51st IEEE Conference on Decision and Control
December 10-13, 2012. Maui, Hawaii, USA

Sign-Perturbed Sums (SPS): A Method for Constructing Exact Finite-Sample Confidence Regions for General Linear Systems

Balázs Csanád Csáji[†‡] Marco C. Campi[*] Erik Weyer[†]

Abstract— **In this paper we propose an algorithm for constructing *non-asymptotic* confidence regions for parameters of *general* linear systems under mild statistical assumptions. The constructed regions are centered around the *prediction error estimate* and are guaranteed to contain the "true" parameter with a user-chosen *exact* probability. Our main assumption is that the noise terms are independent and *symmetrically* distributed about zero, but they do not have to be stationary, nor do their variances and distributions have to be known. The construction of the region is based on the uniform ordering property of some carefully selected *sign-perturbed sums* (SPS) which, as we prove, rigorously guarantees the confidence probability for every finite dataset. The paper also investigates *weighted* estimates and presents a simulation example on an ARMA process that compares our exact confidence regions with the approximate ones based on the asymptotic theory.**

I. INTRODUCTION

One of the core problems of system identification is how to estimate parameters of dynamical systems from noisy measurements [7], [9]. Standard 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 many practical applications, especially in those that involve strong safety, stability or quality constraints, having guaranteed confidence regions, in addition to standard point estimates, is strongly desirable. However, generally the noise characteristics are only *partially known* and the noise may as well have changing intensity through time, i.e., it can be *nonstationary*. Furthermore, in practice we have only a *finite* dataset available. These features make the standard methods inapplicable to deliver *rigorously guaranteed* confidence regions.

A predecessor of the approach presented in this paper is the "Leave-out Sign-dominant Correlation Regions" (LSCR) method, which was developed in [1], [2], [3], [5]. LSCR is a finite-sample system identification algorithm that can build non-asymptotic confidence regions for parameters of various (linear and non-linear) dynamical systems under weak assumptions on the noise. An important theoretical property of LSCR 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 probability level. However, (i) LSCR does not provide confidence regions with *exact* probabilities when more than one parameter is being estimated, moreover, (ii) it does not guarantee the inclusion of a chosen *nominal* (for example, least-squares) estimate.

This paper extends our earlier work on the *Sign-Perturbed Sums* (SPS) algorithm that guarantees non-asymptotic confidence regions around the least-squares estimate for FIR and ARX systems [4]. Here, we generalize our earlier results in two directions: (i) we generalize the method for *general linear systems* such that the constructed finite-sample confidence regions will be centered around the estimates of the *prediction error method* (PEM). Moreover, (ii) we allow *weighting* the measurements depending on, for example, their reliability.

First, we revisit the SPS algorithm for ARX systems to recall some core ideas developed in [4]. Then, we extend SPS to general linear systems and PEM estimates, and also investigate the case when different measurements can have different weights. Finally, we present a simulation example on an ARMA(1,1) process that illustrates the constructed confidence regions. In the appendix, we provide a sketch of the proof of the main theorem, which states that the "true" parameter will be in the confidence region with a user-chosen exact probability.

II. PRELIMINARIES: ARX SYSTEMS

We start by presenting the SPS method for ARX systems, which allows us to demonstrate the main ideas of the method.

A. Problem Setting

Consider the following general SISO ARX system

$$Y_t + \sum_{j=1}^{n_a} a_j^* Y_{t-j} \;\triangleq\; \sum_{j=1}^{n_b} b_j^* U_{t-j} + N_t, \tag{1}$$

where Y_t is the output, U_t is the input and N_t is the noise affecting the system at time t. Y_t, U_t and N_t are real-valued scalars. We assume that the inputs are observed and the orders n_a and n_b are known. Regarding the noise, we only assume that $(N_t)_{t=1}^n$ is a sequence of independent random variables, which are also independent of the input, symmetrically distributed about zero and have densities[1]. No other assumptions are imposed, e.g., the noise can be time-varying with unknown (but symmetric) distributions.

[†]Department of Electrical and Electronic Engineering, Melbourne School of Engineering, The University of Melbourne, 240 Grattan Street, Parkville, Melbourne, Victoria, 3010, Australia, email: bcsaji@unimelb.edu.au and ewey@unimelb.edu.au

[‡]Computer and Automation Research Institute (SZTAKI), Hungarian Academy of Sciences, Kende utca 13–17, Budapest, Hungary, H-1111

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

[1]The density assumption is introduced for simplicity and can be replaced by more general assumptions.

The available data are $(Y_t)_{t=1-n_a}^{n}$ and $(U_t)_{t=1-n_b}^{n-1}$. The goal is to construct a confidence region around the *least-squares* (LS) estimate that is guaranteed to contain the "true" parameter vector θ^* with a user-chosen probability.

