2017 IEEE 56th Annual Conference on Decision and Control (CDC)
December 12-15, 2017, Melbourne, Australia

Finite-Sample System Identification: An Overview and a New Correlation Method

Algo Carè Balázs Csanád Csáji Marco C. Campi Erik Weyer

Abstract— **Finite-sample system identification algorithms can be used to build *guaranteed* confidence regions for unknown model parameters under mild statistical assumptions. It has been shown that in many circumstances these rigorously built regions are comparable in size and shape to those that could be built by resorting to the asymptotic theory. The latter sets are, however, not guaranteed for finite samples and can sometimes lead to misleading results. The general principles behind finite-sample methods make them virtually applicable to a large variety of, even nonlinear, systems. While these principles are simple enough, a rigorous treatment of the attendant technical issues makes the corresponding theory complex and not easy to access. This is believed to be one of the reasons why these methods have not yet received widespread acceptance by the identification community and this paper is meant to provide an easy access point to finite-sample system identification by presenting the fundamental ideas underlying these methods in a simplified manner. We then review three (classes of) methods that have been proposed so far – LSCR (Leave-out Sign-dominant Correlation Regions), SPS (Sign-Perturbed Sums) and PDMs (Perturbed Dataset Methods). By identifying some difficulties inherent in these methods, we also propose in this paper a new sign-perturbation method based on correlation which overcome some of these difficulties.**

I. INTRODUCTION

A fundamental problem in system identification is that of estimating the parameters of partially unknown systems based on noisy observations, [10], [13]. Standard methods in the system identification literature focus on point estimates, that is, they aim at estimating the value of the unknown parameters: classic results guarantee that asymptotically – i.e., when the amount of observations tends to infinity – the parameters can indeed be correctly estimated. However, in general, it is impossible to estimate a parameter with *infinite* precision from a *finite* number of stochastic data, so that a "confidence tag" has to be attached to the point estimate. For this purpose, a confidence region around the estimated parameters is often built. It is well-known that assessing the quality of a non-asymptotic estimate using an asymptotic theory, although popular, may lead to unreliable results, see [7]. On the other hand, making strong assumptions on the probability distribution of the data (e.g., Gaussianity) leads to results that are formally rigorous but of limited practical interest. Motivated by these limitations of standard stochastic[1] identification schemes, non-asymptotic identification methods for building confidence regions that i) are guaranteed when applied to *finite samples* of data and ii) are guaranteed under *minimal assumptions* on the data-generation mechanism have been pursued. The most important examples are the LSCR (Leave-out Sign-dominant Correlation Regions) method [1], the SPS (Sign-Perturbed Sums) method [5] and its generalizations called PDMs (Perturbed Dataset Methods) [9]. These algorithms construct *guaranteed* confidence regions for the unknown model parameters for a large class of dynamical systems, such as general linear systems, [1], [4], and even nonlinear ones [6], under very mild assumptions on the driving noise, or even no assumptions in some specific cases [2]. A difference between LSCR and the latter methods is that regions built by SPS and PDMs contain the true parameter with a probability that is *exact*, while LSCR provides a lower bound in general.

A. Aim of the paper

This paper has two main aims. First, it revisits some crucial ideas in finite-sample system identification and presents them in a unified framework. This is done with the intent of making available to others an easy-to-access point which may foster research in this field. Second, driven by the results highlighted, a new correlation method is proposed which is based on the combination of LSCR and SPS. It builds confidence regions based on correlations, like LSCR, while it applies sign-perturbations with a norm and obtains exact confidence, like SPS. A computational advantage of the new correlation method is that it avoids generating alternative output sequences, which are vital for SPS when handling for example ARX systems. This idea can be easily understood in the light of the unifying approach provided in the paper.

The work of A. Carè was supported by the European Research Consortium for Informatics and Mathematics (ERCIM) and the Australian Research Council (ARC) under Discovery Grant DP130104028. The work of B. Cs. Csáji was supported by the Hung. Sci. Res. Fund (OTKA), pr. no. 113038, the GINOP-2.3.2-15-2016-00002 grant, and by the János Bolyai Research Fellowship, pr. no. BO/00217/16/6. The work of M. C. Campi was partly supported by the University of Brescia under the project H&W "Clafite". Erik Weyer was supported by the ARC Discovery Grant DP130104028.

A. Carè is with the Centrum Wiskunde & Informatica (CWI), Science Park 123, 1098 XG Amsterdam, Netherlands; (email: algocare@gmail.com)

B. Cs. Csáji is with the Institute for Computer Science and Control (SZTAKI), Hungarian Academy of Sciences (MTA), Kende utca 13–17, Budapest, Hungary, 1111; (email: balazs.csaji@sztaki.mta.hu)

M. C. Campi is with the Department of Information Engineering, University of Brescia, Via Branze 38, 25123 Brescia, Italy; (email: marco.campi@unibs.it)

E. Weyer is with Department of Electrical and Electronic Engineering, Melbourne School of Engineering, The University of Melbourne, Melbourne, Victoria, 3010, Australia; (email: ewey@unimelb.edu.au)

[1] Set-membership approaches constitute a different line of research which aims at identifying the region of parameters that are consistent with the observations assuming the noise belongs to some bounded set [11].

