Sign-Perturbed Sums: A New System Identification Approach for Constructing Exact Non-Asymptotic Confidence Regions in Linear Regression Models

Balázs Csanád Csáji, *Member, IEEE*, Marco Claudio Campi, *Fellow, IEEE*, and Erik Weyer, *Member, IEEE*

***Abstract*—We propose a new system identification method, called Sign–Perturbed Sums (SPS), for constructing non-asymptotic confidence regions under mild statistical assumptions. SPS is introduced for linear regression models, including but not limited to FIR systems, and we show that the SPS confidence regions have exact confidence probabilities, i.e., they contain the true parameter with a user-chosen exact probability for any finite data set. Moreover, we also prove that the SPS regions are star convex with the Least-Squares (LS) estimate as a star center. The main assumptions of SPS are that the noise terms are independent and symmetrically distributed about zero, but they can be nonstationary, and their distributions need not be known. The paper also proposes a computationally efficient ellipsoidal outer approximation algorithm for SPS. Finally, SPS is demonstrated through a number of simulation experiments.**

***Index Terms*—Finite sample properties, least squares methods, linear regression models, parameter estimation, statistics, system identification.**

I. Introduction

ESTIMATING parameters of partially unknown systems based on noisy observations is a classical problem in signal processing, system identification, machine learning and statistics. There are several standard methods available, which typically provide point estimates. Given an estimate, it is an intrinsic task to evaluate how close the estimated parameter is to the true one and such evaluation often comes in the form of confidence regions. Confidence regions are especially important for problems involving strict safety, stability or quality guarantees, and serve as a basis for ensuring robustness.

In practice, we only have a finite number of measurements and limited statistical knowledge about the noise, and this strongly restricts the number of methods available for constructing confidence regions, unless we are satisfied with approximate, heuristic solutions. Here, we propose a new statistical parameter estimation approach, called *Sign-Perturbed Sums* (SPS), for constructing finite-sample, quasi distribution-free confidence regions. This paper introduces and analyzes the SPS method for linear regression models including Finite Impulse Response (FIR) and Generalised FIR models.

Linear regression is a classical problem in statistics, which finds applications in many fields. It is a core component of identification, learning and prediction tasks, and a typical application is to estimate parameters of dynamical systems from experimental data, which is one of the fundamental problems of system identification [1]–[5]. Under natural conditions the Least-Squares (LS) method provides a strongly consistent point estimate of the system parameters. Moreover, the parameter estimation error is asymptotically normal, and this property can be used to build approximate confidence regions. However, these regions are based on the Central Limit Theorem, and hence are guaranteed only asymptotically as the number of data points tends to infinity. Therefore, applying the classical system identification theory with finitely many data points results only in heuristic confidence regions, which do not come with strict theoretical guarantees. This calls for alternative approaches that allow us to construct *guaranteed*, *non-asymptotic* confidence regions around the Least-Squares Estimate (LSE) under *mild statistical assumptions*.

The system identification method "Leave-out Sign-dominant Correlation Regions" (LSCR) was developed earlier [6]–[9] by authors of this paper. This method can build guaranteed, non-asymptotic confidence regions for parameters of various (linear and non-linear) dynamical systems under weak assumptions on the noise. LSCR has been applied successfully in various contexts [10], [11], and the topic has been the object of various investigations and related studies [12]–[17]. However, LSCR provides confidence regions with exact probabilities only for scalar parameters, while it offers bounds for the multidimensional case. Furthermore, the inclusion of the LSE in the confidence regions constructed by LSCR is not guaranteed.

Manuscript received May 06, 2014; revised September 16, 2014; accepted October 21, 2014. Date of publication November 10, 2014; date of current version December 04, 2014. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Petr Tichavsky. The work of B. C. Csáji was supported in part by the ARC under Grants DE120102601 and DP0986162, and the János Bolyai Fellowship of the Hungarian Academy of Sciences, under Grant BO/00683/12/6. The work of M. C. Campi was supported in part by the MIUR–Ministero dell'Istruzione, dell'Università e della Ricerca. The work of E. Weyer was supported by the Australian Research Council (ARC) under Discovery Grants DP0986162 and DP130104028.

B. C. Csáji is with MTA SZTAKI, The Institute for Computer Science and Control, Hungarian Academy of Sciences, H-1111 Budapest, Hungary (e-mail: balazs.csaji@sztaki.mta.hu).

M. C. Campi is with Department of Information Engineering, University of Brescia, 25123 Brescia, Italy (e-mail: marco.campi@unibs.it).

E. Weyer is with the Department of Electrical and Electronic Engineering, University of Melbourne, VIC 3010, Australia (e-mail: ewey@unimelb.edu.au).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSP.2014.2369000

The non-asymptotic SPS method, which is presented in this paper, provides exact confidence regions for multidimensional parameter vectors and guarantees the inclusion of the LSE in the confidence set. The exact finite-sample confidence probability is guaranteed even though the knowledge of the particular probability distributions of the noise is not assumed. The main assumptions on the noise terms are that they are independent and have symmetric distributions about zero, however, their distributions can change in each time-step. Regarding regressors, this paper concentrates on the deterministic case, while it is easy to generalize the results to the case where the regressors are random but independent of the noise.

The main contributions of the paper are as follows:

1) The SPS method for building confidence regions for linear regression problems is introduced.
2) The following finite-sample (non-asymptotic) results are proved for SPS under mild statistical assumptions:
   - The probability that the SPS region contains the true parameter is exact for any user-chosen probability.
   - The SPS confidence regions are star convex with the least-squares estimate as a star center.
3) The paper also discusses the practical implementation of SPS and introduces an ellipsoidal outer approximation algorithm which can be efficiently computed.
4) Finally, several experiments are presented that illustrate the SPS method, and compare it with alternatives such as the confidence ellipsoids based on the asymptotic system identification theory.

The structure of the paper is as follows. In Section II the linear regression problem is presented together with the assumptions. Section III gives a short overview of the LS method and its asymptotic theory. Section IV introduces the SPS method, while Section V presents its theoretical properties. In Section VI the ellipsoidal outer approximation algorithm is described followed by several numerical experiments in Section VII. Finally, Section VIII summarizes and concludes the paper. Preliminary versions of the results in this paper can be found in previous conference papers [18]–[20], which also contain additional numerical experiments.

II. Problem Setting

This section presents the linear regression problem and introduces our main assumptions and objectives.

A. Data Generation

Consider the following scalar linear regression system

$$Y_t \triangleq \varphi_t^{\mathrm{T}} \theta^* + N_t, \tag{1}$$

where Y_t is the output, N_t is the noise, φ_t is the regressor, and t is the discrete time index. Parameter θ^* is the true parameter to be estimated. The random variables Y_t and N_t are real-valued, while φ_t and θ^* are d dimensional real vectors. We consider a finite sample of size n which consists of the regressors $\varphi_1, \dots, \varphi_n$ and the outputs $Y_1, \dots, Y_n$.

For simplicity, we will consider deterministic regressors, $\{\varphi_t\}$, in this paper. Note, however, that our results can be easily generalized to the case of random, but exogenous, regressors, namely, to the case when the noise sequence $\{N_t\}$ is independent of the regressor sequence $\{\varphi_t\}$. In that case, our assumptions on the regressors (stated below) must be satisfied almost surely and then the analysis can be traced back to the presented theory by fixing a realization of the regressors (i.e., by conditioning on the σ-algebra generated by the regressors) and applying the presented results realization-wise.

B. Examples

There are many examples in signal processing and control of systems taking the form of (1) [1], [4]. The most common example is the widely used FIR model

$$Y_t = b_1^* U_{t-1} + b_2^* U_{t-2} + \cdots + b_d^* U_{t-d} + N_t,$$

where $\varphi_t = [U_{t-1}, \dots, U_{t-d}]^{\mathrm{T}}$ consists of past inputs and $\theta^* = [b_1^*, \dots, b_d^*]^{\mathrm{T}}$.

More generally, orthogonal functions (w.r.t. the Hardy space $\mathscr{H}_2$) are often used for modeling systems with slowly decaying impulse responses. Their transfer functions can be written as

$$G(z, \theta^*) = \sum_{k=1}^{d} \theta_k^* \, L_k(z, \alpha),$$

where z is the shift operator and $\{L_k(z, \alpha)\}$ is a function expansion with a (fixed) user-chosen parameter α. The regressor in this case is $\varphi_t = [\, L_1(z, \alpha)\, u_t, \, \dots, \, L_d(z, \alpha)\, u_t \,]^{\mathrm{T}}$.

Using $L_k(z, \alpha) = z^{-k}$ corresponds to the standard FIR model while, e.g., a Laguerre model is obtained by using the Laguerre polynomials [1], [21], [22],

$$L_k(z, \alpha) = \frac{1}{z - \alpha} \left(\frac{1 - \alpha z}{z - \alpha} \right)^{k-1}.$$

C. Basic Assumptions

Our assumptions on the regressors and the noise are:

A1 (independence, symmetricity): $\{N_t\}$ is a sequence of independent random variables. Each N_t has a symmetric probability distribution about zero.

A2 (outer product invertibility): $\det(R_n) \neq 0$, *where*

$$R_n \triangleq \frac{1}{n} \sum_{t=1}^{n} \varphi_t \varphi_t^{\mathrm{T}}.$$

The strongest assumption on the noise is that it forms an independent sequence (see Section V-C for comments on how this assumption can be relaxed). Apart from independence, the noise assumptions are rather weak, and the noise terms can be nonstationary with unknown distributions, and there are no moment or density requirements either. The other significant assumption is that the noise must be symmetric. Many standard distributions satisfy this property.

D. Objectives

Our goal is to construct *confidence regions* for the parameter θ^* that have *guaranteed* user-chosen confidence probabilities for finite, and possibly small, number of data points. The constructed regions are quasi distribution-free, as the only assumption on the noise distribution is A1. This is important since in practice the knowledge about the noise distribution is limited. Additionally, the confidence regions should contain the *least-squares* point estimate.

