (but not necessarily chosen) input signal, independent of $(N_t)_t$.

The available data for the construction of the confidence region are $(Y_t)_{t=1-n_a}^n$ and $(U_t)_{t=1-n_b}^{n-1}$.

Construction of the Confidence Regions:

- (1) Using the data, compute the prediction errors (i.e., reconstruct the noise for parameter θ)

$$\hat{N}_t(\theta) \triangleq Y_t - \varphi_t^T \theta, \quad t = 1, \dots, n. \quad (42)$$

- (2) Select an integer m and construct m sequences $(\alpha_i) = (\alpha_{i,t})_{t=1}^n$, $i = 0, \dots, m-1$, with the following properties. Let $\alpha_{0,1}, \dots, \alpha_{0,N} = 1, \dots, 1$ be the sequence of all 1s. Every element of the other sequences takes the values -1 or 1 with probability $1/2$ each. Each such sequence is i.i.d. and their elements are also independent of the other variables in the system.

Build sequences of sign-perturbed prediction errors

$$(\alpha_{i,t} \hat{N}_t(\theta))_{t=1}^n, \quad i = 0, \dots, m-1. \quad (43)$$

Use the sign-perturbed prediction errors to construct perturbed alternatives of the outputs

$$\bar{Y}_t(\theta, \alpha_i) = -a_1 \bar{Y}_{t-1}(\theta, \alpha_i) - \dots - a_{n_a} \bar{Y}_{t-n_a}(\theta, \alpha_i) + b_1 U_{t-1} + \dots + b_{n_b} U_{t-n_b} + \alpha_{i,t} \hat{N}_t(\theta), \quad (44)$$

with using the initial conditions $\bar{Y}_t(\theta, \alpha_i) = Y_t$, for $t = 1 - n_a, \dots, 0$; and form the regressors

$$\bar{\varphi}_t(\theta, \alpha_i) = [-\bar{Y}_{t-1}(\theta, \alpha_i), \dots, -\bar{Y}_{t-n_a}(\theta, \alpha_i), \quad (45)$$

$$U_{t-1}, \dots, U_{t-n_b}]^T, \quad i = 0, \dots, m-1. \quad (46)$$

- (3) Compute the following functions

$$Z_i(\theta) \triangleq \|S_i(\theta)\|_2^2 + \varepsilon_i, \quad (47)$$

for $i = 1, \dots, m-1$, where $S_i(\theta)$ is as in (28)-(29), while $(\varepsilon_i)_{i=0}^{m-1}$ are “small” random variables, introduced to break ties. They are conditionally i.i.d. and continuous given $\sigma\{(U_t)_t, (\pm \hat{N}_t(\theta))_t\}$ (cf. Appendix).

- (4) Let $R_0(\theta)$ be the rank of $Z_0(\theta)$ in the ordering of $(Z_i(\theta))_{i=0}^{m-1}$ i.e., $R_0(\theta)$ denotes the number of $Z_i(\theta)$'s that are larger than $Z_0(\theta)$. The confidence region, Θ_m^q , for the true parameters θ^* is constructed as follows

$$\Theta_m^q \triangleq \{ \theta \in \mathbb{R}^d : R_0(\theta) \geq q \}. \quad (48)$$

In other words, Θ_m^q is the region in parameter space where there are at least q functions larger than $Z_0(\theta)$.

This confidence region has the following property

Theorem 1. Under Assumptions 1-3, the probability that θ^* is in the confidence region Θ_m^q is exactly $1 - q/m$.

A sketch of the proof is given in the Appendix.

Remarks. The idea behind the construction is that, corresponding to θ^* , $Z_0(\theta^*)$ will only be among the largest $Z_i(\theta^*)$ functions with a small probability, and hence the values of θ for which this happens are excluded. On the other hand, $Z_0(\theta)$ grows faster than the other $Z_i(\theta)$, $i \neq 0$, functions as $\|\theta - \theta^*\|_2^2$ gets larger, and thus values different from θ^* will eventually be excluded from the region.

The parameters q and m are user-chosen, so the probability $1 - q/m$ is under complete control of the user.

The least-squares estimate, $\hat{\theta}_{LS}$, is obtained from

$$\sum_{t=1}^n \varphi_t(Y_t - \varphi_t^T \theta) = 0, \quad (49)$$

and hence $\hat{\theta}_{LS}$ has the property that $S_0(\hat{\theta}_{LS}) = 0$. Thus, it is always in the constructed region, assuming, e.g., that each α_i sequence is different and the noises are continuous.