B. Structure of the paper

In Section II, the fundamental idea behind finite-sample identification methods based on the *sign-perturbation idea* is revisited and presented in a simplified manner. Then, in Section III, we consider known methods in the light of the framework of Section II, these are LSCR, SPS and PDMs. We show that some of the drawbacks in the existing methods can be overcome by a new, correlation-based approach, which is presented and also applied to a bilinear system in Section IV. Finally, in Section V, we present a brief summary of properties which should be taken into account when finite sample methods are designed or evaluated. Conclusions are drawn in Section VI.

II. Fundamentals of Finite-Sample Identification Methods

We first introduce the goal of exact, finite-sample identification methods, and then describe the *sign-perturbation approach* for building confidence regions. We aim at isolating the main idea and highlight the fundamental principles.

A. Problem set-up

Consider a sample of n *output* measurements $Y_1,\dots,Y_n$. We represent this sequence as a vector $\mathbf{Y}_n = (Y_1, Y_2, ..., Y_n)$. The vector $\mathbf{Y}_n$ depends on the vector $\mathbf{U}_n = (U_1, U_2, ..., U_n)$ of (past) *measured inputs*, on the vector $\mathbf{W}_n = (W_1, W_2, ...W_n)$ of (past) *nonmeasured inputs* (noise), and possibly on some auxiliary set of *initial conditions* $\mathscr{I}$ through a function $\mathbb{F}$,

$$\mathbf{Y}_n \triangleq \mathbb{F}(\mathbf{U}_n, \mathbf{W}_n, \mathscr{I}). \tag{1}$$

Consider now a family of functions $\{\mathbb{F}(\mathbf{U}_n, \mathbf{W}_n, \mathscr{I}; \theta)\}$ parameterized by means of θ and assume that the system function $\mathbb{F}(\mathbf{U}_n, \mathbf{W}_n, \mathscr{I})$ is obtained for one value of θ, say $\theta = \theta^*$.[2] We are interested in constructing methods for building a confidence region $\widehat{\Theta}_n \subseteq \mathbb{R}^d$ that contains the correct θ^* with a user-chosen probability p, namely[3]

$$\mathbb{P}\{\theta^* \in \widehat{\Theta}_n\} = p. \tag{2}$$

Clearly, there is no unique way to build confidence regions so that (2) is satisfied: our goal is presenting well-principled and useful methods.

B. Assumptions

The system is assumed to be invertible w.r.t. the noise:

Assumption 1: For any value of θ, relation $\mathbf{Y}_n \triangleq \mathbb{F}(\mathbf{U}_n, \mathbf{W}_n, \mathscr{I}; \theta)$ is noise invertible in the sense that, given the values of $\mathbf{Y}_n$, $\mathbf{U}_n$, $\mathscr{I}$, vector $\mathbf{W}_n$ can be recovered. ⋆

Example 1: Consider an ARX model

$$Y_t = a_1 Y_{t-1} + \cdots + a_{n_a} Y_{t-n_a} + b_1 U_{t-1} + \cdots + b_{t-n_b} U_{t-n_b} + W_t.$$

Assuming that the given initial conditions, $\mathscr{I}$, contain the terms $U_0,\dots,U_{1-n_b}$ and $Y_0,\dots,Y_{1-n_a}$, the noise vector $\mathbf{W}_n$ can be reconstructed from $\mathbf{Y}_n$ and $\mathbf{U}_n$ by making explicit the ARX equation with respect to the noise term. ⋆

Noise invertibility is a very mild condition. At times, however, one does not know the initial conditions $\mathscr{I}$ so that only part of $\mathbf{W}_n$ can be reconstructed. For instance, in the ARX example not knowing $\mathscr{I}$ prevents the reconstruction of the first terms of $\mathbf{W}_n$. To streamline the presentation, this aspect is glossed over here and we assume that the whole $\mathbf{W}_n$ can be reconstructed; the interested reader is referred to the papers cited in the introduction for more discussion.

In the sequel, the reconstructed noise is indicated with $\widehat{\mathbf{W}}_n(\theta)$, where θ indicates explicitly that the model with parameter θ has been used. Clearly, $\widehat{\mathbf{W}}_n(\theta^*) = \mathbf{W}_n$.

Assumption 2: The noise $\mathbf{W}_n$ is jointly symmetric about zero, i.e., $(W_1,\dots,W_n)$ has the same *joint* probability distribution as $(\sigma_1 W_1,\dots,\sigma_n W_n)$ for all possible sign-sequences, $\sigma_i \in \{+1,-1\}$, $i = 1,\dots,n$. ⋆

Note that in Assumption 2 neither stationarity nor independence is assumed. If the noise sequence is independent, then Assumption 2 is equivalent to say that each noise term W_t has a symmetric probability distribution about zero.

Remark 1 (Beyond the symmetric noise assumption): There are methods in the literature that rely on *no assumptions on the noise*. These methods assume *symmetry of the input* instead, see e.g., [2]. The ideas outlined in this paper can be applied to these methods with minor modifications. For relaxation of the symmetry assumption see also [3] and the references therein.

C. Exact guarantees through sign-perturbation

To simplify notation, given a vector $\mathbf{v}_n = (v_1,\dots,v_n)$ and a vector of signs $\mathbf{s}_n = (\sigma_1,\dots,\sigma_n) \in \{+1,-1\}^n$, we denote the corresponding *sign-perturbed vector* by $\mathbf{s}_n[\mathbf{v}_n] \triangleq (\sigma_1 v_1,\dots,\sigma_n v_n)$.