We will see that the SPS method proposed in this paper provides finite-sample confidence regions that have an exact user-chosen probability to contain θ^*. Despite the generality of our assumptions, the confidence regions are well-shaped and, in standard cases, they are similar in size to the regions that would be constructed with the full knowledge of the statistical characteristics of the noise.

III. Least Squares and its Asymptotic Theory

Before we present the SPS approach, we briefly recall the LS method and its associated asymptotic theory as they are used in later sections.

A. Least-Squares Estimate (LSE)

To find the LSE, we introduce the predictors

$$\hat{Y}_t(\theta) \triangleq \varphi_t^{\mathrm{T}}\theta.$$

The prediction errors for a given θ are

$$\varepsilon_t(\theta) \triangleq Y_t - \hat{Y}_t(\theta) = Y_t - \varphi_t^{\mathrm{T}}\theta,$$

and the LSE is found by minimizing the sum of the squared prediction errors, that is,

$$\hat{\theta}_n \triangleq \operatorname*{arg\,min}_{\theta\in\mathbb{R}^d} \sum_{t=1}^{n} \varepsilon_t^2(\theta) = \operatorname*{arg\,min}_{\theta\in\mathbb{R}^d} \sum_{t=1}^{n} (Y_t - \varphi_t^{\mathrm{T}}\theta)^2.$$

The solution can be found by solving the normal equation,

$$\sum_{t=1}^{n} \varphi_t\,\varepsilon_t(\theta) = \sum_{t=1}^{n} \varphi_t(Y_t - \varphi_t^{\mathrm{T}}\theta) = 0, \tag{2}$$

which has the analytic solution (assuming A2)

$$\hat{\theta}_n = \left(\sum_{t=1}^{n} \varphi_t\varphi_t^{\mathrm{T}}\right)^{-1} \left(\sum_{t=1}^{n} \varphi_t Y_t\right).$$

B. Asymptotic Confidence Regions

For zero mean independent and identically distributed (i.i.d.) noise, the LS estimation error is asymptotically Gaussian under mild conditions. More precisely, $\sqrt{n}\,(\hat{\theta}_n - \theta^*)$ converges in distribution to the Gaussian distribution with zero mean and covariance $\Gamma \triangleq \sigma^2 R^{-1}$, where σ^2 is the variance of the noise, and R is the limit of $R_n = \frac{1}{n}\sum_{t=1}^{n}\varphi_t\varphi_t^{\mathrm{T}}$ as $n \to \infty$ assuming this limit exists and is positive definite. As a consequence, $\frac{n}{\sigma^2}(\hat{\theta}_n - \theta)^{\mathrm{T}} R\,(\hat{\theta}_n - \theta)$ converges in distribution to the χ^2 distribution with $\dim(\theta^*) = d$ degrees of freedom [1].

Replacing R with R_n and σ^2 with the estimate

$$\hat{\sigma}_n^2 \triangleq \frac{1}{n-d} \sum_{t=1}^{n} \varepsilon_t^2(\hat{\theta}_n), \tag{3}$$

an approximate confidence region can be built as

$$\widetilde{\Theta}_n \triangleq \left\{ \theta \in \mathbb{R}^d : (\theta - \hat{\theta}_n)^{\mathrm{T}} R_n\,(\theta - \hat{\theta}_n) \le \frac{\mu\hat{\sigma}_n^2}{n} \right\}, \tag{4}$$

where the probability that θ^* is in the confidence region $\widetilde{\Theta}_n$ is approximately $F_{\chi^2(d)}(\mu)$, where $F_{\chi^2(d)}$ is the cumulative distribution function of the χ^2 distribution with d degrees of freedom. However, this confidence region based on the asymptotic system identification theory does not come with rigorous theoretical guarantees for the finite sample case, and therefore should only be used as a heuristic.

IV. The Sign-Perturbed Sums (SPS) Method

In this section, we first motivate, and then formally introduce, the *Sign-Perturbed Sums* (SPS) method for constructing confidence regions with guaranteed finite sample properties.

A. Intuitive Idea

The LS estimate is obtained as the solution to the normal (2). This equation can be re-written as

$$\sum_{t=1}^{n} \varphi_t\varphi_t^{\mathrm{T}}\tilde{\theta} + \sum_{t=1}^{n} \varphi_t N_t = 0,$$

where $\tilde{\theta} \triangleq \theta^* - \theta$. It is clear that the uncertainty in the LSE comes from the noise $\{N_t\}$, and, if each N_t were zero, then $\hat{\theta}_n = \theta^*$. In order to construct a confidence region, we should somehow evaluate the uncertainty of the estimate. One way of doing this would be to assume a particular probability distribution of the noise and propagate this distribution through the above formula to get a distribution of the estimation error. Then, the distribution of the estimation error can be used to construct the confidence region. However, we want to avoid such an approach as it needs strong prior assumptions on the noise, which makes it unattractive for practical purposes.

We follow another approach and try to exploit the information in the data as much as possible while assuming minimal prior statistical knowledge about the noise. Our core assumption is the *symmetry* of the noise. We introduce $m-1$ *sign-perturbed sums*

$$\begin{aligned} H_i(\theta) &= \sum_{t=1}^{n} \varphi_t\alpha_{i,t}(Y_t - \varphi_t^{\mathrm{T}}\theta) \\ &= \sum_{t=1}^{n} \alpha_{i,t}\varphi_t\varphi_t^{\mathrm{T}}\tilde{\theta} + \sum_{t=1}^{n} \alpha_{i,t}\varphi_t N_t, \end{aligned}$$

$i = 1,\ldots,m-1$, where $\{\alpha_{i,t}\}$ are random signs, i.e., i.i.d. random variables that take on the values ±1 with probability 1/2 each. That is, we perturb the sign of the prediction errors in the normal equation. For a given θ, we can also calculate the value for the case when no sign-perturbations are used, which we call the *reference sum*,

$$H_0(\theta) = \sum_{t=1}^{n} \varphi_t (Y_t - \varphi_t^{\mathrm{T}} \theta) = \sum_{t=1}^{n} \varphi_t \varphi_t^{\mathrm{T}} \tilde{\theta} + \sum_{t=1}^{n} \varphi_t N_t .$$

A comparison of the $H_0(\theta)$ and $H_i(\theta)$ functions can be done by using a norm $\|\cdot\|$. In the sequel, if not otherwise stated, $\|\cdot\|$ will refer to the 2-norm, i.e., $\|x\|^2 = x^{\mathrm{T}} x$.

For $\theta = \theta^*$, these sums can be simplified to

$$\begin{aligned} H_0(\theta^*) &= \sum_{t=1}^{n} \varphi_t N_t, \\ H_i(\theta^*) &= \sum_{t=1}^{n} \alpha_{i,t} \varphi_t N_t = \sum_{t=1}^{n} \pm \varphi_t N_t, \end{aligned}$$

where in the last equation we have written ± instead of $\alpha_{i,t}$ for intuitive understanding. $H_0(\theta^*)$ and $H_i(\theta^*)$ have the *same distribution* since $\{N_t\}$ are independent and symmetric. Therefore, there is no reason why a particular $\|H_j(\theta^*)\|^2$ should be bigger or smaller than another $\|H_i(\theta^*)\|^2$ and the probability that a particular $\|H_j(\theta^*)\|^2$ is the kth largest one in the ordering of $\{\|H_i(\theta^*)\|^2\}_{i=0}^{m-1}$ will be the same for all j, including $j = 0$ (the case of the reference sum, i.e., where there are no sign-perturbations). As j can take on m different values, this probability is exactly $1/m$.

However, for "*large enough*" $\|\tilde{\theta}\|$, we will have that

$$\left\| \sum_{t=1}^{n} \varphi_t \varphi_t^{\mathrm{T}} \tilde{\theta} + \sum_{t=1}^{n} \varphi_t N_t \right\|^2 > \left\| \sum_{t=1}^{n} \pm \varphi_t \varphi_t^{\mathrm{T}} \tilde{\theta} + \sum_{t=1}^{n} \pm \varphi_t N_t \right\|^2,$$

with " *high probability*". In fact, $\sum_{t=1}^{n} \varphi_t \varphi_t^{\mathrm{T}} \tilde{\theta}$ on the left-hand side increases faster than $\sum_{t=1}^{n} \pm \varphi_t \varphi_t^{\mathrm{T}} \tilde{\theta}$ on the right-hand side. Hence, for $\|\tilde{\theta}\|$ large enough, $\|H_0(\theta)\|^2$ dominates in the ordering of $\{\|H_i(\theta)\|^2\}$.

From these intuitions, the general idea is to construct the confidence region based on the *rankings* of the functions $\{\|H_i(\theta)\|^2\}$ and leave out those θ parameters for which $\|H_0(\theta)\|^2$ "*dominates*" the other functions.

In the formal construction of the SPS method, functions $\{H_i(\theta)\}$ will be modified with a term, $R_n^{-1/2}$, that helps to shape the region and an $1/n$ factor to increase the numerical stability, that is, we will use $S_i(\theta) = R_n^{-1/2} \frac{1}{n} H_i(\theta)$ instead of $H_i(\theta)$, cf. the pseudocode in Table II. However, this does not affect the core idea of the construction. Next, we provide the formal construction of SPS, followed by results stating some finite-sample properties of the obtained confidence sets.

B. Confidence Region Construction

The SPS method is in two parts. The first part, which we call initialization, sets the main global parameters of SPS and generates the random objects needed for the construction. In the initialization, the user provides the desired confidence probability p. The second part evaluates an indicator function, which can be called for a particular parameter value θ to decide whether it is included in the confidence region.