4. CONCLUDING REMARKS

In this paper, a new system identification approach, called *Sign-Perturbed Sums* (SPS), has been proposed that allows the construction of guaranteed *non-asymptotic* confidence regions for the parameters of dynamical systems under mild statistical assumptions. The constructed confidence regions always contain the the *least-squares* estimate and the true parameters with *exact* and user-chosen probabilities, therefore, no conservatism is introduced.

The theoretical tool behind our approach is that certain random variables are “uniformly ordered”, which allows us to perturb the signs of the reconstructed noise sequences and build confidence regions by ranking the rerandomized sums of some carefully selected functions. Depending on how these functions are selected, we can arrive at confidence regions for different identification methods. This fact has been demonstrated for FIR and ARX systems. Moreover, a numerical experiment has shown that the constructed region is comparable in size and shape with the standard, asymptotic confidence ellipsoid.

REFERENCES

- M.C. Campi and E. Weyer. Guaranteed non-asymptotic confidence regions in system identification. *Automatica*, 41:1751–1764, 2005.
- M.C. Campi and E. Weyer. Non-asymptotic confidence sets for the parameters of linear transfer functions. *IEEE Trans. on Automatic Control*, 55:2708–2720, 2010.

- M.C. Campi, S. Ko, and E. Weyer. Non-asymptotic confidence regions for model parameters in the presence of unmodelled dynamics. *Automatica*, 45:2175–2186, 2009.
- M. Dalai, E. Weyer, and M.C. Campi. Parameter identification for nonlinear systems: Guaranteed confidence. *Automatica*, 43:1418–1425, 2007.
- O. Kallenberg. *Foundations of Modern Probability*. Springer, 2nd edition, 2001. ISBN 978-0-387-95313-7.
- M. Ledoux and M. Talagrand. *Probability in Banach Spaces*. Springer, 1991. ISBN 3-540-52013-9.
- L. Ljung. *System Identification: Theory for the User*. Prentice-Hall, Upper Saddle River, 2nd edition, 1999.
- T. Söderström and P. Stoica. *System Identification*. Prentice Hall, 1989. ISBN 0-13-881236-5.

APPENDIX

In this appendix, we provide definitions and background information as well as a sketch of the proof of Theorem 1.

Symmetric Variables

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a *probability space*, where Ω is the sample space, \mathcal{F} is the σ -algebra of events and \mathbb{P} is the probability measure. We call an $(\mathbb{R}^d$ -valued) random variable X *symmetric* (about the origin, the zero vector), if

$$\forall A \in \mathcal{F} : \mathbb{P}(X \in A) = \mathbb{P}(-X \in A). \quad (50)$$

Many standard distributions, e.g., Gaussian, Laplace, Cauchy-Lorentz, Bernoulli, Binomial, Students t and uniform are symmetric (e.g., assuming zero mean).

If X_1, \dots, X_n are $(\mathbb{R}^d$ -valued) random variables, we introduce their *sign-independent* σ -algebra as follows

$$\sigma\{\pm X_1, \dots, \pm X_n\} \triangleq \{A \cup -A : A \in \sigma\{X_1, \dots, X_n\}\},$$

where $\sigma\{X_1, \dots, X_n\}$ is the σ -algebra generated by random variables X_1, \dots, X_n . Note that $\sigma\{\pm X\}$ gives us all information about X up to a ± 1 multiplication. It is easy to see that if X is real-valued, then $\sigma\{\pm X\} = \sigma\{|X|\}$.

Note that $(\varepsilon_i)_{i=0}^{m-1}$ in (47) is only needed to resolve ties. It is sufficient that $(\varepsilon_i)_{i=0}^{m-1}$ are i.i.d. *conditionally* on the available data, more precisely, given $\sigma\{(U_i)_{i=1}^n, (\pm \hat{N}_i(\theta))_{i=1}^n\}$. In this way, we can make sure that they only break ties and do not change the order of $Z_i(\theta)$'s inadvertently.

Random Signs

We define *random signs* as symmetric ± 1 valued Bernoulli variables: they take $+1$ and -1 with probability $1/2$ each.

A characterization of symmetric random variables can be given as: X is symmetric if and only if it has the same distribution as $\beta \cdot X$, where β is a random sign that is independent of X (see Ledoux and Talagrand [1991]).

In order to prove the confidence probability, we will need the following definition of uniform ordering.

Definition 1. A finite sequence of real-valued random variables Z_0, \dots, Z_{m-1} is said to be “uniformly ordered” if, for all permutations i_0, \dots, i_{m-1} of $0, \dots, m-1$, we have that

$$\mathbb{P}(Z_{i_0} < Z_{i_1} < \dots < Z_{i_{m-1}}) = \frac{1}{m!}. \quad (51)$$

Note that if Z_0, \dots, Z_{m-1} are uniformly ordered, then they are almost surely pairwise not-equal. Moreover, Z_i takes position j in the ordering with probability exactly $1/m$.