B. Least-Squares Estimate

The system (1) can be written in a linear regression form

$$Y_t \,=\, \varphi_t^{\mathrm{T}} \theta^* + N_t, \tag{2}$$

where the regressor φ_t and the parameter θ^* is defined as

$$\varphi_t \,\triangleq\, (-Y_{t-1}, \dots, -Y_{t-n_a}, U_{t-1}, \dots, U_{t-n_b})^{\mathrm{T}}, \tag{3}$$
$$\theta^* \,\triangleq\, \left(a_1^*, \dots, a_{n_a}^*, b_1^*, \dots, b_{n_b}^*\right)^{\mathrm{T}}. \tag{4}$$

A generic parameter $\theta \in \mathbb{R}^d$, $d \triangleq n_a + n_b$, is denoted by

$$\theta \,=\, (a_1, \dots, a_{n_a}, b_1, \dots, b_{n_b})^{\mathrm{T}}. \tag{5}$$

For a given θ and t, the *predictor error* at time t is

$$\hat{N}_t(\theta) \,\triangleq\, Y_t - \varphi_t^{\mathrm{T}} \theta. \tag{6}$$

The least-squares estimate for (2) can be found by minimizing the sum of the squared prediction errors,

$$\hat{\theta}_{\mathrm{LS}} \,\triangleq\, \operatorname*{arg\,min}_{\theta \in \mathbb{R}^d} \sum_{t=1}^{n} \hat{N}_t^2(\theta) \,=\, \operatorname*{arg\,min}_{\theta \in \mathbb{R}^d} \sum_{t=1}^{n} \left(Y_t - \varphi_t^{\mathrm{T}} \theta \right)^2,$$

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

$$\sum_{t=1}^{n} \varphi_t \hat{N}_t(\hat{\theta}_{\mathrm{LS}}) \,=\, \sum_{t=1}^{n} \varphi_t (Y_t - \varphi_t^{\mathrm{T}} \hat{\theta}_{\mathrm{LS}}) \,=\, 0. \tag{7}$$

It is well-known [9] that $\hat{\theta}_{\mathrm{LS}}$ can be explicitly written as

$$\hat{\theta}_{\mathrm{LS}} = \left(\frac{1}{n} \sum_{t=1}^{n} \varphi_t \varphi_t^{\mathrm{T}} \right)^{-1} \left(\frac{1}{n} \sum_{t=1}^{n} \varphi_t Y_t \right), \tag{8}$$

assuming the invertability of the matrix $\frac{1}{n}\sum_{t=1}^{n} \varphi_t \varphi_t^{\mathrm{T}}$.

C. Asymptotic Confidence Ellipsoids

It is also known that under some moment conditions on the noise sequence, such as the Lindeberg condition [8], the error of the LS estimates is *asymptotically* normal, that is

$$\sqrt{n}(\hat{\theta}_n - \theta^*) \xrightarrow{\;d\;} N(0, \Gamma) \quad \text{as} \quad n \to \infty, \tag{9}$$

where $\hat{\theta}_n$ is the LS estimate using n data points, $\xrightarrow{\;d\;}$ denotes convergence in distribution and $N(0, \Gamma)$ denotes the normal distribution with zero mean and covariance matrix

$$\Gamma \,\triangleq\, \sigma_0^2 \left(\mathbb{E}\left[\varphi_0 \varphi_0^{\mathrm{T}} \right] \right)^{-1}, \tag{10}$$

with σ_0^2 being the variance of the noise, where the noise and input signals are here assumed to be stationary.

This result allows the construction of *asymptotic* confidence ellipsoids. In fact, from (9) one obtains [7] that

$$n \,\|\, \hat{\theta}_n - \theta^* \,\|_{\Gamma}^2 \xrightarrow{\;d\;} \chi^2(d) \quad \text{as} \quad n \to \infty, \tag{11}$$

where $\|\, \hat{\theta}_n - \theta^* \,\|_{\Gamma}^2 = (\hat{\theta}_n - \theta^*)^{\mathrm{T}}\, \Gamma^{-1} \,(\hat{\theta}_n - \theta^*)$, and $\chi^2(d)$ denotes the χ^2 distribution with $\dim(\theta^*) = d$ degrees of freedom. Matrix Γ is, of course, not known in practice, but it can be estimated from the data. By using estimates for σ_0^2 and $\mathbb{E}\left[\varphi_0 \varphi_0^{\mathrm{T}}\right]$, a confidence region is obtained as

$$\Theta_n^{\alpha} \,\triangleq\, \left\{\, \theta \in \mathbb{R}^d \,:\, \|\, \theta - \hat{\theta}_n \,\|_{\Phi_n}^2 \,\leq\, \alpha\, \hat{\sigma}_n / n \,\right\}, \tag{12}$$

where the probability that θ^* is not in Θ_n^{α} can be computed as the α-level of the $\chi^2(d)$ distribution, and where

$$\Phi_n \,\triangleq\, \frac{1}{n} \sum_{t=1}^{n} \varphi_t \varphi_t^{\mathrm{T}}, \quad \text{and} \quad \hat{\sigma}_n \,\triangleq\, \frac{1}{n} \sum_{t=1}^{n} \hat{N}_t^2(\hat{\theta}_n). \tag{13}$$