TABLE I

```
PSEUDOCODE: SPS-INITIALIZATION
1. Given a (rational) confidence probability $p \in (0,1)$,
   set integers $m > q > 0$ such that $p = 1 - q/m$;
2. Calculate the outer product
       $R_n \triangleq \frac{1}{n} \sum_{t=1}^{n} \varphi_t \varphi_t^{\mathrm{T}},$
   and find a factor $R_n^{1/2}$ such that
       $R_n^{1/2} R_n^{1/2\mathrm{T}} = R_n;$
3. Generate $n\,(m-1)$ i.i.d. random signs $\{\alpha_{i,t}\}$ with
       $\mathbb{P}(\alpha_{i,t} = 1) = \mathbb{P}(\alpha_{i,t} = -1) = \frac{1}{2},$
   for $i \in \{1, \dots, m-1\}$ and $t \in \{1, \dots, n\}$;
4. Generate a random permutation $\pi$ of the set
   $\{0, \dots, m-1\}$, where each of the $m!$ possible
   permutations has the same probability $1/(m!)$
   to be selected.
```

TABLE II

```
PSEUDOCODE: SPS-INDICATOR ( $\theta$ )
1. For the given $\theta$, compute the prediction errors
   for $t \in \{1, \dots, n\}$
       $\varepsilon_t(\theta) \triangleq Y_t - \varphi_t^{\mathrm{T}} \theta;$
2. Evaluate
       $S_0(\theta) \triangleq R_n^{-\frac{1}{2}} \frac{1}{n} \sum_{t=1}^{n} \varphi_t \varepsilon_t(\theta),$
       $S_i(\theta) \triangleq R_n^{-\frac{1}{2}} \frac{1}{n} \sum_{t=1}^{n} \alpha_{i,t}\, \varphi_t \varepsilon_t(\theta),$
   for $i \in \{1, \dots, m-1\}$;
3. Order scalars $\{\|S_i(\theta)\|^2\}$ according to $\succ_\pi$;
4. Compute the rank $\mathcal{R}(\theta)$ of $\|S_0(\theta)\|^2$ in the ordering,
   where $\mathcal{R}(\theta) = 1$ if $\|S_0(\theta)\|^2$ is the smallest in the
   ordering, $\mathcal{R}(\theta) = 2$ if $\|S_0(\theta)\|^2$ is the second
   smallest, and so on.
6. Return 1 if $\mathcal{R}(\theta) \le m - q$, otherwise return 0.
```

The pseudocode for the initialization is given in Table I. The permutation π in point 4 is only used in the indicator function to break ties, and decide which function $\|S_i(\theta)\|^2$ or $\|S_j(\theta)\|^2$ is the "*larger*" if $\|S_i(\theta)\|^2$ and $\|S_j(\theta)\|^2$ take on the same value. More precisely, given m real numbers $\{Z_i\}$, $i = 0, \dots, m-1$, we define a strict total order $\succ_\pi$ by

$$\begin{aligned} Z_k \succ_\pi Z_j \qquad & \text{if and only if} \\ & (\, Z_k > Z_j \,) \;\text{ or }\; (\, Z_k = Z_j \;\text{ and }\; \pi(k) > \pi(j) \,) . \end{aligned}$$

Note that π is a bijection (one-to-one correspondence) from $\{0, \dots, m-1\}$ to itself, thus, for $k \neq j$, $\pi(k)$ and $\pi(j)$ are two different integers in $\{0, \dots, m-1\}$.

After SPS is initialized, the indicator function given in Table II can be called to decide whether a particular parameter value θ is included in the confidence region.

Using this construction, we can define the p-level *SPS confidence region* as follows

$$\widehat{\Theta}_n \triangleq \left\{ \theta \in \mathbb{R}^d : \text{SPS-INDICATOR}(\,\theta\,) = 1 \right\} .$$

Observe that the LS estimate, $\hat{\theta}_n$, has by definition the property that $S_0(\hat{\theta}_n) = 0$. Therefore, the LSE is included in the SPS confidence region, assuming that it is non-empty.[1]

The SPS method in the form developed in this section lends itself nicely to problems where the indicator function should only be evaluated for finitely many values of θ. This happens for example in certain hypothesis testing or change detection problems. Later, we will discuss ways to make SPS suitable for problems where one wishes to represent the whole SPS confidence regions compactly.

V. Theoretical Results

A. Exact Confidence

The most important property of the SPS method is that the regions it generates have *exact* confidence probabilities for any *finite* sample. The following theorem holds.

Theorem 1: Assuming A1 and A2, the confidence probability of the constructed confidence region is exactly p, *that is,*

$$\mathbb{P}\big(\theta^* \in \widehat{\Theta}_n\big) = 1 - \frac{q}{m} = p.$$

A formal proof of Theorem 1 can be found in Appendix A. Interestingly, the proof does not depend on the applied norm, and the result keeps its validity regardless of the norm used in step 3 in the SPS indicator function when constructing $\widehat{\Theta}_n$. Since the confidence probability is exact, no conservatism is introduced. Moreover, the statistical assumptions imposed on the noise are mild, e.g., knowledge of the particular noise distribution is not assumed, the noise can change in each time step, and there are no moment or density assumptions.

The simulation examples in Section VII also demonstrate that, when the noise is stationary, the SPS confidence regions compare in size with the approximate confidence regions obtained by applying the asymptotic system identification theory, while, unlike asymptotic regions, the SPS regions maintain their guaranteed validity even for nonstationary noise patterns.

B. Star Convexity

Earlier we observed that the LSE is in the SPS confidence region. The next theorem makes our claim that the SPS regions are built around the LS estimate more precise.

Recall that set $\mathcal{X} \subseteq \mathbb{R}^d$ is called *star convex* if there is a *star center* $c \in \mathbb{R}^d$, such that

$$\forall x \in \mathcal{X},\ \forall \beta \in [0,1] : \beta\, x + (1-\beta)\, c \in \mathcal{X}.$$

All convex sets are star convex, but the converse is not true.

It is easy to construct examples that show that, in general, the SPS confidence regions are not convex. For example, if $q = 1$, the SPS region is the union of ellipsoids, and it is typically nonconvex. On the other hand, as the next theorem demonstrates, the SPS confidence regions are star convex.

Theorem 2: Assuming A1 and A2, the SPS confidence regions are star convex with the LS estimate as a star center.

The proof of Theorem 2 is given in Appendix B.

This result not only shows that the SPS regions are centered around the LS estimate, but it also provides a basis for finding the boundary of the SPS region. In fact, one can search rays from the LS estimate outwards for the first point which is not in the SPS region, and by the star convexity property this will be a boundary point of the SPS region.

C. A More General Algorithm: Block SPS

The fundamental assumption regarding the noise terms is that they are symmetric about zero and independent. Theoretically, it is easy to relax the independence assumption and allow dependent noises as long as their signs are independent. Moreover, robustness against the independence assumption can be boosted by using a modified SPS method where the random signs $\{\alpha_{i,t}\}$ are kept at the same value $+1$ or -1 for blocks of T consecutive time instants before the sign is again randomly drawn. The only difference is in point 3 of the SPS-Initialization algorithm, which now becomes (assuming, for simplicity, that n/T is an integer)

3'. Generate $\frac{n}{T}\,(m-1)$ i.i.d. random signs $\{\bar{\alpha}_{i,k}\}$ with

$$\mathbb{P}(\bar{\alpha}_{i,k} = 1) = \mathbb{P}(\bar{\alpha}_{i,k} = -1) = \tfrac{1}{2},$$

for $i \in \{1,\dots,m-1\}$, $k \in \{1,\dots,n/T\}$, and let

$$\alpha_{i,(k-1)T+j} = \bar{\alpha}_{i,kT}$$

for $i \in \{1,\dots,m-1\}$, $k \in \{1,\dots,n/T\}$, and $j \in \{1,\dots T\}$.

If the noise is dependent and T is larger than the dominant time constant of the noise dynamics, then the blocks of T consecutive noise terms act approximately as independent noise terms so that the result in Theorem 1 holds approximately.

When instead this modified Block SPS method is applied to systems where the noise is actually independent, the exact confidence result in Theorem 1 remains valid as can be seen from an inspection of the proof. Moreover, for independent noise, the regions constructed by SPS and Block SPS methods are not very different and Block SPS is only marginally worse. See Section VII-E for a simulation example.

D. Comparison to Bootstrap

A common feature in SPS and bootstrap approaches is that randomization is an essential ingredient.

In case of linear regression problems [23]–[25], we are interested in the distribution of the noise, however, we do not have a direct access to the noise samples. In order to overcome this difficulty, bootstrap typically uses the prediction errors and works with the sample of residuals, $\{\varepsilon_t(\theta)\}$, as an estimate of the noise, instead of the sample of the (unobserved) noise terms, $\{N_t\}$. One can use residuals for different parameters to test various hypotheses [25] or, as is common, one can work with the prediction errors of a nominal estimate [23], [24], such as the LS residuals, $\{\varepsilon_t(\hat{\theta}_n)\}$.

The latter case corresponds to work with the new data set(s) $\{\varphi_\kappa, Y'_\kappa\}$ where Y'_κ is generated by $Y'_\kappa = \varphi_\kappa^{\mathrm{T}}\hat{\theta}_n + \tilde{\varepsilon}_\kappa$, with

[1] There is a positive, but negligible, probability that the SPS confidence region is empty. This happens, if for at least $m-q$ indices i, the $\{\alpha_{i,t}\}$ sequences are sequences of all $+1$s or all -1s, and $m-q$ or more of the corresponding $\{S_i\}$ functions are ranked smaller than S_0 by π.

$\tilde{\varepsilon}_\kappa$ randomly selected (uniformly, with replacement) from $\{\varepsilon_t(\hat{\theta}_n)\}$. Various statistics can then be calculated from the resampled data set(s). An alternative way of generating data set(s) is pairs bootstrap [24], where data sets are generated by random selection (with replacement) of regressor-output pairs, (φ_τ, Y_τ), and the bootstrap samples are built from these pairs.

The domain of applicability of bootstrap is larger than SPS, since there is no assumption that the noise is symmetric and there are also bootstrap methods that can handle correlated noise. Moreover, bootstrap can be applied to non-linear models. On the other hand, while there are theoretical asymptotic results for bootstrap methods, there are few finite sample results and hence the results based on bootstrap are in most cases only approximate in the finite sample case. For example, if the approach for estimating the covariance matrix of the LS estimate, found in Ch.9 of Efron and Tibshirani's classical book [23], is combined with an assumption that the estimate has a Gaussian distribution, then the confidence ellipsoids of asymptotic system identification theory are obtained and they are approximate and not exact in a finite sample setting.