Now, we state some lemmas (mostly without proofs, due to space limitations) which will be used in the main proof.

Lemma 1. Let X be a symmetric \mathbb{R}^d -valued random variable and let β be a random sign, independent of X . Then, β and βX are independent and, of course, $X = \beta \cdot (\beta X)$.

Lemma 2. Let $\alpha, \beta_1, \dots, \beta_k$ be i.i.d. random signs. Then, random variables $\alpha, \alpha \cdot \beta_1, \dots, \alpha \cdot \beta_k$ are i.i.d. random signs.

The next lemma shows that probabilities can be calculated by considering all realizations, if they are unanimous.

Lemma 3. Let X and Y be two independent, \mathbb{R}^d -valued and \mathbb{R}^k -valued random variables, respectively. Let us consider a measurable function $g : \mathbb{R}^d \times \mathbb{R}^k \rightarrow \mathbb{R}$ and a measurable set $A \subseteq \mathbb{R}$. Then, if for all $x \in \mathbb{R}^d$ we have $\mathbb{P}(g(x, Y) \in A) = p$, we also have $\mathbb{P}(g(X, Y) \in A) = p$.

Uniform Ordering Lemma

Lemma 4. Let Z_0, \dots, Z_{m-1} be real-valued, i.i.d., continuous random variables. Then, they are uniformly ordered.

Proof. Let us consider a function g that gives an ascending order of Z_0, \dots, Z_{m-1} , i.e., it provides a permutation i_0, i_1, \dots, i_{m-1} , such that $Z_{i_0} \leq Z_{i_1} \leq \dots \leq Z_{i_{m-1}}$. Then, function $g(Z_0, \dots, Z_{m-1})$ can be treated as a discrete random variable with $m!$ possible outcomes, where an outcome represent a particular ordering of the Z_i 's. Because g depends only on the ordering of the variables, if z_0, \dots, z_{m-1} is a realization, where the z_i 's are pairwise not equal, which has probability one, $g(z_0, \dots, z_{m-1}) \neq g(z_{i_0}, \dots, z_{i_{m-1}})$ if $(0, \dots, m-1) \neq (i_0, \dots, i_{m-1})$, where (i_0, \dots, i_{m-1}) is a permutation of $(0, \dots, m-1)$. Moreover, $g(z_{i_0}, \dots, z_{i_{m-1}}) = g(z_{j_0}, \dots, z_{j_{m-1}})$ if and only if $(i_0, \dots, i_{m-1}) = (j_0, \dots, j_{m-1})$. But, since Z_0, \dots, Z_{m-1} are i.i.d., the distribution of $g(Z_{i_0}, \dots, Z_{i_{m-1}})$ is the same for all permutations. There are $m!$ permutations, thus, each of its outcomes must have probability $1/(m!)$. \square

Proof Sketch of Theorem 1

By its construction, the confidence region Θ_m^q contains the true parameters θ^* if and only if $R_0(\theta^*) \geq q$. It means that $Z_0(\theta^*)$ should take one of the positions $0, \dots, m-q-1$ in the ordering of $(Z_i(\theta^*))_{i=0}^{m-1}$. We will prove that the $Z_i(\theta^*)$'s are *uniformly ordered*, which means that $Z_0(\theta^*)$ takes each position in the ordering with probability $1/m$, hence, its rank will be at least q with probability $1-q/m$. To show that the $Z_i(\theta^*)$'s are uniformly ordered, we start by fixing an arbitrary realization of the inputs, $(u_i)_{i=1}^n$, and henceforth we will condition on this realization.

Assuming the (conditional) uniform ordering of $Z_i(\theta^*)$, we have that $\mathbb{P}(\theta^* \in \Theta_m^q) = 1-q/m$, given $(u_i)_{i=1}^n$. Since, this result is independent of the realization, Lemma 3 shows that it also holds without fixing the realization.

To complete the proof, we have to show that, after the realization $(u_i)_{i=1}^n$ was fixed, $Z_i(\theta^*)$ are uniformly ordered.

Let $g(\cdot)$ be the (measurable) function defined as

$$g(\tilde{x}) \triangleq \left\| \bar{\Phi}_n^{-\frac{1}{2}}(\tilde{x}) \sum_{t=1}^n \bar{\varphi}_t(\tilde{x}) x_t \right\|_2^2, \quad (52)$$

where $\tilde{x} = [x_1, \dots, x_n]^T$, and $\bar{\varphi}_t(\tilde{x})$, $\bar{\Phi}_n^{-\frac{1}{2}}(\tilde{x})$ are the "constructed" regressors and covariance estimates using θ^* and applying noise term $N_t = x_t$ as well as the fixed inputs $U_t = u_t$, for all t . Thus, $\bar{\varphi}_t(N_1, \dots, N_n) = \varphi_t$.