Here we shall make two observations: (i) the confidence region Θ_n^{α} is *stochastic*, since it depends on $\hat{\theta}_n$, $\hat{\sigma}_n$ and Φ_n, which are random elements. Moreover, (ii) region Θ_n^{α} is only an approximated confidence region, and it does not have rigorous guarantees. As Monte Carlo simulation studies show [7], this approach is usually not applicable to small samples.

D. Non-Asymptotic Confidence Regions

There are several applications in which it is desirable to have *rigorously guaranteed* confidence regions even for *finite*, possibly small, datasets. In this section we present the Sign-Perturbed Sums (SPS) method to build *non-asymptotic* confidence regions for the LS estimate. The constructed confidence regions are guaranteed to contain the "true" parameter with a user-chosen *exact* probability, hence no conservativism is introduced. In addition, only mild assumptions on the noise terms are made: the noise terms are *independent* of each other (and of the inputs), and *symmetrically* distributed about zero. Note that each noise term can have different distribution and no knowledge about the particular distributions are assumed.

Now, we present the pseudo-code of the algorithm that decides whether a given θ is included in a p-level confidence region. Probability p is a user-chosen parameter.

PSEUDO-CODE: IS-INCLUDED (θ, p)

1) Given p, which is assumed to be a rational number, set integers $m \geq 1$ and $q \geq 0$ such that $p = 1 - q/m$;
2) Compute
$$\hat{N}_t(\theta) \,\triangleq\, Y_t - \varphi_t^{\mathrm{T}} \theta, \quad 1 \leq t \leq n; \tag{14}$$
3) Generate $n \cdot (m-1)$ independent random signs with
$$\mathbb{P}(\alpha_{it} = 1) \,=\, \mathbb{P}(\alpha_{it} = -1) \,=\, \frac{1}{2}, \tag{15}$$
where $1 \leq i \leq m-1$ and $1 \leq t \leq n$;
4) Create $m-1$ sequences of sign-perturbed noise terms:
$$(\alpha_{it}\, \hat{N}_t(\theta))_{t=1}^{n}, \tag{16}$$
using the prediction errors, where $1 \leq i \leq m-1$;
5) Use the sign-perturbed prediction errors to construct perturbed version of the outputs:
$$\bar{Y}_t(\theta, \alpha_i) \,\triangleq\, -\sum_{j=1}^{n_a} a_j \bar{Y}_{t-j}(\theta, \alpha_i) +$$

$$+ \sum_{j=1}^{n_b} b_j U_{t-j} + \alpha_{it} \hat{N}_t(\theta), \quad (17)$$

where $1 \le i \le m-1$ and $1 \le t \le n$, using the initial conditions $\bar{Y}_t(\theta, \alpha_i) \triangleq Y_t$, for $1 - n_a \le t \le 0$;

6) Construct the perturbed version of the regressors:

$$\bar{\varphi}_t(\theta, \alpha_i) \triangleq \quad (18)$$
$$(-\bar{Y}_{t-1}(\theta, \alpha_i), \dots, -\bar{Y}_{t-n_a}(\theta, \alpha_i), U_{t-1}, \dots, U_{t-n_b})^{\mathrm{T}},$$

where $1 \le i \le m-1$ and $1 \le t \le n$;

7) Compute the perturbed covariance estimates:

$$\bar{\Phi}_n(\theta, \alpha_i) \triangleq \frac{1}{n} \sum_{t=1}^{n} \bar{\varphi}_t(\theta, \alpha_i) \bar{\varphi}_t^{\mathrm{T}}(\theta, \alpha_i), \quad (19)$$

where $1 \le i \le m-1$ (assume, they have full rank);

8) Evaluate the sign-perturbed sums below at parameter θ:

$$S_0(\theta) \triangleq \Phi_n^{-\frac{1}{2}} \sum_{t=1}^{n} \varphi_t \hat{N}_t(\theta), \quad (20)$$

$$S_i(\theta) \triangleq \bar{\Phi}_n^{-\frac{1}{2}}(\theta, \alpha_i) \sum_{t=1}^{n} \alpha_{it} \, \bar{\varphi}_t(\theta, \alpha_i) \hat{N}_t(\theta), \quad (21)$$

where $0 \le i \le m-1$ and Φ_n is defined as in (13);