Consider any function Z that takes as input two vectors of length N and the parameter θ. Example of such functions are given later in the paper. Sign-perturbation methods are based on comparing a reference function defined as

$$Z_0(\theta) \triangleq Z(\mathbf{U}_n, \widehat{\mathbf{W}}_n(\theta), \theta),$$

with $m-1$ "sign-perturbed" functions defined as

$$Z_i(\theta) \triangleq Z(\mathbf{U}_n, \mathbf{s}_n^{(i)}[\widehat{\mathbf{W}}_n(\theta)], \theta),$$

for $i = 1,\dots,m-1$, where $\mathbf{s}_n^{(1)},\dots,\mathbf{s}_n^{(m-1)}$ are $m-1$ user-generated sign vectors of independent random signs, whose elements are $+1$ or -1 with $1/2$ probability each.

Precisely, the construction of the confidence region $\widehat{\Theta}_n$ for θ^* is based on *ranking* $Z_0(\theta)$ with respect to $Z_i(\theta)$, $i = 1,\dots,m-1$. To this goal, one first selects two integers h_1 and h_2 with $h_1 \le h_2$ in the range $1,2,\dots,m$. Then, for any value

[2] This amounts to require that the structure of the system is known while its parameters are not.

[3] In the language of hypothesis testing, p is the probability of type one error, i.e., that the true θ^* is not in the constructed region; the type two error cannot instead be kept under control similarly since a θ that is close enough to θ^* is hard to remove. Instead of enforcing limits on type two errors, in finite-sample system identification one asks that $\widehat{\Theta}_n$ becomes smaller and converges toward θ^* as N increases, see below for more details.

of θ, the numbers $Z_i(\theta)$, $i=0,1,,\ldots,m-1$, are sorted in increasing order. If so happens that $Z_0(\theta)$ is in the position h_1 or h_1+1 or $\ldots$ or h_2, then that θ belongs to $\widehat{\Theta}_n$, in the opposite it does not. For example, say that $m=10$, so that there are 10 functions $Z_i(\theta)$, $i=0,1,\ldots,9$. Take $h_1=1$ and $h_2=3$. For a given θ, if it happens that $Z_0(\theta)$ is the smallest of all functions $Z_i(\theta)$, $i=0,1,\ldots,9$, or the second smallest or the third smallest, then this θ is included in $\widehat{\Theta}_n$, otherwise it is not.[4] Under some additional minor details as hinted at below, the following result holds.

Claim 1: Call $R(\theta)$ the rank of $Z_0(\theta)$ among $\{Z_i(\theta),\, i=0,\ldots,m-1\}$, i.e., if $Z_0(\theta)$ is the smallest, then $R(\theta)=1$, if $Z_0(\theta)$ is the second smallest, then $R(\theta)=2$, and so on. The confidence region defined as

$$\widehat{\Theta}_n \triangleq \{\theta\in\mathbb{R}^d : h_1\le R(\theta)\le h_2\}$$

is such that $\mathbb{P}\{\theta^*\in\widehat{\Theta}_n\} = (h_2-h_1+1)/m$. $\star$

This result is in the form of (2), where $p=(h_2-h_1+1)/m$. Note that h_2-h_1+1 is the number of positions in the ordering that $Z_0(\theta)$ is allowed to take over the total number m of positions. The proof of this result requires some mathematical underpinning to deal with a number of details including the possibility of having ties and possible correlation issues between the system measurable input and the nonmeasurable noise. The exact manner to approach these issues is given in the papers cited in the introduction, while we here only remark that the fundamental idea behind this result is almost straightforward and can be explained as follows. Under the assumption that $\theta=\theta^*$, functions $\{Z_i(\theta^*)\}$ become

$$\begin{aligned} Z_0(\theta^*) &\triangleq Z(\mathbf{U}_n,\mathbf{W}_n,\theta^*),\\ Z_i(\theta^*) &\triangleq Z(\mathbf{U}_n,\mathbf{s}_n^{(i)}[\mathbf{W}_n],\theta^*).\end{aligned}$$

The only difference between these m random variables is that the argument $\mathbf{W}_n$ in the first is replaced by $\mathbf{s}_n^{(i)}[\mathbf{W}_n]$ in the others. However, $\mathbf{W}_n$ and $\mathbf{s}_n^{(i)}[\mathbf{W}_n]$ are random variables having the same distribution because of Assumption 2. Hence, there is no reason why, among the variables $Z_0(\theta^*)$ and $Z_i(\theta^*)$, $i=1,\ldots,m-1$, one should have a larger chance than the others to be in the first or in the second or ... in any other particular position, and in fact each has the same probability $1/m$ to be in any position. Since in Claim 1 $\widehat{\Theta}_n$ is determined by including a given θ if $Z_0(\theta)$ ranks in one among h_2-h_1+1 positions, then θ^* is included with probability $(h_2-h_1+1)/m$. This argument is not rigorous because of tie-breaks and many other minor issues, but the fundamental idea that has been explained here goes through.

Clearly, Claim 1 is not the end of the story, as one would also like to construct a region $\widehat{\Theta}_n$ that is well shaped and converges toward θ^* as n increases. Moreover, of no minor importance is the issue of the computational complexity associated to constructing $\widehat{\Theta}_n$. In the next section, we present existing methods, namely LSCR (Leave-out Sign-dominant Correlation Regions), SPS (Sign-Perturbed Sums) and PDMs (Perturbed Dataset Methods), and cast them within the setup of this section and also discuss the issue of the region shape and the computational complexity associated to these methods. This sheds light on the pros and cons of these various techniques in a comparative way, which is the first goal of this paper. Then, in the following section we introduce a new correlation method which combines some advantages of the above-mentioned approaches.