VI. Ellipsoidal Approximation Algorithm

Given a particular value of θ, it is easy to check whether θ is in the confidence region. All we have to do is to calculate the $\{\|S_i(\theta)\|^2\}$ functions for that θ and compare them. Hence the SPS confidence regions can be constructed by checking each parameter value on a grid. However, this approach is computationally demanding and suffers from the " *curse of dimensionality*". Here, we present an approximation algorithm for SPS that can be efficiently computed (i.e., in polynomial time) and offers a compact representation in the form of ellipsoidal over-bounds. An alternative approach based on interval analysis has also been proposed [26], [27].

A. Ellipsoidal Outer Approximation

Expanding $\|S_0(\theta)\|^2$, we find that it can be written as

$$\begin{aligned}\|S_0(\theta)\|^2 &= \left[\frac{1}{n}\sum_{t=1}^{n}\varphi_t(Y_t-\varphi_t^{\mathrm{T}}\theta)\right]^{\mathrm{T}} R_n^{-1}\left[\frac{1}{n}\sum_{t=1}^{n}\varphi_t(Y_t-\varphi_t^{\mathrm{T}}\theta)\right]\\ &= \left[\frac{1}{n}\sum_{t=1}^{n}\varphi_t\varphi_t^{\mathrm{T}}(\theta-\hat{\theta}_n)\right]^{\mathrm{T}} R_n^{-1}\left[\frac{1}{n}\sum_{t=1}^{n}\varphi_t\varphi_t^{\mathrm{T}}(\theta-\hat{\theta}_n)\right]\\ &= (\theta-\hat{\theta}_n)^{\mathrm{T}}R_n(\theta-\hat{\theta}_n),\end{aligned}$$

For the purpose of finding an ellipsoidal over-bound we can ignore the random ordering used when $\|S_0(\theta)\|^2$ and $\|S_i(\theta)\|^2$ are equal, and consider the set given by those values of θ at which q of the $\|S_i(\theta)\|^2$ are larger *or equal* to $\|S_0(\theta)\|^2$, i.e.,

$$\widehat{\Theta}_n \subseteq \left\{\,\theta\in\mathbb{R}^d : (\theta-\hat{\theta}_n)^{\mathrm{T}}R_n(\theta-\hat{\theta}_n) \le r(\theta)\,\right\},$$

where $r(\theta)$ is the q th largest value of functions $\{\|S_i(\theta)\|^2\}$, $i=1,\dots,m-1$. The idea is now to seek an over-bound by replacing $r(\theta)$ with a parameter independent r. This outer approximation will hence have the *same shape* and *orientation* as the asymptotic confidence ellipsoid (4), but it will have a different volume. The outer approximation is a guaranteed confidence region for finitely many data points. Moreover, it will have a compact representation, since it is characterized in terms of $\hat{\theta}_n, R_n$ and r.

B. Convex Programming Formulation

Comparing $\|S_0(\theta)\|^2$ with one single $\|S_i(\theta)\|^2$ function, we have

$$\begin{aligned}&\{\,\theta : \|S_0(\theta)\|^2 \le \|S_i(\theta)\|^2\,\}\\ &\qquad\subseteq \left\{\theta : \|S_0(\theta)\|^2 \le \max_{\theta:\|S_0(\theta)\|^2\le\|S_i(\theta)\|^2}\|S_i(\theta)\|^2\right\}.\end{aligned}$$

Relation $\|S_0(\theta)\|^2 \le \|S_i(\theta)\|^2$ can be rewritten as

$$\begin{aligned}&(\theta-\hat{\theta}_n)^{\mathrm{T}}R_n(\theta-\hat{\theta}_n)\\ &\le \left[\frac{1}{n}\sum_{t=1}^{n}\alpha_{i,t}\varphi_t(Y_t-\varphi_t^{\mathrm{T}}\theta)\right]^{\mathrm{T}} R_n^{-1}\left[\frac{1}{n}\sum_{t=1}^{n}\alpha_{i,t}\varphi_t(Y_t-\varphi_t^{\mathrm{T}}\theta)\right]\\ &= \theta^{\mathrm{T}}Q_iR_n^{-1}Q_i\theta - 2\,\theta^{\mathrm{T}}Q_iR_n^{-1}\psi_i + \psi_i^{\mathrm{T}}R_n^{-1}\psi_i,\end{aligned}$$

where matrix Q_i and vector ψ_i are defined as

$$\begin{aligned}Q_i &\triangleq \frac{1}{n}\sum_{t=1}^{n}\alpha_{i,t}\varphi_t\varphi_t^{\mathrm{T}},\\ \psi_i &\triangleq \frac{1}{n}\sum_{t=1}^{n}\alpha_{i,t}\varphi_tY_t.\end{aligned}$$

Noting that

$$\max_{\theta:\|S_0(\theta)\|^2\le\|S_i(\theta)\|^2}\|S_i(\theta)\|^2 = \max_{\theta:\|S_0(\theta)\|^2\le\|S_i(\theta)\|^2}\|S_0(\theta)\|^2$$

and using the notation $z \triangleq R_n^{\frac{1}{2}\mathrm{T}}(\theta-\hat{\theta}_n)$, the quantity $\max_{\theta:\|S_0(\theta)\|^2\le\|S_i(\theta)\|^2}\|S_i(\theta)\|^2$ can be obtained as the value of the following optimization problem

$$\begin{aligned}&\text{maximize} && \|z\|^2\\ &\text{subject to} && z^{\mathrm{T}}A_iz + 2z^{\mathrm{T}}b_i + c_i \le 0,\end{aligned}\qquad (5)$$

where A_i, b_i and c_i are defined as

$$\begin{aligned}A_i &\triangleq I - R_n^{-\frac{1}{2}}Q_iR_n^{-1}Q_iR_n^{-\frac{1}{2}\mathrm{T}},\\ b_i &\triangleq R_n^{-\frac{1}{2}}Q_iR_n^{-1}(\psi_i - Q_i\hat{\theta}_n),\\ c_i &\triangleq -\,\psi_i^{\mathrm{T}}R_n^{-1}\psi_i + 2\hat{\theta}_n^{\mathrm{T}}Q_iR_n^{-1}\psi_i - \hat{\theta}_n^{\mathrm{T}}Q_iR_n^{-1}Q_i\hat{\theta}_n.\end{aligned}$$

This program is not convex in general. However, it can be shown [28, Appendix B] that *strong duality* holds, so that the value of the above optimization program is equal to the value of its dual which can be formulated as

$$\begin{aligned}&\text{minimize} \quad \gamma\\ &\text{subject to} \quad \lambda \ge 0\\ &\begin{bmatrix} -I+\lambda A_i & \lambda b_i\\ \lambda b_i^{\mathrm{T}} & \lambda c_i+\gamma\end{bmatrix} \succeq 0,\end{aligned}\qquad (6)$$

TABLE III

Pseudocode: SPS-Outer-Approximation

```
1. Compute the least-squares estimate,
   $\hat{\theta}_n = R_n^{-1}\left[\frac{1}{n}\sum_{t=1}^{n}\varphi_t Y_t\right]$;
2. For $i \in \{1,\dots,m-1\}$, solve the optimization
   problem (5), and let $\gamma_i^*$ be the optimal value;
3. Let $r$ be the $q$th largest $\gamma_i^*$ value;
4. The outer approximation of the SPS confidence
   region is given by the ellipsoid
   $\widehat{\widehat{\Theta}}_n = \{\,\theta \in \mathbb{R}^d : (\theta - \hat{\theta}_n)^{\mathrm{T}} R_n (\theta - \hat{\theta}_n) \leq r\,\}$.
```

where “$\succeq 0$” denotes that a matrix is positive semidefinite. This program is convex, and can be easily solved using, e.g., Yalmip [29] and a solver such as SDPT3.

Letting γ_i^* be the value of program (6), we now have

$$\{\theta : \|S_0(\theta)\|^2 \leq \|S_i(\theta)\|^2\} \subseteq \{\theta : \|S_0(\theta)\|^2 \leq \gamma_i^*\}.$$

Thus,

$$\widehat{\Theta}_n \subseteq \widehat{\widehat{\Theta}}_n \triangleq \left\{\theta \in \mathbb{R}^d : (\theta - \hat{\theta}_n)^{\mathrm{T}} R_n (\theta - \hat{\theta}_n) \leq r\right\},$$

where $r = q$th largest value of γ_i^*, $i = 1,\dots,m-1$.

$\widehat{\widehat{\Theta}}_n$ is the sought outer approximation. It is clear that

$$\mathbb{P}\left(\theta^* \in \widehat{\widehat{\Theta}}_n\right) \geq 1 - \frac{q}{m} = p,$$

for any finite n.

The pseudocode for computing $\widehat{\widehat{\Theta}}_n$ is given in Table III.

VII. Simulation Examples

In this section we illustrate SPS with numerical examples. The confidence regions constructed by SPS are compared with those obtained using asymptotic system identification theory, and, when the noise is i.i.d. Gaussian, with the ellipsoids based on the F-distribution. The effect of using norms other than the 2-norm as well as the presence of unmodelled dynamics are also studied. The block SPS algorithm is illustrated on an example where the assumptions on the noise are not satisfied.

A. Second Order FIR System

We consider a second order data generating FIR system

$$Y_t = b_1^* U_{t-1} + b_2^* U_{t-2} + N_t,$$

where $b_1^* = 0.7$ and $b_2^* = 0.3$ are the true system parameters and $\{N_t\}$ is a sequence of i.i.d. Laplacian random variables with zero mean and variance 0.1. The input signal is given by

$$U_t = 0.75\, U_{t-1} + V_t,$$

where $\{V_t\}$ is a sequence of i.i.d. Gaussian random variables with zero mean and variance 1.

The model class is the class of second order FIR models, and hence the predictor is given by

$$\hat{Y}_t(\theta) = b_1 U_{t-1} + b_2 U_{t-2} = \varphi_t^{\mathrm{T}}\theta,$$

where $\theta = [\,b_1, b_2\,]^{\mathrm{T}}$ is the model parameter, and $\varphi_t = [\,U_{t-1}, U_{t-2}\,]^{\mathrm{T}}$.