9) Generate $(\varepsilon_i)_{i=0}^{m-1}$ "small" continuous random variables (they are included used to break possible ties);

10) Compute

$$Z_i(\theta) \triangleq \|S_i(\theta)\|_2^2 + \varepsilon_i, \quad (22)$$

where $0 \le i \le m-1$ and $\|\cdot\|_2$ is the Euclidean norm;

11) Order scalars $(Z_i(\theta))_{i=0}^{m-1}$ in a descending order and let $R_0(\theta)$ be the rank of $Z_0(\theta)$, i.e., the number of those $Z_i(\theta)$'s that are *larger* than $Z_0(\theta)$;

12) Return "true" if $R_0(\theta) \ge q$, otherwise return "false";

Applying the method above, we can construct a random *confidence region* for the "true" parameter θ^*, as follows

$$\Theta_m^q \triangleq \left\{ \theta \in \mathbb{R}^d \, : \, R_0(\theta) \ge q \right\}. \quad (23)$$

Furthermore, the probability that $\theta^* \in \Theta_m^q$ is

$$\mathbb{P}\left(\theta^* \in \Theta_m^q\right) = \mathbb{P}\left(R_0(\theta^*) \ge q\right) = 1 - \frac{q}{m} = p, \quad (24)$$

note that it is an *exact* probability (cf., Theorem 1).

Therefore, for any given (rational) probability $p \in (0,1)$, we can construct a confidence region that contains the unknown "true" parameter θ^* with *exact* probability p.

E. How does this work?

In this section we provide some intuition behind the construction. The related theoretical result, Theorem 1, is stated in Section III-D.

Before we continue, we formalize the definition of *uniform ordering*, since it is one of the key concepts of our method.

Definition 1: A finite sequence of real-valued random variables $Z_0, \dots, Z_{m-1}$ is called "uniformly ordered", if, for all possible permutations $i_0, \dots, i_{m-1}$ of their indexes, we have that $\mathbb{P}(Z_{i_0} < Z_{i_1} < \cdots < Z_{i_{m-1}}) = 1/(m!)$.

Note that if $Z_0, \dots, Z_{m-1}$ are uniformly ordered, this implies that they are also almost surely pairwise non-equal; and, for all i, j, Z_i takes position j with probability $1/m$.

Observe that

$$S_i(\theta^*) = \bar{\Phi}_n^{-\frac{1}{2}}(\theta^*, \alpha_i) \sum_{t=1}^{n} \alpha_{it} \, \bar{\varphi}_t(\theta^*, \alpha_{it}) \, N_t, \quad (25)$$

for all $0 \le i \le m-1$, where for $i = 0$ we let $\alpha_{0t} = 1$ for all t. Based on this expression, and since $(N_t)_{t=1}^n$ are independent and symmetric, in the Appendix it is shown that the variables

$$Z_i(\theta^*) \triangleq \|S_i(\theta^*)\|_2^2 + \varepsilon_i, \quad 0 \le i \le m-1, \quad (26)$$

are *uniformly ordered*. This means that $Z_0(\theta^*)$ takes position i in the ordering of variables $(Z_i(\theta^*))_{i=0}^{m-1}$ with probability $1/m$, and this implies that the exact probability that $\theta^* \in \Theta_m^q$ is $p = 1 - q/m$ since $R_0(\theta^*) \ge q$ means that $Z_0(\theta^*)$ takes one of the positions $0, \dots, m-q-1$ in the ordering.

On the other hand, $Z_0(\theta)$ grows faster than the other $Z_i(\theta)$, $i \neq 0$, functions for values of θ away from θ^*, and thus values different from θ^* will eventually be excluded from the confidence region.

Finally, note that $\hat{\theta}_{\mathrm{LS}}$ solves the normal equations (7) so that $S_0(\hat{\theta}_{\mathrm{LS}}) = 0$. This implies that the LS estimate will be included in the confidence region for all p, provided the ε_i's are chosen small enough and $\alpha_i \neq \mathbb{1}$ for $i \neq 0$.

III. General Linear Systems

The SPS method is extended in this section to general linear systems.

A. Problem Setting

Let us consider the following general linear system [7]

$$A(z^{-1}) \, Y_t = \frac{B(z^{-1})}{F(z^{-1})} \, U_t + \frac{C(z^{-1})}{D(z^{-1})} \, N_t, \quad (27)$$

where U_t is the input, Y_t is the output, N_t is the noise, and A, B, C, D and F are polynomials in z^{-1}, the backward shift operator ($z^{-1} Y_t = Y_{t-1}$). The coefficients of A, B, C, D and F are $(a_k^*)_{k=1}^{n_a}$, $(b_k^*)_{k=1}^{n_b}$, $(c_k^*)_{k=1}^{n_c}$, $(d_k^*)_{k=1}^{n_d}$ and $(f_k^*)_{k=1}^{n_f}$, respectively. We use $\theta^* = (a_1^*, ..., a_{n_a}^*, b_1^*, ..., b_{n_b}^*, c_1^*, ..., c_{n_c}^*, d_1^*, ..., d_{n_d}^*, f_1^*, ..., f_{n_f}^*)^{\mathrm{T}}$.