[4] A subtle issue may arise in case two $Z_i(\theta)$ functions take the same value. In this case, a suitable tie-break rule can be applied, and this aspect is discussed in the literature cited in the introduction while we neglect this aspect here because it would stray us too much into unnecessary details.

III. Revisiting Existing Finite-Sample Methods

In this section, we revisit three existing finite-sample approaches using the framework introduced in Section II.

A. The LSCR method

In its randomized formulation [2], LSCR fits into the framework of Section II where the function $Z_0(\theta)$ is simply defined as a sum of *error correlation terms*, such as, e.g., $\widehat{W}_t(\theta)\widehat{W}_{t-k}(\theta)$, or of *input-error correlation terms* such as, e.g., $\widehat{W}_t(\theta)U_{t-k}$, while the perturbed functions $Z_i(\theta)$ are obtained by replacing in the definition of $Z_0(\theta)$ the components of $\widehat{\mathbf{W}}_n(\theta)$ with the components of $\mathbf{s}_n^{(i)}[\widehat{\mathbf{W}}_n(\theta)]$. Consider, for example, $Z_0(\theta)=-\sum_{t=2}^n \widehat{W}_t(\theta)\widehat{W}_{t-1}(\theta)$. Then, for each θ, the ranking of Z_0 among $\{Z_0,\ldots,Z_{m-1}\}$ is equivalent to the ranking of 0 (the constant zero function) among $\{0,Z_1-Z_0,\ldots,Z_{m-1}-Z_0\}$. Note that Z_i-Z_0 is a sum of the kind $\sum_{t=2}^n \alpha_t\widehat{W}_t(\theta)\widehat{W}_{t-1}(\theta)$, where α_t is equal to 0 or 2 with equal probability: this is the *random subsampling idea* of [2].

Consistency results for LSCR are based on proving that in the long run, sums like $\sum_{t=2}^n \alpha_t\widehat{W}_t(\theta)\widehat{W}_{t-1}(\theta)$, for every $\theta\neq\theta^*$, tend to become large in absolute value, and therefore *every* $\theta\neq\theta^*$ will eventually be excluded from the region. However, in order to get consistency results, focusing on one sum only is not enough. For example, for ARMA(n_a,n_w) systems, the LSCR region is obtained by intersecting various regions $\widehat{\Theta}_n^{(k)}$, each of which constructed by considering a sum of the kind $\sum_{t=k+1}^n \widehat{W}_t(\theta)\widehat{W}_{t-k}(\theta)$ for different values of k.

In some cases, using different kinds of correlations such as input-error correlations or even higher order correlations is advisable, [1], [6]. Note that if every region $\widehat{\Theta}_n^{(k)}$ is guaranteed to include the true parameter θ^* with exact probability p, then the intersection $\widehat{\Theta}_n=\cap_{k=1}^{\bar{k}}\widehat{\Theta}_n^{(k)}$ includes θ^* with probability *at least* $1-(\bar{k}(1-p))$, by the union bound, which is a source of conservatism.

B. The SPS method

Consider a system in linear regression form as $Y_t=\varphi_t^\top\theta^*+W_t$, where φ_t is a function of $U_1,\ldots,U_t$ and W_t is the symmetric noise. Given n samples $Y_1,\ldots,Y_n$ and the corresponding regressors $\varphi_1,\ldots,\varphi_n$, the least-squares estimate $\hat{\theta}_{\mathrm{LS}}$ is obtained by minimizing $L(\theta)=\sum_{t=1}^n \widehat{W}_t^2(\theta)$, where $\widehat{W}_t(\theta)=Y_t-\widehat{Y}_t(\theta)$, and $\widehat{Y}_t(\theta)\triangleq\varphi_t^\top\theta$. $\hat{\theta}_{\mathrm{LS}}$ is the solution

(unique, under some technical conditions) of $\nabla_\theta L(\theta) = \sum_{t=1}^{n} \varphi_t \widehat{W}_t(\theta) = 0$.

1) SPS with exogenous regressors: In the prototypical SPS algorithm, under the assumption that the regressors $\{\varphi_t\}$ do not depend on outputs (i.e., regressors are exogenous), a normed version of $\nabla_\theta L(\cdot)$ is chosen as the reference element and thus $Z_0(\theta) = \|\sum_{t=1}^{n} \varphi_t \widehat{W}_t(\theta)\|_R^2$, where $\|\cdot\|_R^2$ is a suitably rescaled Euclidean norm, and $Z_i(\theta)$ is obtained by replacing $\widehat{\mathbf{W}}_n(\theta)$ with $\mathbf{s}_n^{(i)}[\widehat{\mathbf{W}}_n(\theta)]$. Note that, by construction, $Z_0(\hat{\theta}_{\mathrm{LS}}) = 0 \le Z_i(\hat{\theta}_{\mathrm{LS}})$, so that when $h_1 = 1$ the SPS region includes $\widehat{\theta}_{\mathrm{LS}}$. Moreover, the errors in all the components of θ are taken simultaneously into account by the norm. This idea will be henceforth referred to as the "*norm trick*".