Based on $n = 25$ data points $(\varphi_t, Y_t) = ([\,U_{t-1},\ U_{t-2}\,]^{\mathrm{T}}, Y_t)$, $t = 1,\dots,25$, we want to find a 95% confidence region for $\theta^* = [b_1^*,\ b_2^*]^{\mathrm{T}}$. Following the SPS procedure we first compute the matrix

$$R_{25} = \frac{1}{25}\sum_{t=1}^{25}\begin{bmatrix}U_{t-1}\\ U_{t-2}\end{bmatrix}[U_{t-1},\ U_{t-2}],$$

and find a factor $R_{25}^{\frac{1}{2}}$ such that $R_{25}^{\frac{1}{2}}R_{25}^{\frac{1}{2}\mathrm{T}} = R_{25}$.

Then we compute the reference sum

$$S_0(\theta) = R_{25}^{-\frac{1}{2}}\frac{1}{25}\sum_{t=1}^{25}\begin{bmatrix}U_{t-1}\\ U_{t-2}\end{bmatrix}(Y_t - b_1 U_{t-1} - b_2 U_{t-2}),$$

and the 99 sign perturbed sums, $i = 1,\dots,99$,

$$S_i(\theta) = R_{25}^{-\frac{1}{2}}\frac{1}{25}\sum_{t=1}^{25}\alpha_{i,t}\begin{bmatrix}U_{t-1}\\ U_{t-2}\end{bmatrix}(Y_t - b_1 U_{t-1} - b_2 U_{t-2}),$$

where $\alpha_{i,t}$ are i.i.d. random signs. Moreover, we generate a random permutation π to break possible ties.

The confidence region is constructed as the values of θ for which at least 5 of the $\|S_i(\theta)\|^2$, $i = 1,\dots,99$, functions are larger than $\|S_0(\theta)\|^2$. Here $m = 100$ and $q = 5$ and it follows from Theorem 1 that the constructed region contains the true parameter with exact probability $1 - \frac{5}{100} = 95\%$.

The SPS confidence region is shown in Fig. 1 together with the outer approximation and the confidence region based on the asymptotic system identification theory.

It can be observed that the non-asymptotic SPS region is similar in shape to, and not too different in size from, the asymptotic confidence region, while it has the advantage that it is guaranteed to contain the true parameter with exact probability 95%, unlike the ones based on asymptotic results.

Next, the number of data points were increased to $n = 400$, still with $m = 100$ and $q = 5$, and the confidence regions in Fig. 2 were obtained. As can be seen, the differences between the various regions get smaller, and the SPS confidence region concentrates around the true parameter as n increases.

Finally, the effect of using $\|\cdot\|_1$ or $\|\cdot\|_\infty$ norms, instead of $\|\cdot\|_2$, was considered. Recalling that the SPS regions are exact for each norm, the theory can be used to establish a precise confidence for each construction. The obtained confidence regions are illustrated in Fig. 3.

B. Choice of m and q

The probability of the SPS confidence region is $1 - q/m$ and hence there are many choices of q and m that give the same probability. Based on experience, selecting q and m too low increases the stochastic volatility in the SPS construction, and,

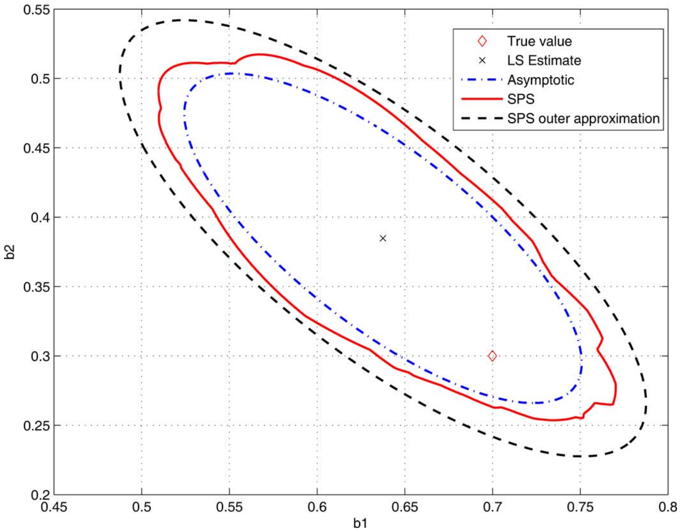

Fig. 1. 95% confidence regions, $n = 25$, $m = 100$.

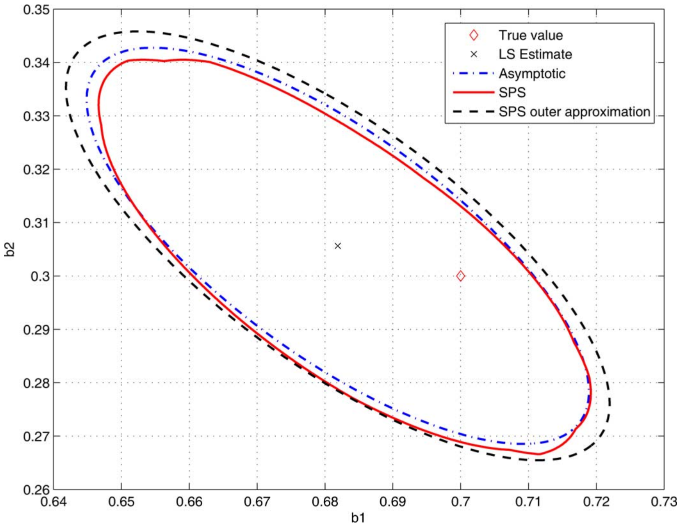

Fig. 2. 95% confidence regions, $n = 400$, $m = 100$.

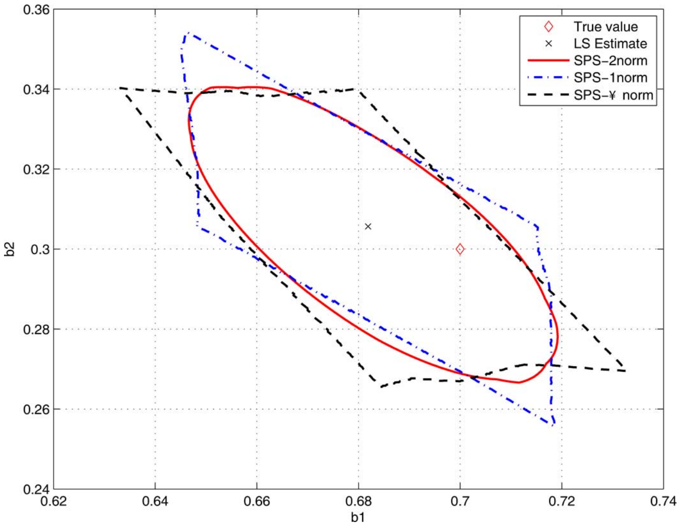

Fig. 3. 95% confidence regions using various norms, $n = 400$, $m = 100$.

TABLE IV
AVERAGE AREA, $n = 25$

m	20	60	100	200	400	600
average area	0.1041	0.0837	0.0806	0.0788	0.0778	0.0777

consequently, the average area of the confidence regions tends to be larger. However, a saturation effect occurs so that pushing q and m beyond certain values has no practical benefit. This is illustrated in Table IV where the average area of the 95% confidence regions based on 500 simulations of the same system as in the previous section with $n = 25$ is evaluated. As we can see, the average area decreases as m increases, but there is little reduction in the average area by increasing m beyond 200.

In all simulation examples above Laplacian noise was used which is heavy-tailed. However, very similar results were obtained with, for example, uniform and Gaussian noises [20].

C. Comparing With Exact Confidence Ellipsoids Based on the F-Distribution

In the special case that the noise $\{N_t\}$ is a sequence of i.i.d. Gaussian random variables, the quantity

$$\frac{n}{d}\frac{1}{\hat{\sigma}_n^2}(\theta^* - \hat{\theta}_n)^{\mathrm{T}} R_n (\theta^* - \hat{\theta}_n)$$

is distributed according to an $F(d, n-d)$ distribution where d is the number of parameters in θ and $\hat{\sigma}_n^2$ is given by (3).

In this case

$$\widetilde{\Theta}_n \triangleq \left\{ \theta \in \mathbb{R}^d : (\theta - \hat{\theta}_n)^{\mathrm{T}} R_n (\theta - \hat{\theta}_n) \leq \frac{\mu d \hat{\sigma}_n^2}{n} \right\}$$

contains the true parameter value with exact probability $F_F(\mu)$ where $F_F(\mu)$ is the cumulative distribution of an $F(d, n-d)$ distributed random variable.

SPS constructs exact confidence regions under much weaker conditions, and even when the noise is i.i.d. Gaussian we do not loose much since the SPS confidence regions are comparable to those obtained using the F-distribution as the following results show.

We consider the same system as in the previous section with $n = 25$ data points. This time $\{N_t\}$ was a sequence of i.i.d. zero mean Gaussian random variables with variance 0.1. Fig. 4 shows the 95% confidence regions we obtained in four simulation trials, and similar plots with $n = 200$ are shown in Fig. 5. Table V gives the average area of the confidence ellipsoids based on 1000 Monte Carlo simulations with $n = 25$ and $n = 200$. The average area increases by 20% ($n = 25$) and 6% ($n = 200$), when the prior information about the noise is reduced from i.i.d. Gaussian to independent and symmetrically distributed.

D. Undermodelling

The true data generating system is now given by

$$Y_t = b_1^* U_{t-1} + b_2^* U_{t-2} + b_3^* U_{t-3} + N_t,$$

where $b_1^* = 0.7$, $b_2^* = 0.3$ and $b_3^* = 0.21$ are the true system parameters. N_t and U_t are as in Section VII-A.