This system can be written as

$$Y_t \triangleq G(z^{-1}; \theta^*) \, U_t + H(z^{-1}; \theta^*) \, N_t, \quad (28)$$

where $G(z^{-1}; \theta^*)$ and $H(z^{-1}; \theta^*)$ are rational transfer functions. We make the following five assumptions:

Assumption 1: The "true" system that generates the data is in the model class, i.e., it has the form (28). The orders of the polynomials are known;

Assumption 2: The transfer function $H(z^{-1}; \theta^*)$ has a stable inverse. Moreover, $G(0; \theta^*) = 0$ and $H(0; \theta^*) = 1$;

Assumption 3: $(N_t)_{t=1}^n$ is an *independent* (but not necessarily identically distributed) noise sequence (not observed), where each N_t is *symmetrically* distributed about zero;

Assumption 4: $(U_t)_{t=1}^{n-1}$ is an *observed* (but not necessarily chosen) input signal, independent of $(N_t)_{t=1}^n$;

Assumption 5: The system is initialized with $Y_t = N_t = U_t = 0$, $t \leq 0$.

The available data are $(Y_t)_{t=1}^n$ and $(U_t)_{t=1}^{n-1}$, and the goal is to construct a confidence region that includes the prediction error estimate and it is guaranteed to contain the "true" parameter θ^* with a user-chosen *exact* probability.

B. Prediction Error Estimate

For simplicity, here we only consider the *quadratic cost criterion*. The prediction errors can be calculated from relation [9]

$$\hat{N}_t(\theta) \triangleq H^{-1}(z^{-1};\theta)(Y_t - G(z^{-1};\theta)\,U_t). \qquad (29)$$

Note that $\hat{N}_t(\theta^*) = N_t$.

The *prediction error estimate* for (28) is found by minimizing the sum of the squared prediction errors,

$$\hat{\theta}_{\mathrm{PEM}} \triangleq \operatorname*{arg\,min}_{\theta\in\mathcal{M}} \sum_{t=1}^{n} \hat{N}_t^2(\theta),$$

where $\mathcal{M}$ is the class of allowed models.

In general the PEM estimate does not have a closed-form solution. It can be found, e.g., by using the equation

$$\sum_{t=1}^{n} \psi_t(\hat{\theta}_{\mathrm{PEM}})\hat{N}_t(\hat{\theta}_{\mathrm{PEM}}) = 0, \qquad (30)$$

where $\psi_t(\theta)$ is the gradient of the prediction error,

$$\psi_t(\theta) \triangleq \frac{d}{d\theta}\hat{N}_t(\theta). \qquad (31)$$

In case of ARX systems $\psi_t(\theta)$ is simply φ_t, yielding (7).

These gradients can be directly calculated in terms of the defining polynomials. This immediately leads to formulas for most known models (e.g., AR, ARMAX, Box-Jenkins).

The gradients of the prediction errors are [7]

$$\frac{\partial}{\partial a_k}\hat{N}_t(\theta) = \frac{D(z^{-1})}{C(z^{-1})}\,Y_{t-k}, \qquad (32)$$

$$\frac{\partial}{\partial b_k}\hat{N}_t(\theta) = -\frac{D(z^{-1})}{C(z^{-1})F(z^{-1})}\,U_{t-k}, \qquad (33)$$

$$\frac{\partial}{\partial c_k}\hat{N}_t(\theta) = \frac{D(z^{-1})B(z^{-1})}{C(z^{-1})C(z^{-1})F(z^{-1})}\,U_{t-k} - -\frac{D(z^{-1})A(z^{-1})}{C(z^{-1})C(z^{-1})}\,Y_{t-k}, \qquad (34)$$

$$\frac{\partial}{\partial d_k}\hat{N}_t(\theta) = \frac{A(z^{-1})}{C(z^{-1})}\,Y_{t-k} - -\frac{B(z^{-1})}{C(z^{-1})F(z^{-1})}\,U_{t-k}, \qquad (35)$$

$$\frac{\partial}{\partial f_k}\hat{N}_t(\theta) = \frac{D(z^{-1})B(z^{-1})}{C(z^{-1})F(z^{-1})F(z^{-1})}\,U_{t-k}. \qquad (36)$$

C. Sign-Perturbed Sums

In order to extend our confidence region construction, we need to re-define the sign-perturbed sums. We again apply perturbed versions of the outputs that are

$$\bar{Y}_t(\theta,\alpha_i) \triangleq G(z^{-1};\theta)\,U_t + H(z^{-1};\theta)\,(\alpha_{it}\hat{N}_t(\theta)), \qquad (37)$$

where α_{it} are random signs as previously.