2) SPS for ARX systems: Some difficulties arise when φ_t depends on past outputs, as it is in autoregressive systems. In this case simply using φ_t in both the reference Z_0 and the perturbed Z_i functions is not a valid option, because it would invalidate the key symmetry argument behind Claim 1. In fact, through past inputs, φ_t depends on noise terms and these noise terms have to undergo the sign perturbation in the Z_i functions. A solution to this problem is to "reconstruct" alternative output sequences based on the available information. Given any triplet of the kind $(\mathbf{U}_n', \mathbf{W}_n', \theta)$, the knowledge of $\mathbb{F}$ can be used to define an alternative output $\widetilde{\mathbf{Y}}_n$ as $\widetilde{\mathbf{Y}}_n \triangleq \mathbb{F}(\mathbf{U}_n', \mathbf{W}_n', \mathscr{I}; \theta)$, cf. (1). Using $\widetilde{\mathbf{Y}}_n$, also alternative regressors $\{\widetilde{\varphi}_t\}$ can be constructed that include elements of $\widetilde{\mathbf{Y}}_n$ instead of the actual output $\mathbf{Y}_n$. Finally, the Z function for a generic triple $(\mathbf{U}_n', \mathbf{W}_n', \theta)$ is defined as

$$Z(\mathbf{U}_n', \mathbf{W}_n', \theta) \triangleq \left\| \sum_{t=1}^{n} \widetilde{\varphi}_t W_t' \right\|_R^2.$$

Then, as usual, $Z_0(\theta) = Z(\mathbf{U}_n, \widehat{\mathbf{W}}_n(\theta), \theta)$. In Z_0, the values of $\widetilde{\varphi}_t$ and $\widetilde{\mathbf{Y}}_n$ are computed using θ and $(\mathbf{U}_n, \widehat{\mathbf{W}}_n(\theta))$. Therefore, by (1) and the invertibility assumption, the values of $\widetilde{\mathbf{Y}}_n$ *coincide* with the observed output values of $\mathbf{Y}_n$ for every θ, and $\widetilde{\varphi}_t = \varphi_t$. On the other hand, the Z_i's are obtained by replacing $\widehat{\mathbf{W}}_n(\theta)$ with $\mathbf{s}_n^{(i)}[\widehat{\mathbf{W}}_n(\theta)]$, so that $\widetilde{\varphi}_t$ and $\widetilde{\mathbf{Y}}_n$ are now reconstructed by using $\mathbf{s}_n^{(i)}[\widehat{\mathbf{W}}_n(\theta)]$ instead of the actual error $\widehat{\mathbf{W}}_n(\theta)$. Thus, denoting by $\widetilde{\mathbf{Y}}_n^{(i)}(\theta)$ the i-th reconstructed alternative output sequence, that is,

$$\widetilde{\mathbf{Y}}_n^{(i)}(\theta) = \mathbb{F}(\mathbf{U}_n, \mathbf{s}_n^{(i)}[\widehat{\mathbf{W}}_n(\theta)], \mathscr{I}; \theta), \tag{3}$$

we have that $\widetilde{\mathbf{Y}}_n^{(i)}(\theta) \neq \mathbf{Y}_n$ in general. It can be proven that with this approach Claim 1 remains rigorously valid [4].

C. Perturbed Dataset Methods

PDMs form an interesting class of methods that leave many degrees of freedom to the user and fit also situations where the joint symmetry assumption is replaced by other conditions such as arbitrary i.i.d. sequences. In these methods the *alternative output*, (3), plays the crucial role: a "perturbed dataset", in the terminology of [9], is any pair $(\mathbf{U}_n, \widetilde{\mathbf{Y}}_n^{(i)}(\theta))$. We focus here on a stimulating idea briefly mentioned in [9].

1) Bootstrap-style PDMs: Let functions Z_0 and $\{Z_i\}$ be

$$Z_0(\theta) \triangleq \|\theta - \widehat{\theta}_n(\mathbf{U}_n, \mathbf{Y}_n)\|_R^2,$$
$$Z_i(\theta) \triangleq \|\theta - \widehat{\theta}_n(\mathbf{U}_n, \widetilde{\mathbf{Y}}_n^{(i)}(\theta))\|_R^2,$$

where $\widehat{\theta}_n(\cdot)$ is a point-estimator. Claim 1 applies to this context. Moreover, in Z_0, function $\widehat{\theta}_n(\cdot)$ computes an estimate of θ^* based on the original input-output dataset, $(\mathbf{U}_n, \mathbf{Y}_n)$; hence, $Z_0(\theta^*) = \|\theta^* - \widehat{\theta}_n(\mathbf{U}_n, \mathbf{Y}_n)\|_R^2$ tends to be small for large n. On the other hand, for each other Z_i function, $\widehat{\theta}_n(\cdot)$ computes an estimate based on the *perturbed dataset* $(\mathbf{U}_n, \widetilde{\mathbf{Y}}_n^{(i)}(\theta))$; hence, $\widehat{\theta}_n(\mathbf{U}_n, \widetilde{\mathbf{Y}}_n^{(i)}(\theta))$ is an estimate of θ and $Z_i(\theta^*) = \|\theta^* - \widehat{\theta}_n(\mathbf{U}_n, \widetilde{\mathbf{Y}}_n^{(i)}(\theta))\|_R^2$ does not converge to zero as $n \to \infty$. Hence, by selecting $h_1 = 1$ one singles out in the long run the true θ^*.

It can be proved that, for FIR and ARX systems, by choosing $\widehat{\theta}_n(\cdot)$ as the least-squares estimator, the suggested method builds the same region as SPS. This is not true in the case of general linear systems with the prediction error estimator. In that case, one difficulty of the bootstrap PDM is that it is computationally intensive. In fact, computing $Z_i(\theta)$, for $i = 1, \ldots, m-1$, for any fixed θ, requires to calculate $\widehat{\theta}_n(\mathbf{U}_n, \widetilde{\mathbf{Y}}_n^{(i)}(\theta))$. Consequently, for every θ, one has to solve $m-1$ non-convex optimization problems.[5]

IV. A New Correlation Approach

In this section we introduce a new finite-sample identification method that combines some of the previous ideas into a new algorithm with improved properties.