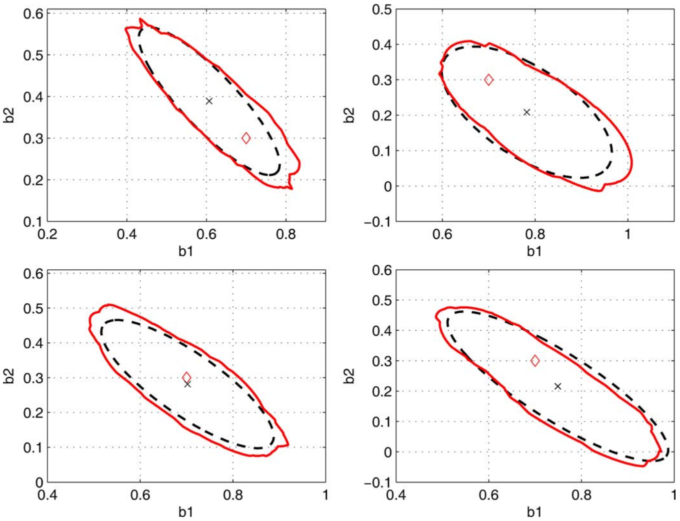

Fig. 4. 95% confidence regions, $n = 25$, $m = 100$. The solid line gives the SPS region. The dashed line gives the confidence ellipsoid based on the F-distribution.

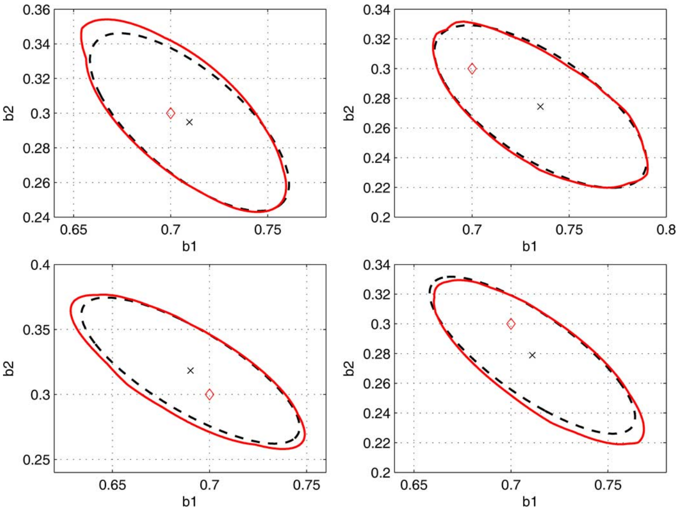

Fig. 5. 95% confidence regions, $n = 200$, $m = 100$. The solid line gives the SPS region. The dashed line gives the confidence ellipsoid based on the F-distribution.

TABLE V
AVERAGE AREA.

	SPS	F-distribution
$n = 25$	0.07876	0.065658
$n = 200$	0.00689	0.00650

The model class is still the class of all second order FIR systems with predictors

$$\hat{Y}_t(\theta) = b_1 U_{t-1} + b_2 U_{t-2}.$$

In this case the model class is not rich enough to contain the true system. The asymptotic least squares estimate as the number of data points tends to infinity is the value of $\theta = [b_1, b_2]^{\mathrm{T}}$ such that $E(Y_t - \hat{Y}_t(\theta))^2$ is minimised. For the input signal used these values are $\hat{b}_1^* = b_1^* = 0.7$ and $\hat{b}_2^* = b_2^* + 0.75 b_3^* = 0.4575$. A Monte Carlo simulation with 1000000 run found that the 95% confidence region contained the asymptotic LS estimate with empirical probability 0.9509 with $n = 25, m = 100$ and $q = 5$, which shows that in this example SPS exhibits robustness with respect to undermodelling. A typical result is shown in Fig. 6.

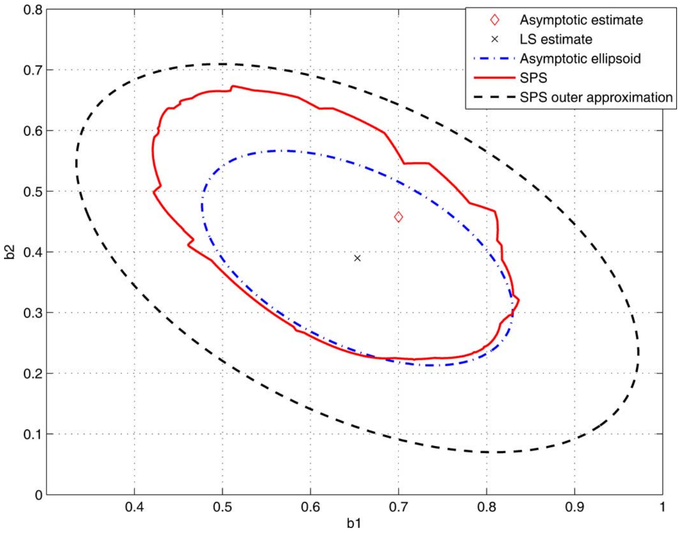

Fig. 6. 95% confidence regions, $n = 25$. The true system is a third order system, while the model is second order.

TABLE VI
EMPIRICAL PROBABILITIES BASED ON 10^6 MONTE CARLO SIMULATIONS

SPS	Block SPS	Asymptotic theory
0.888	0.944	0.883

Assumptions on the Noise are Not Satisfied

In this example, the assumptions on the noise are not satisfied. The system is the same as in the previous sections, but the noise is now the autoregressive process

$$N_t = 0.3 N_{t-1} + \sqrt{1 - 0.3^2}\, W_t,$$

where $\{W_t\}$ is i.i.d. Gaussian with variance 0.1. $n = 200$ data points are available and the aim is as before to generate a 95% confidence region. Due to the correlation in the noise both standard SPS and asymptotic system identification theory fail to produce confidence regions with the required probability. However, by using block SPS as described in Section V-C where the random signs kept their values 1 or -1 for 10 consecutive values we got a confidence probability much closer to the desired 95% than by using standard SPS or asymptotic system identification theory as shown in Table VI.

Examples of confidence regions obtained in four simulation trials are shown in Fig. 7. This demonstrates that block SPS works well also when the assumptions on the noise are not satisfied, and although the results are not precisely guaranteed anymore, it still gives good approximations.

Further, we see that we do not loose too much by using block SPS instead of the standard SPS when the assumptions on the noise are satisfied. Table VII shows the average area of the 95% confidence sets based on 1000 Monte Carlo simulations when

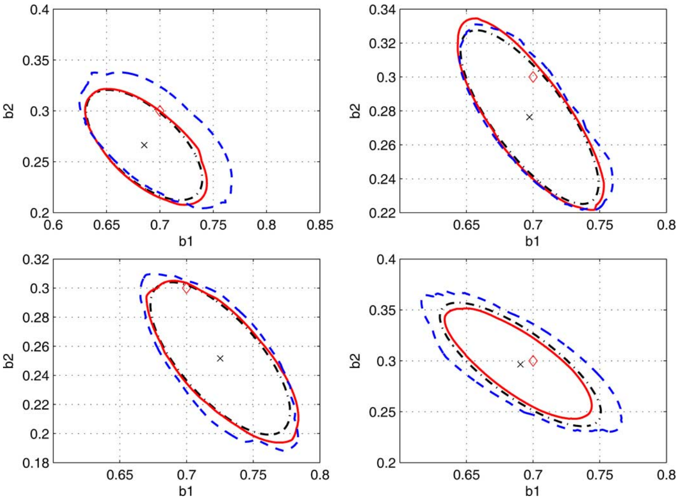

Fig. 7. 95% confidence regions, $n = 200$, $m = 100$. The solid line gives the standard SPS region and the dashed line gives the block SPS region. The dash dotted line gives the confidence ellipsoid based on asymptotic system identification theory.

TABLE VII
AVERAGE AREA, $n = 200$

SPS	Block SPS
0.00682	0.00743

the noise was i.i.d. Gaussian with variance 0.1. The average area only increased with 8.9%.

E. Higher Order System

In this experiment, the data generating system is an eight order FIR system, that is,

$$Y_t = b_1^* U_{t-1} + b_2^* U_{t-2} + b_3^* U_{t-3} + b_4^* U_{t-4} + b_5^* U_{t-5} \\ + b_6^* U_{t-6} + b_7^* U_{t-7} + b_8^* U_{t-8} + N_t,$$

with $\theta^* = [0.7, 0.3, 0.21, 0.2, 0.15, 0.25, 0.1, 0.05]^{\mathrm{T}}$. Processes $\{U_t\}$ and $\{N_t\}$ are the same as in Section VII-A.

We ran 1000 Monte Carlo simulations with $n = 200$, 800 and 3200, and computed the ellipsoidal over-bound for the 95% confidence region using $m = 100$ and $q = 5$.

The computational time for computing a single ellipsoidal over-bound with 3200 data points was around 23 seconds on a standard laptop using Yalmip and SDPT3, showing that the computational burden of the approximation is quite modest.

Table VIII gives the relative increase per dimension of the ellipsoidal over-bound as compared with the ellipsoid of the asymptotic theory. The confidence ellipsoids based on asymptotic theory are smaller than the ones based on the ellipsoidal over-bound, but the difference gets smaller as n increases, and the probability is guaranteed using SPS while it is not when using asymptotic theory.

VIII. SUMMARY AND CONCLUSION

In this paper a new system identification method called *Sign-Perturbed Sums* (SPS) has been introduced. SPS allows the construction of guaranteed non-asymptotic confidence regions, which are built around the least-squares estimate and contain the true system parameter with a user-chosen exact probability for any finite data set. SPS works under mild statistical assumptions on the system noise, and it is *non-conservative*, i.e., its confidence probability is exact. In addition, it was shown that the SPS confidence regions are *star convex* with the LS estimate as a star center.

TABLE VIII
VOLUMES. 8TH ORDER SYSTEM

Data points	Relative increase per dimension of the ellipsoidal over-bound
$n = 200$	1.78
$n = 800$	1.34
$n = 3200$	1.17

Evaluating whether a given parameter value θ belongs to the SPS confidence region is a task that can be carried out at low computational cost. This makes SPS an effective method to apply when only a finite number of candidate θ values have to be tested. On the other hand, finding the precise boundary of the SPS set can be computationally demanding in general. In order to overcome this issue, an algorithm has been introduced that provides an outer approximation of the SPS region in the form of an ellipsoid. It was demonstrated that such over-bound can be efficiently computed by convex programming methods.