As we saw above, $\psi_t(\theta)$ can be treated as a linear filtered version of the outputs and the inputs, that is

$$\psi_t(\theta) = W_0(z^{-1};\theta)\,Y_t + W_1(z^{-1};\theta)\,U_t, \qquad (38)$$

where W_0 and W_1 are vector-valued.

We use them to define perturbed versions of $\psi_t(\theta)$ as

$$\bar{\psi}_t(\theta,\alpha_i) \triangleq W_0(z^{-1};\theta)\,\bar{Y}_t(\theta,\alpha_i) + W_1(z^{-1};\theta)\,U_t, \qquad (39)$$

where the difference is that we filter the perturbed outputs.

Finally, the sign-perturbed sums for θ are defined as

$$S_0(\theta) \triangleq \Psi_n^{-\frac{1}{2}}(\theta)\sum_{t=1}^{n}\psi_t(\theta)\hat{N}_t(\theta), \qquad (40)$$

$$S_i(\theta) \triangleq \bar{\Psi}_n^{-\frac{1}{2}}(\theta,\alpha_i)\sum_{t=1}^{n}\alpha_{it}\,\bar{\psi}_t(\theta,\alpha_i)\hat{N}_t(\theta), \qquad (41)$$

where $0 \leq i \leq m-1$, and Ψ_n and $\bar{\Psi}_n(\theta,\alpha_i)$ are defined as

$$\Psi_n(\theta) \triangleq \sum_{t=1}^{n}\psi_t(\theta)\psi_t^{\mathrm{T}}(\theta), \qquad (42)$$

$$\bar{\Psi}_n(\theta,\alpha_i) \triangleq \sum_{t=1}^{n}\bar{\psi}_t(\theta,\alpha_i)\bar{\psi}_t^{\mathrm{T}}(\theta,\alpha_i), \qquad (43)$$

which is the perturbed version of the covariance estimate. If the model is ARX, these sums are the same as (20) and (21).

Applying the previous method with the new sign-perturbed sums, (40) and (41), we again arrive at a confidence region

$$\Theta_m^q \triangleq \left\{\,\theta\in\mathbb{R}^d : R_0(\theta) \geq q\,\right\}. \qquad (44)$$

Note that for $\theta = \theta^*$ *all* sums $S_i(\cdot)$, $i = 0,\ldots,m-1$, take the form

$$S_i(\theta^*) = \bar{\Psi}_n^{-\frac{1}{2}}(\theta^*,\alpha_i)\sum_{t=1}^{n}\alpha_{it}\,\bar{\psi}_t(\theta^*,\alpha_i)N_t. \qquad (45)$$

Moreover, $S_i(\theta^*)$, $i \neq 0$, can be constructed from $S_0(\theta^*)$ by replacing each occurrence of N_t with $\alpha_{it}N_t$, which has the same distribution as N_t as N_t is symmetric. This ensures the uniform ordering property. Based on this feature the desired confidence probability $p = 1-q/m$ can be rigorously guaranteed (see Theorem 1).

We know that the PEM estimate, $\hat{\theta}_{\mathrm{PEM}}$, satisfies (30), thus we have that $S_0(\hat{\theta}_{\mathrm{PEM}}) = 0$, which guarantees the inclusion of $\hat{\theta}_{\mathrm{PEM}}$ in Θ_m^q, under the same conditions as before.

D. Weighted Measurements

It is often useful to assign different weights to different measurements reflecting the reliability of the data.

In the case that we want to build a confidence region around a weighted nominal estimate, we resize the random signs α_{it} according to the given weights. In fact, this approach works more generally, since symmetric random variables can be applied with great generality. In order to do this, we need to refine the sign-perturbed sums

$$S_0(\theta) \triangleq \Psi_n^{-\frac{1}{2}}(\theta) \sum_{t=1}^{n} |\alpha_{it}| \psi_t(\theta) \hat{N}_t(\theta), \tag{46}$$

$$S_i(\theta) \triangleq \bar{\Psi}_n^{-\frac{1}{2}}(\theta, \bar{\alpha}_i) \sum_{t=1}^{n} \alpha_{it}\, \bar{\psi}_t(\theta, \bar{\alpha}_i) \hat{N}_t(\theta), \tag{47}$$

where $\bar{\alpha}_i$ denotes the sign-vector $(\text{sign}(\alpha_{i1}), \dots, \text{sign}(\alpha_{in}))$ and α_{it} are now arbitrary symmetric random variables with the property that $(\alpha_{it})_{i=0}^{m-1}$ are i.i.d. for any fixed t; however, they can have different distributions for different t indexes. If α_{it} are random signs, we get back (40) and (41).