A. Motivations

As we saw, LSCR is based on a correlation idea (combined with subsampling) which leads to a flexible and easy to implement algorithm. It is also computationally light, as unlike SPS and PDMs, LSCR does not require the generation of alternative, perturbed input-output datasets. However, the confidence bound resulting from intersecting individually exact regions makes LSCR conservative for high dimensional parameters.

SPS and PDMs evaluate the errors in all parameters simultaneously (norm-trick) and construct confidence regions having exact confidences. Unfortunately, the generation of alternative input-output datasets is required to ensure exact confidence in the case of more general systems. As a consequence, these methods can become difficult to analyze and computationally expensive or even impractical, especially when they involve hard optimization steps, as it is the case for bootstrap-style PDMs.

Here we aim at defining a new class of methods that exploits the correlation idea of LSCR, which makes the method computable, together with the norm trick of SPS, which makes the confidence of the constructed regions exact. One goal with this section is to stimulate further research in this direction.

[5] An interesting direction of research about PDMs is whether the estimator $\widehat{\theta}_n(\cdot)$ can be successfully replaced by an approximated estimator that is easy-to-compute.

B. Sign-perturbed correlation regions

The main idea of the new finite-sample method, called *Sign-Perturbed Correlation Regions* (SPCR), is as follows. Instead of defining a different Z function for each correlation and then intersecting the resulting regions as in LSCR, we stack the correlation sums into a vector and compute a single scalar "summary" of them by introducing a suitable norm.

Here we will present the method for ARX systems with the notations used in Example 1. Besides Assumptions 1 and 2, we also suppose that the system operates in open-loop, i.e., that the inputs $\{U_t\}$ and the noises $\{N_t\}$ are independent.

For a generic couple of input and noise vectors $\mathbf{U}'_n$ and $\mathbf{W}'_n$, we introduce the correlation vectors defined for every $t = 1,\dots,n$ as

$$\mathbf{C}_t(\mathbf{U}'_n,\mathbf{W}'_n) \triangleq (W'_tW'_{t-1},\dots,W'_tW'_{t-k},W'_tU'_t,\dots,W'_tU'_{t-l+1})^{\mathrm{T}},$$

where k and l are user-chosen parameters, typically $k+l \geq n_a+n_b$. We assume, for simplicity, that the given initial conditions allow us to compute the correlation vector, $\mathbf{C}_t(\mathbf{U}'_n,\mathbf{W}'_n)$, for all $t = 1,\dots,n$.

As we saw in Section II, the fundamental component of such methods is the Z function, which for SPCR is

$$Z(\mathbf{U}'_n,\mathbf{W}'_n,\theta) \triangleq \left\|\mathbf{Q}^{-\frac{1}{2}}(\mathbf{U}'_n,\mathbf{W}'_n)\frac{1}{n}\sum_{t=1}^{n}\mathbf{C}_t(\mathbf{U}'_n,\mathbf{W}'_n)\right\|^2,$$

where $\mathbf{Q}$ is a "scaling" matrix defined as

$$\mathbf{Q}(\mathbf{U}'_n,\mathbf{W}'_n) \triangleq \frac{1}{n}\sum_{t=1}^{n}\mathbf{C}_t(\mathbf{U}'_n,\mathbf{W}'_n)\mathbf{C}_t^{\mathrm{T}}(\mathbf{U}'_n,\mathbf{W}'_n),$$

which is assumed to be invertible, for convenience. As in the case of SPS, the "shaping" matrix $\mathbf{Q}$ has the role of balancing the action of the norm with respect to the variability of the different components. Note that the so defined Z is a function of $\mathbf{U}'_n,\mathbf{W}'_n$ only, that is, the third argument (the system parameter θ) is not used for computing the value of Z, and we can omit it. Finally, we define $Z_0(\theta) = Z(\mathbf{U}_n,\widehat{\mathbf{W}}_n(\theta))$ and $Z_i(\theta) = Z(\mathbf{U}_n,\mathbf{s}_n^{(i)}[\widehat{\mathbf{W}}_n(\theta)])$, which depend on θ only through the reconstructed noise $\widehat{\mathbf{W}}_n(\theta)$.

The confidence region construction is the same as before with $h_1 = 1$,

$$\widehat{\Theta}_n \triangleq \{\,\theta \in \mathbb{R}^{n_a+n_b} : R(\theta) \leq h_2\,\}.$$

Note that SPCR is a *class* of methods where different constructions correspond to different choices of (k,l). For more general (especially nonlinear) systems, it may be useful to also include higher-order correlations in $\{\mathbf{C}_t\}$ [6].

C. Properties of SPCR confidence regions

It is easy to see that the SPCR methods fit into the framework of Section II and Claim 1 holds. Therefore, the confidence regions constructed by SPCR are non-conservative, namely their confidence probabilities are *exactly* h_2/m.