Simulation experiments demonstrated that the SPS method works well, and that the confidence regions have similar size and shape as the heuristic ellipsoids of the asymptotic theory or the exact ellipsoids based on the F-distribution when the noise is i.i.d. Gaussian.

In this paper we assumed that the regressors are deterministic. While it is easy to generalize our results to the case of random regressors that are independent of the noise, the extension to the case where the regressors can depend on the noise terms is non-trivial. Generalizing the method to that case is of high practical importance. We leave this to further work, noting that some preliminary results addressing this issue were presented in earlier conference papers [18], [19].

APPENDIX A
PROOF OF THEOREM 1: EXACT CONFIDENCE

We begin with a definition and some lemmas.

Definition 1: Let $Z_1, \ldots, Z_k$ be a finite collection of random variables and $\succ$ a strict total order. If for all permutations $i_1, \ldots, i_k$ of indices $1, \ldots, k$ we have

$$\mathbb{P}(Z_{i_k} \succ Z_{i_{k-1}} \succ \ldots \succ Z_{i_1}) = \frac{1}{k!},$$

then we call $\{Z_i\}$ uniformly ordered w.r.t. order $\succ$.

Lemma 1: Let $\alpha, \beta_1, \ldots, \beta_k$ be i.i.d. random signs, then the random variables $\alpha, \alpha \cdot \beta_1, \ldots, \alpha \cdot \beta_k$ are i.i.d. random signs.

Proof: Let $c_0, c_1, \ldots, c_k$ be a fixed vector of signs, i.e., $c_i \in \{-1, 1\}$. Then, we have

$$\begin{aligned}
&\mathbb{P}(\alpha = c_0,\, \alpha\beta_1 = c_1,\, \ldots,\, \alpha\beta_k = c_k) \\
&\quad = \mathbb{P}(\alpha = c_0,\, \beta_1 = c_0 c_1,\, \ldots,\, \beta_k = c_0 c_k) \\
&\quad = \mathbb{P}(\alpha = c_0)\, \mathbb{P}(\beta_1 = c_0 c_1) \ldots\, \mathbb{P}(\beta_k = c_0 c_k) \\
&\quad = \mathbb{P}(\alpha = c_0)\, \mathbb{P}(\alpha\beta_1 = c_1) \ldots\, \mathbb{P}(\alpha\beta_k = c_k),
\end{aligned}$$

where we have used that the original collection was independent, and that $\alpha\beta_i$ and β_i has the same probability mass function, i.e., for all signs $a, b \in \{-1, 1\}$: $\mathbb{P}(\beta_i = b) = \mathbb{P}(\alpha\beta_i = a) = 1/2$. □

Lemma 2: 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 we have $\mathbb{P}(\, g(x, Y) \in A\,) = p$, for all (constant) $x \in \mathbb{R}^d$, then we also have $\mathbb{P}(\, g(X, Y) \in A\,) = p$.

Proof: Define $\mathbb{I}_A$ as follows

$$\mathbb{I}_A(x, y) \triangleq \begin{cases} 1 & \text{if } g(x, y) \in A, \\ 0 & \text{otherwise,} \end{cases}$$

which is the indicator function of the event that $g(X, Y) \in A$. Now, let us define function $i_A : \mathbb{R}^d \to \mathbb{R}$ as $i_A(x) \triangleq \mathbb{E}\,[\,\mathbb{I}_A(x, Y)\,]$, where $x \in \mathbb{R}^d$ is a constant, therefore, $i_A(x)$ is a number (non-random). We know that for all $x \in \mathbb{R}^d$ we have $i_A(x) = \mathbb{P}(\, g(x, Y) \in A\,) = p$. Then, by applying the properties of the conditional expectation [30], we have that

$$\begin{aligned} \mathbb{P}(\, g(X, Y) \in A\,) &= \mathbb{E}\,[\,\mathbb{I}_A(X, Y)\,] \\ = \mathbb{E}\,[\,\mathbb{E}\,[\,\mathbb{I}_A(X, Y) \,|\, X\,]\,] &= \mathbb{E}\,[\,i_A(X)\,] = \mathbb{E}\,[\,p\,] = p, \end{aligned}$$

which completes the proof of the lemma. □

The following lemma highlights an important property of the $\succ_\pi$ relation that was introduced in Section IV.

Lemma 3: Let $Z_1, \ldots, Z_k$ be real-valued, i.i.d. random variables. Then, they are uniformly ordered w.r.t. $\succ_\pi$.

Proof: Since $\succ_\pi$ is a total order which resolves ties, there is a *unique* ordering for all realizations $Z_1(\omega), \ldots, Z_k(\omega)$ and $\pi(\omega)$. Therefore, the events $Z_{i_k} \succ_\pi \ldots \succ_\pi Z_{i_1}$ define a complete system of events. There are $k!$ such orderings, thus, in order to complete the proof we need to show that each such ordering has the same probability. Let us select two orderings $Z_{i_k} \succ_\pi \ldots \succ_\pi Z_{i_1}$ and $Z_{j_k} \succ_\pi \ldots \succ_\pi Z_{j_1}$, which have probabilities, p_i and p_j, respectively. We will show that $p_i = p_j$. First, we define some new random variables $Z'_{i_1} \triangleq Z_{j_1}, \ldots, Z'_{i_k} \triangleq Z_{j_k}$. Then, $Z'_{i_k} \succ_\pi \ldots \succ_\pi Z'_{i_1}$ has the same probability as $Z_{j_k} \succ_\pi \ldots \succ_\pi Z_{j_1}$, since the corresponding variables are the same and π is completely *symmetric* with respect to the indices. Because $\{Z_k\}$ are i.i.d., $Z'_1, \ldots, Z'_k$ has the same distribution as $Z_1, \ldots, Z_k$. Then, $Z'_{i_k} \succ_\pi \ldots \succ_\pi Z'_{i_1}$ must have the same probability as $Z_{i_k} \succ_\pi \ldots \succ_\pi Z_{i_1}$, namely p_i. Thus, $p_j = p_i$. □

Proof of Theorem 1: By construction, parameter θ^* is in the confidence region if $\mathcal{R}(\theta^*) \leq m - q$. This means that $\|S_0(\theta^*)\|^2$ takes one of the positions $1, \ldots, m-q$ in the ascending order (w.r.t. $\succ_\pi$) of variables $\{\|S_i(\theta^*)\|^2\}$. We are going to prove that the $\{\|S_i(\theta^*)\|^2\}$ are *uniformly ordered,* hence $\|S_0(\theta^*)\|^2$ takes each position in the ordering with probability $1/m$, thus its rank is at most $m - q$ with probability $1 - q/m$.

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

$$S_i(\theta^*) = R_n^{-\frac{1}{2}} \frac{1}{n} \sum_{t=1}^{n} \alpha_{i,t}\, \varphi_t N_t,$$

for all $i \in \{0, \ldots, m-1\}$, where $\alpha_{0,t} \triangleq 1$, $t \in \{1, \ldots, n\}$.

Therefore, all the $S_i(\cdot)$ functions depend on the perturbed noise sequence, $\{\alpha_{i,t} N_t\}$, via the *same* function for all i, which we denote by $S(\alpha_{i,1} N_1, \ldots, \alpha_{i,n} N_n) \triangleq S_i(\theta^*)$.

Since each N_t is symmetric, we know that $sign(N_t)$ and $|N_t|$ are independent. Then, for all i and t, we introduce $\gamma_{i,t} \triangleq \alpha_{i,t}\, sign(N_t)$. Using that $\{\alpha_{i,t}\}$ are i.i.d. random signs, independent of the other random elements, $\{N_t\}$ are independent, and applying Lemma 1, it follows that $\{\gamma_{i,t}\}$ are not only independent of $\{|N_t|\}$, but also i.i.d. random signs.

After fixing a *realization* of $\{|N_t|\}$, called $\{v_t\}$, we define the real-valued variables $\{Z_i\}$ by

$$Z_i \triangleq \|S(\gamma_{i,1} v_1, \ldots, \gamma_{i,n} v_n)\|^2.$$

We know that, if the same (measurable) function is applied to each element of an i.i.d. sample, then the result will also be i.i.d.. Therefore, the $\{Z_i\}$ are i.i.d. random variables. Consequently, Lemma 3 can be applied to show that $\{Z_i\}$ are in fact uniformly ordered with respect to relation $\succ_\pi$.

So far we have proved the theorem assuming that the absolute values of the noises are constant, namely, the uniform ordering property was achieved by fixing a realization of $\{|N_t|\}$. However, the probabilities obtained are *independent of the particular realization* of $\{|N_t|\}$, hence, Lemma 2 can be applied to relax fixing the realization (i.e., in Lemma 2, X plays the role of $\{|N_t|\}$ and Y incorporates the other random variables), and obtain the unconditional uniform ordering property of $\{\|S_i(\theta^*)\|^2\}$, from which the theorem follows. □

Appendix B
Proof of Theorem 2: Star Convexity

Let $\Phi \triangleq [\,\varphi_1, \ldots, \varphi_n\,]^{\mathrm{T}}$, $D_i \triangleq 1/n \cdot Diag(\alpha_{i,1}, \ldots, \alpha_{i,n})$, $i \in \{1, \ldots, m-1\}$, where $Diag(\cdot)$ is the diagonal matrix with its arguments on the main diagonal, and $Q_i \triangleq \Phi^{\mathrm{T}} D_i \Phi$. Notice that $\|S_i(\theta)\|^2$ and $\|S_0(\theta)\|^2$ are quadratic forms with Hessian $Q_i R_n^{-1} Q_i$ and R_n, respectively. The following lemma relates these Hessians.

Lemma 4: For all i, we have $R_n \succeq Q_i R_n^{-1} Q_i$ in the Löwner partial ordering, i.e., $R_n - Q_i R_n^{-1} Q_i$ is positive semidefinite.