Our main result, using (46) and (47), can be stated as

Theorem 1: For general linear systems, under Assumptions 1-5, the probability that θ^* is in Θ_m^q is exactly $1-q/m$.

A proof of this theorem can be found in the Appendix.

IV. Simulation Example

In this section we demonstrate the SPS method through a simulation experiment. We consider the ARMA process

$$Y_t + a^* Y_{t-1} = N_t + c^* N_{t-1}, \tag{48}$$

where the "true" parameter is $\theta^* = (a^*,\, c^*)$. The filter of the noise is $C(z^{-1};\theta)\, N_t = N_t + c\, N_{t-1}$. To apply the previous results, we need the inverse of polynomial $C(z^{-1};\theta)$, that is

$$C^{-1}(z^{-1};\theta) = \sum_{k=0}^{\infty} (-1)^k\, c^k\, z^{-k}, \tag{49}$$

which we use to define the prediction errors as

$$\hat{N}_t(\theta) = C^{-1}(z^{-1};\theta)\, (Y_t + a\, Y_{t-1}), \tag{50}$$

where $\theta = (a,\, c)$ is a generic parameter.

The perturbed versions of the outputs are

$$\bar{Y}_t(\theta, \alpha_i) = -a\, \bar{Y}_{t-1}(\theta, \alpha_i) + \alpha_{i,t}\, \hat{N}_t(\theta) + c\, \alpha_{i,t-1} \hat{N}_{t-1}(\theta),$$

for $1 \leq i \leq m-1$ and $1 \leq t \leq n$, where α_{it} are random signs. Finally, we can calculate $\bar{\psi}_t(\theta, \alpha_i)$,

$$\bar{\psi}_t(\theta, \alpha_i) = \begin{bmatrix} C^{-1}(z^{-1};\theta)\, \bar{Y}_{t-1}(\theta, \alpha_i) \\ C^{-2}(z^{-1};\theta) A^{-1}(z^{-1};\theta)\, \alpha_{i,t}\, \bar{Y}_{t-1}(\theta, \alpha_i) \end{bmatrix},$$

which can be used to define functions (40) and (41).

A numerical experiment with the ARMA system (48) is demonstrated in Figure 1. The unknown "true" parameter was $\theta^* = (-0.7,\, 0.3)$, we used $n = 500$ observations, where N_t were i.i.d. zero mean Gaussian random variables with variance one, and we aimed at a $99\,\%$ confidence region around the PEM estimate. Therefore, we set $m = 100$ and $q = 1$. The result of the simulation shows that the SPS confidence region, which has rigorously guaranteed *exact* probabilty, is comparable in size and shape to the confidence ellipsoids of the asymptotic theory.

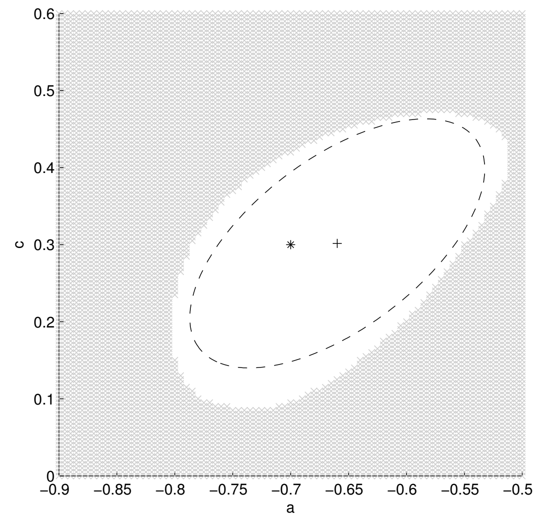

Fig. 1. The $99\,\%$ SPS confidence region (blank part), the "true" parameter "$\star$", the PEM estimate "$+$" with its $99\,\%$ asymptotic confidence ellipsoid (dashed)

V. Concluding Remarks

System identification algorithms with guaranteed finite-sample properties are of high practical importance. In this paper we extended our Sign-Perturbed Sums (SPS) method and showed that it can build confidence regions for general linear systems around the prediction error estimate. The "true" parameter is guaranteed to be included in the region with a user-chosen exact probability for any finite dataset.

Our main assumptions are: (i) The "true" system is in the model class; (ii) The noise sequence is independent and symmetrically distributed about zero.

The SPS method does not require the knowledge of the particular noise distributions, which can even change over time and have unknown variances.

References

[1] M.C. Campi and E. Weyer. Guaranteed non-asymptotic confidence regions in system identification. *Automatica*, 41:1751–1764, 2005.

[2] M.C. Campi and E. Weyer. Non-asymptotic confidence sets for the parameters of linear transfer functions. *IEEE Transasctions on Automatic Control*, 55:2708–2720, 2010.

[3] 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.