Another nice property of SPCR is the inclusion of certain point-estimates. Assume, for simplicity, that $l+k = n_a+n_b$, then the correlation-type [10] point-estimate $\hat{\theta}$ satisfying

$$\frac{1}{n}\sum_{t=1}^{n}\mathbf{C}_t(\mathbf{U}_n,\widehat{\mathbf{W}}_n(\hat{\theta})) = 0,$$

is included in $\widehat{\Theta}_n$, since $Z_0(\hat{\theta}) = 0 \leq Z_i(\hat{\theta})$, for all i. For example, if $k = 0$ and $l = n_a+n_b$ we can guarantee the inclusion of an *instrumental variable* estimate, if the inputs are chosen as instrumental variables. In this case, the previously introduced IV-SPS [14] is a special case of SPCR. Other properties of SPS and LSCR are expected to carry over to SPCR, see also Sections V and VI.

D. Simulation example

Assume that the true system generating the output sequence $\{Y_t\}$ is a bilinear system [12] defined as

$$Y_t \triangleq a^*Y_{t-1} + b^*U_t + \frac{1}{2}U_tN_t + N_t,$$

for $t = 1,\dots,n$, with $a^* = 0.7$ and $b^* = 1$, with zero initial conditions. Notice that this system has the structure

$$Y_t \triangleq a^*Y_{t-1} + b^*U_t + W_t,$$

with $W_t = \frac{1}{2}U_tN_t + N_t$. Sequence $\{U_t\}$ is the measured input generated by $U_t \triangleq 0.5U_{t-1} + V_t$, with zero initial conditions, where $\{V_t\}$ is i.i.d. Gaussian with zero mean and unit variance. The noise sequence $\{N_t\}$ is i.i.d. Laplacian with zero mean and unit variance, independent of $\{U_t\}$.

Define

$$\widehat{Y}_t(\theta) \triangleq aY_{t-1} + bU_t.$$

Assuming we have a sample of $Y_1,\dots,Y_n$ and $U_1,\dots,U_n$, and using the zero initial conditions, we have that the residuals $\widehat{W}_t(\theta) \triangleq Y_t - \widehat{Y}_t(\theta)$ are well-defined for all $t \leq n$.

We apply SPCR with $k = l = 2$ and we assume that $n > 2$, for convenience, and leave out from the sum those vectors which surely contain some zero correlations. Thus, the reference ($i = 0$) and sign-perturbed functions ($i = 1,\dots,m-1$) are

$$Z_i(\theta) \triangleq \left\|\mathbf{Q}_i^{-\frac{1}{2}}(\theta)\frac{1}{n-2}\sum_{t=3}^{n}\begin{bmatrix}\sigma_{i,t-1}\widehat{W}_{t-1}(\theta)\\ \sigma_{i,t-2}\widehat{W}_{t-2}(\theta)\\ U_t\\ U_{t-1}\end{bmatrix}\sigma_{i,t}\widehat{W}_t(\theta)\right\|^2,$$

where $\sigma_{0,t} = 1$, for all t, while, for $i \neq 0$, $\{\sigma_{0,t}\}$ are i.i.d. random signs, as before. Matrix $\mathbf{Q}_i(\theta)$ is

$$\mathbf{Q}_i(\theta) \triangleq \frac{1}{n-2}\sum_{t=3}^{n}\begin{bmatrix}\sigma_{i,t-1}\widehat{W}_{t-1}(\theta)\\ \sigma_{i,t-2}\widehat{W}_{t-2}(\theta)\\ U_t\\ U_{t-1}\end{bmatrix}\begin{bmatrix}\sigma_{i,t-1}\widehat{W}_{t-1}(\theta)\\ \alpha_{i,t-2}\widehat{W}_{t-2}(\theta)\\ U_t\\ U_{t-1}\end{bmatrix}^{\mathrm{T}}\widehat{W}_t^2(\theta),$$

and is almost surely invertible, for $i = 0,\dots,m-1$.

It is easy to check that variables $\widehat{W}_t(\theta^*) = \frac{1}{2}U_tN_t + N_t$, $t = 1,\dots,n$, are jointly symmetric (use that $\{N_t\}$ are i.i.d. and symmetric, and $\{U_t\}$ is independent of $\{N_t\}$). Hence, the assumptions of Section II are satisfied and SPCR delivers *rigorously* guaranteed confidence regions, with exact probability of containing the true parameter values (a^*,b^*).

Figure 1 presents confidence regions built by SPCR for increasing number of observations, $n = 50,200,400$. The regions were built with $p = 0.95$, $m = 100$, and $h_2 = 95$. The figure is indicative of the phenomenon that the SPCR regions are well-shaped and shrink around the true parameter.

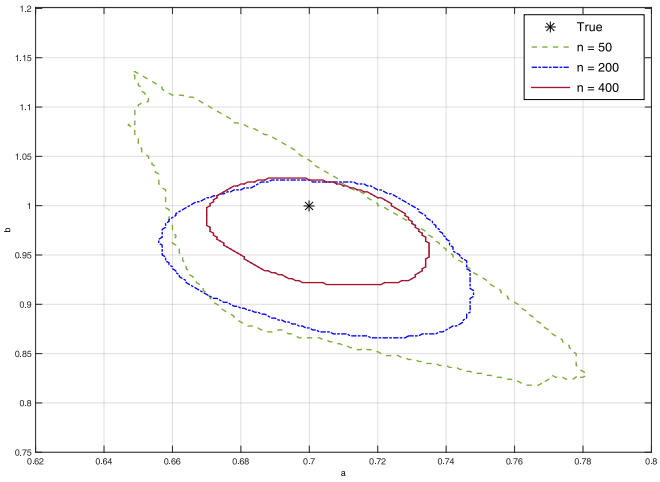

Fig. 1. 95% confidence regions built by SPCR with $k = 2$ and $l = 2$.