Then, we can write variables $(Z_i(\theta^*))_{i=0}^{m-1}$ in the form

$$Z_0 \triangleq Z_0(\theta^*) = g(N_1, \dots, N_n) + \varepsilon_0, \quad (53)$$

$$Z_i \triangleq Z_i(\theta^*) = g(\alpha_{i,1}N_1, \dots, \alpha_{i,n}N_n) + \varepsilon_i, \quad (54)$$

where $i \in \{1, \dots, m-1\}$ and $(\alpha_{i,t})_{i,t=1,1}^{m-1,n}$ are as before.

Since $(N_t)_{t=1}^n$ are symmetric, using Lemma 1, we have

$$N_t = \beta_t(\beta_t N_t) = \beta_t W_t, \quad (55)$$

for all $t \in \{1, \dots, n\}$, where $W_t \triangleq \beta_t N_t$ and $(\beta_t)_{t=1}^n$ are random signs independent of $(N_t)_{t=1}^n$ and, as it was shown by Lemma 1, also independent of $(V_t)_{t=1}^n$. Then,

$$Z_0 = g(\beta_1 V_1, \dots, \beta_n V_n) + \varepsilon_0, \quad (56)$$

$$Z_i = g(\alpha_{i,1} \beta_1 V_1, \dots, \alpha_{i,n} \beta_n V_n) + \varepsilon_i, \quad (57)$$

which can be written in a unified form as

$$Z_i = g(\gamma_{i1} V_1, \dots, \gamma_{in} V_n) + \varepsilon_i, \quad (58)$$

where, for all t , $\gamma_{0t} \triangleq \beta_t$ and, for all $i \neq 0$, $\gamma_{it} \triangleq \alpha_{i,t} \beta_t$.

Now, as shown by Lemma 2, $(\gamma_{it})_{i,t=0,1}^{m-1,n}$ is an i.i.d. collection of random signs, it is also independent of $(V_t)_{t=1}^n$.

By fixing a realization of $(V_t)_{t=1}^n$, called $(v_t)_{t=1}^n$, we have

$$Z'_i \triangleq g(\gamma_{i1} v_1, \dots, \gamma_{in} v_n) + \varepsilon_i, \quad (59)$$

where $(v_t)_{t=1}^n$ are deterministic constants. We continue our investigation by conditioning on this fixed realization.

Random variables $(g(\gamma_{i1} v_1, \dots, \gamma_{in} v_n))_{i=0}^{m-1}$ are i.i.d., since $(\gamma_{it})_{i,t=0,1}^{m-1,n}$ are i.i.d. Moreover, since $(\varepsilon_i)_{i=0}^{m-1}$ are conditionally i.i.d. and continuous, given $\sigma\{(\pm N_i)_{i=1}^n\}$, random variables $(Z'_i)_{i=0}^{m-1}$ are real-valued, i.i.d. and continuous (note that constants $(v_t)_{t=1}^n$ only provide information about the realization of the noise sequence $(N_t)_{t=1}^n$ up to ± 1 multiplications). Therefore, Lemma 4 can be applied to show that variables $(Z'_i)_{i=0}^{m-1}$ are uniformly ordered.

Since this uniform ordering can be obtained *independently of the realization* of the sequence $V \triangleq (V_t)_{t=1}^n$, the statement of the theorem follows. This last step can be made more precise, as follows. First, let us decompose $\varepsilon \triangleq (\varepsilon_i)_{i=0}^{m-1}$ to the form of $\varepsilon = r(V, \delta)$, where r is a measurable function and δ is a real-valued random variable, uniform on $(0, 1)$, independent of V . The validity of this decomposition is supported by Theorem 6.10 of Kallenberg [2001]. Then, let $Q \triangleq (\gamma, \delta)$, where $\gamma \triangleq (\gamma_{it})_{i,t=0,1}^{m-1,n}$. As we have shown above, if we fix a realization of V , then the probability of a particular ordering of $(Z'_i)_{i=0}^{m-1}$ is $1/(m!)$ independently of the realization. By letting $p(Z_0, \dots, Z_{m-1})$ denote the function that provides the ordering of the variables (cf. Lemma 4), we can write this function as $\hat{p}(V, Q)$, because the ordering only depends on V and Q . Finally, since V and Q are independent, we can apply Lemma 3 to show that the result also holds without fixing the realization. \square