[4] B.Cs. Csáji, M.C. Campi, and E. Weyer. Non-asymptotic confidence regions for the least-squares estimate. In *Proceedings of the 16th IFAC Symposium on System Identification*, 2012.

[5] M. Dalai, E. Weyer, and M.C. Campi. Parameter identification for nonlinear systems: guaranteed confidence. *Autom.*, 43:1418–1425, 2007.

[6] O. Kallenberg. *Foundations of Modern Probability*. Springer-Verlag, New York, 2nd edition, 2001.

[7] L. Ljung. *System Identification: Theory for the User*. Prentice-Hall, Upper Saddle River, 2nd edition, 1999.

[8] A. N. Shiryaev. *Probability*. Springer, 2 edition, 1995.

[9] T. Söderström and P. Stoica. *System Identification*. Prentice Hall International, Hertfordshire, UK, 1989.

APPENDIX

Here, we provide a sketch of the proof of Theorem 1.

We say that a random variable X is *symmetric* or *symmetrically distributed about zero*, if X and $-X$ have the same distribution. A special case of symmetric variables is the *random sign*, which is a ± 1 valued Bernoulli random variable, i.e., it takes $+1$ and -1 with probability $1/2$ each.

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

Lemma 1: Let X be a symmetric real-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 $\beta_1, \ldots, \beta_k$ be independent symmetric random variables, α is a random sign, independent of $\beta_1, \ldots, \beta_k$. Then, $\alpha \cdot \beta_1, \ldots, \alpha \cdot \beta_k$ are independent symmetric random variables with the same distribution as $\beta_1, \ldots, \beta_k$.

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

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

A proof of this "uniform ordering lemma" is given in [4].

Proof Sketch of Theorem 1

By definition, parameter θ^* is in the confidence region if $R_0(\theta^*) \geq q$. It means that $Z_0(\theta^*)$ should take one of the positions $0, \ldots, m-q-1$ in the ordering. 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 the $Z_i(\theta^*)$'s, we have that $\mathbb{P}\left(\theta^* \in \Theta_m^q\right) = 1 - q/m$, 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, the $Z_i(\theta^*)$'s are uniformly ordered.

First, we highlight that, for $\theta = \theta^*$, *all* $S_i(\cdot)$ have the form

$$S_i(\theta^*) = \bar{\Psi}_n^{-\frac{1}{2}}(\theta^*, \bar{\xi}_i) \sum_{t=1}^{n} \xi_{it}\, \bar{\psi}_t(\theta^*, \bar{\xi}_i) N_t, \qquad (51)$$

where $\xi_{it} = \alpha_{it}$ if $i \neq 0$ and $\xi_{it} = |\alpha_{it}|$ otherwise, and the notations are the same as in (47). Therefore, all the $S_i(\cdot)$'s depend on the perturbed noises, $(\xi_{it} N_t)_{t=1}^n$, via the *same* function for all i, which we denote by $S_i(\theta^*) = S(\xi_{i1} N_1, \ldots, \xi_{in} N_n)$.

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

$$Z_0 \;\triangleq\; Z_0(\theta^*) = g(|\alpha_{01}| N_1, \ldots, |\alpha_{0n}| N_n) + \varepsilon_0, \quad (52)$$

$$Z_i \;\triangleq\; Z_i(\theta^*) = g(\alpha_{i1} N_1, \ldots, \alpha_{in} N_n) + \varepsilon_i, \qquad (53)$$

where $i \in \{1, \ldots, m-1\}$ and $g(\cdot) = \|S(\cdot)\|_2^2$. Since $(N_t)_{t=1}^n$ are symmetric, using Lemma 1, we have $N_t = \beta_t(\beta_t N_t) = \beta_t V_t$, for all $t \in \{1, \ldots, n\}$, where $V_t \triangleq \beta_t N_t$ 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, for all i,

$$Z_i \;=\; g(\gamma_{i1} V_1, \ldots, \gamma_{in} V_n) + \varepsilon_i, \qquad (54)$$

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

Now, as shown by Lemma 2, $(\gamma_{it})_{i=0}^{m-1}$ are i.i.d. random variables for all t, and they are 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

$$Z_i' \;\triangleq\; g(\gamma_{i1} v_1, \ldots, \gamma_{in} v_n) + \varepsilon_i, \qquad (55)$$

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

Random variables $(g(\gamma_{i1} v_1, \ldots, \gamma_{in} v_n))_{i=0}^{m-1}$ are i.i.d., since $(\gamma_{it})_{i=0}^{m-1}$ are i.i.d., for all t. 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 [6]. 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 $p(Z_0, \ldots, Z_{m-1})$ denote the function that provides the ordering of the variables (viz., takes values from a set with size $m!$), we can write this function as $\widehat{p}(V, Q)$, 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 a particular realization. $\square$