V. Desirable properties of finite-sample methods

Now, we return to the general overview of finite-sample methods and list some of the most important properties that one wants to achieve by suitably designing the Z function.

- *Inclusion of a point-estimate*: Confidence regions can help to assess the quality of point-estimates and, e.g., to determine how robust a design that is based on them should be. We know that SPS builds its confidence regions around the least-squares (LS) estimate, while SPCR can guarantee the inclusion of correlation-type estimates.
- *Consistency*: For any false parameter value, $\theta' \neq \theta^*$, the probability of $\theta' \in \widehat{\Theta}_n$ should decrease as the sample size, n, increases. Asymptotically, the coverage probability of any such false θ' should be zero. Some consistency results are available for LSCR [1] and SPS [15], and can be easily obtained for some bootstrap-style PDMs. It is yet to be proven whether SPCR inherits this property.
- *Favorable topology*: The constructed confidence region, $\widehat{\Theta}_n$, should have good topological properties. We know, for example, that the SPS confidence regions are star-convex (and hence also connected) with the LS estimate as a star centre, assuming exogenous regressors.
- *Weak computability*: Deciding whether a candidate θ belongs to $\widehat{\Theta}_n$ should be computationally easy. LSCR, SPS and SPCR are all weakly computable in that sense, even for endogenous regressors; but this may not hold for bootstrap-style PDMs, for which evaluating the Z function can quickly become too complex.
- *Strong computability*: Calculating a representation of $\widehat{\Theta}_n$ or an approximation of it should be computationally feasible. An ellipsoidal outer-approximation for SPS with exogenous regressors can be constructed efficiently by solving convex optimization problems [5]. Inner- and outer-approximations can also be built using interval-analysis, see [8] for LSCR and SPS.

VI. Conclusions

Finite-sample system identification methods are practically important as they provide rigorously guaranteed results under mild statistical assumptions. This paper has been prepared to foster research in this important field by providing an easy access-point to the neophyte. First, fundamental ideas behind finite-sample identification methods have been analyzed. Three existing approaches were revisited: LSCR, SPS and PDMs. Finally, a new non-asymptotic identification algorithm, SPCR, was suggested based on the idea of combining LSCR and SPS. SPCR has the flexibility and computational advantages of LSCR combined with the exact confidence of SPS. Finally, some essential properties of the aforementioned finite-sample identification methods were discussed.

We believe that SPCR is promising for the identification of complex systems, including nonlinear ones. Many results that were previously proved in the context of LSCR [1], [6] and SPS [3], [5] can be used for analyzing and extending this new correlation-type method. For example, in virtue of [1], we can argue that the consistency of the method can be improved by suitably prefiltering the input signal.

References

[1] Marco C. Campi and Erik Weyer. Guaranteed non-asymptotic confidence regions in system identification. *Automatica*, 41(10):1751–1764, 2005.

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

[3] Algo Carè, Balázs Cs. Csáji, and Marco C. Campi. Sign-perturbed sums (SPS) with asymmetric noise: Robustness analysis and robustification techniques. In *Proceedings of the 55th IEEE Conference on Decision and Control (CDC)*, 2016.

[4] Balázs Cs. Csáji, Marco C. Campi, and Erik Weyer. Sign-Perturbed Sums (SPS): A method for constructing exact finite-sample confidence regions for general linear systems. In *CDC*, pages 7321–7326, 2012.

[5] Balázs Cs. Csáji, Marco C. Campi, and Erik Weyer. Sign-Perturbed Sums: A new system identification approach for constructing exact non-asymptotic confidence regions in linear regression models. *IEEE Transactions on Signal Processing*, 63(1):169–181, 2015.

[6] Marco Dalai, Erik Weyer, and Marco C. Campi. Parameter identification for non-linear systems: guaranteed confidence regions through LSCR. *Automatica*, 43:1418–1425, 2007.

[7] Simone Garatti, Marco C. Campi, and Sergio Bittanti. Assessing the quality of identified models through the asymptotic theory – when is the result reliable? *Automatica*, 40(8):1319–1332, 2004.

[8] Michel Kieffer and Eric Walter. Guaranteed characterization of exact non-asymptotic confidence regions as defined by LSCR and SPS. *Automatica*, 50(2):507–512, 2014.

[9] Sándor Kolumbán, István Vajk, and Johan Schoukens. Perturbed datasets methods for hypothesis testing and structure of corresponding confidence sets. *Automatica*, 51:326–331, 2015.

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

[11] Mario Milanese, John Norton, Hélène Piet-Lahanier, and Éric Walter. *Bounding approaches to system identification*. New York, NY, USA: Springer, 2013.

[12] Ronald R. Mohler. *Bilinear control processes: with applications to engineering, ecology and medicine*. Academic Press, Inc., 1973.

[13] Torsten Söderström and Petre Stoica. *System Identification*. Prentice Hall International, Hertfordshire, UK, 1989.

[14] Valerio Volpe, Balázs Cs. Csáji, Algo Carè, Erik Weyer, and Marco C. Campi. Sign-perturbed sums (SPS) with instrumental variables for the identification of ARX systems. In *Proceedings of the 54th IEEE Conference on Decision and Control (CDC)*, 2015.

[15] Erik Weyer, Marco C. Campi, and Balázs Cs. Csáji. Asymptotic properties of SPS confidence regions. *Automatica*, 82:287 – 294, 2017.