Proof: Since $R_n \succ 0$, a Schur complement argument, [31], shows that the lemma statement is equivalent to the positive semidefiniteness of the matrix

$$B_i \triangleq \begin{bmatrix} R_n & Q_i \\ Q_i & R_n \end{bmatrix} = \begin{bmatrix} \frac{1}{n}\Phi^{\mathrm{T}}\Phi & \Phi^{\mathrm{T}} D_i \Phi \\ \Phi^{\mathrm{T}} D_i \Phi & \frac{1}{n}\Phi^{\mathrm{T}}\Phi \end{bmatrix}.$$

Matrix B_i can be decomposed as

$$B_i = \begin{bmatrix} \Phi^{\mathrm{T}} & 0 \\ 0 & \Phi^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} \frac{1}{n} I & D_i \\ D_i & \frac{1}{n} I \end{bmatrix} \begin{bmatrix} \Phi & 0 \\ 0 & \Phi \end{bmatrix},$$

so that B_i is positive semidefinite if and only if the middle matrix is [32]. Resorting again to a Schur complement argument, the middle matrix is positive semidefinite if $\frac{1}{n} I - n\, D_i D_i$ is positive semidefinite, which is clearly true since this latter matrix is in fact the zero matrix. □

Introduce now the set

$$\mathcal{E}_i \triangleq \{\, \theta \in \mathbb{R}^d : \|S_0(\theta)\|^2 \le \|S_i(\theta)\|^2 \,\}.$$

Since in Lemma 4 we have proven that the Hessian of $\|S_0(\theta)\|^2$ is no smaller than that of $\|S_i(\theta)\|^2$, it follows that $\|S_0(\theta)\|^2 - \|S_i(\theta)\|^2$ is a convex function, and $\{\, \theta \in \mathbb{R}^d : \|S_0(\theta)\|^2 - \|S_i(\theta)\|^2 \le 0 \,\} = \mathcal{E}_i$ is a convex set. Moreover, $\hat{\theta}_n \in \mathcal{E}_i$ since $\|S_0(\hat{\theta}_n)\|^2 = 0$. Likewise, one can prove that set

$$\bar{\mathcal{E}}_i \triangleq \{\, \theta \in \mathbb{R}^d : \|S_0(\theta)\|^2 < \|S_i(\theta)\|^2 \,\}$$

is convex or empty. Moreover, when it is not empty, it certainly contains $\hat{\theta}_n$. Indeed, $\|S_0(\hat{\theta}_n)\|^2 = 0$, so that for $\hat{\theta}_n$ not to be in $\bar{\mathcal{E}}_i$ it must be that $\|S_i(\hat{\theta}_n)\|^2 = 0$ too. In addition, since $\|S_0(\theta)\|^2$ and $\|S_i(\theta)\|^2$ are quadratic forms, also their derivatives in $\hat{\theta}_n$ must be zero. Now, if by contradiction we assume that $\bar{\mathcal{E}}_i \neq \emptyset$, then there is a point $\bar{\theta} \in \bar{\mathcal{E}}_i$ and it holds that $\|S_0(\bar{\theta})\|^2 < \|S_i(\bar{\theta})\|^2$. Over the line segment connecting $\hat{\theta}_n$ to $\bar{\theta}$, function $\|S_i(\theta)\|^2 - \|S_0(\theta)\|^2$ then grows from 0 to a positive value, and since it moves away from $\hat{\theta}_n$ with zero slope, it must have in some point in between $\hat{\theta}_n$ to $\bar{\theta}$ a positive curvature, a fact that contradicts Lemma 4.

We are now ready to establish the result in Theorem 2.

Let

$$\mathcal{E}_i^{\pi} \triangleq \{\, \theta \in \mathbb{R}^d : \|S_0(\theta)\|^2 \prec_{\pi} \|S_i(\theta)\|^2 \,\}.$$

Notice that, depending on π, the set $\mathcal{E}_i^{\pi}$ is either $\mathcal{E}_i$ or $\bar{\mathcal{E}}_i$. Therefore, $\mathcal{E}_i^{\pi}$ is either empty or it is a convex set containing $\hat{\theta}_n$. Next, note that the SPS confidence region $\widehat{\Theta}_n$ can be written as

$$\widehat{\Theta}_n = \bigcup_{\substack{\mathcal{I} \subseteq \mathcal{M} \\ |\mathcal{I}| = q}} \bigcap_{i \in \mathcal{I}} \mathcal{E}_i^{\pi},$$

where $\mathcal{M} = \{1, \ldots, m-1\}$ and $|\cdot|$ denotes cardinality. To prove this note that if we take a finite index set of size q, $\{i_1, \ldots, i_q\}$, then the intersection of the sets $\mathcal{E}_{i_j}^{\pi}$, $j \in \{1, \ldots, q\}$, contains all the parameter values θ for which $\|S_0(\theta)\|^2$ is less than (w.r.t. $\prec_{\pi}$) all the functions $\|S_i(\theta)\|^2$ that have indexes from the index set $\{i_1, \ldots, i_q\}$. When union is taken over all possible index sets, one obtains the set of parameter values θ for which $\|S_0(\theta)\|^2$ is less than (w.r.t. $\prec_{\pi}$) at least q other functions $\|S_i(\theta)\|^2$, which is exactly the definition of the SPS confidence region. Finally, the theorem claim follows since unions and intersections of star convex sets having a common star center are themselves star convex with the same center. □

References

[1] L. Ljung, *System Identification: Theory for the User*, 2nd ed. Upper Saddle River, NJ, USA: Prentice-Hall, 1999.

[2] L. Ljung and T. Glad, *Modeling of Dynamic Systems*. Upper Saddle River, NJ, USA: Prentice Hall, 1994.

[3] L. Ljung, “Perspectives on system identification,” *Annu. Rev. Contr.*, vol. 34, no. 1, pp. 1–12, 2010.

[4] T. Söderström and P. Stoica, *System Identification*. Hertfordshire, U.K.: Prentice Hall Int., 1989.

[5] M. Gevers, “A personal view of the development of system identification: A 30-year journey through an exciting field,” *IEEE Control Syst.*, vol. 26, no. 6, pp. 93–105, Dec. 2006.

[6] M. C. Campi and E. Weyer, “Guaranteed non-asymptotic confidence regions in system identification,” *Automatica*, vol. 41, pp. 1751–1764, 2005.

[7] M. C. Campi and E. Weyer, “Non-asymptotic confidence sets for the parameters of linear transfer functions,” *IEEE Trans. Autom. Control*, vol. 55, no. 12, pp. 2708–2720, Dec. 2010.

[8] M. C. Campi, S. Ko, and E. Weyer, “Non-asymptotic confidence regions for model parameters in the presence of unmodelled dynamics,” *Automatica*, vol. 45, pp. 2175–2186, 2009.

[9] M. Dalai, E. Weyer, and M. C. Campi, “Parameter identification for non-linear systems: Guaranteed confidence regions through LSCR,” *Automatica*, vol. 43, pp. 1418–1425, 2007.

[10] E. Weyer, S. Ko, and M. C. Campi, “Finite sample properties of system identification with quantized output data,” in *Proc. 48th IEEE Conf. Decision Contr.*, 2009, pp. 1532–1537.

[11] E. Weyer, S. Ko, and M. C. Campi, “A randomised subsampling method for change detection,” in *Proc. SafeProcess’09 Conf.*, Barcelona, Spain, 2009.

[12] J. Schoukens, Y. Rolain, G. Vandersteen, and R. Pintelon, “Study of small data set efficiency losses in system identification: The FIR case,” in *Proc. 11th IFAC Int. Workshop Adapt. Learn. Contr. Signal Process.*, 2013, pp. 68–73.

[13] F. Dabbene, M. Sznaier, and R. Tempo, “Probabilistic optimal estimation with uniformly distributed noise,” *IEEE Trans. Autom. Control*, vol. 59, no. 8, pp. 2113–2127, Aug. 2014.

[14] J. C. Aguero, C. R. Rojas, H. Hjalmarsson, and G. C. Goodwin, “Accuracy of linear multiple-input multiple-output (MIMO) models obtained by maximum likelihood estimation,” *Automatica*, vol. 48, pp. 632–637, 2012.

[15] O. N. Granichin, “The nonasymptotic confidence set for parameters of a linear control object under an arbitrary external disturbance,” *Autom. Remote Contr.*, vol. 73, no. 1, pp. 20–30, 2012.

[16] A. J. den Dekker, X. Bombois, and P. M. J. Van den Hof, “Finite sample confidence regions for parameters in prediction error identification using output error,” in *Proc. IFAC World Congr.*, 2008, pp. 5024–5029.

[17] H. Hjalmarsson and B. Ninness, “Least-squares estimation of a class of frequency functions: A finite sample variance expression,” *Automatica*, vol. 42, pp. 589–600, 2006.

[18] B. C. Csáji, M. C. Campi, and E. Weyer, “Non-asymptotic confidence regions for the least-squares estimate,” in *Proc. 16th IFAC Symp. Syst. Identificat.*, 2012, pp. 227–232.

[19] B. C. Csáji, M. C. Campi, and E. Weyer, “A method for constructing exact finite-sample confidence regions for general linear systems,” in *Proc. 51st IEEE Conf. Decision Contr.*, 2012, pp. 7321–7326.

[20] E. Weyer, B. C. Csáji, and M. C. Campi, “Guaranteed non-asymptotic confidence ellipsoids for FIR systems,” in *Proc. 52st IEEE Conf. Decision Contr.*, 2013, pp. 7162–7167.

[21] P. M. J. Van den Hof, P. S. C. Heuberger, and J. Bokor, “System identification with generalized orthonormal basis functions,” *Automatica*, vol. 31, pp. 1821–1834, 1995.

[22] P. Van den Hof and B. Ninness, “System identification with generalized orthonormal basis functions,” in *Modelling and Identification with Rational Orthogonal Basis Functions*. London, U.K.: Springer-Verlag, 2005, pp. 61–102.

[23] B. Efron and R. Tibshirani, *An Introduction to the Bootstrap*. New York, NY, USA: Chapman & Hall, 1993.

[24] R. Davidson and J. G. MacKinnon, “Bootstrap methods in econometrics,” in *Palgrave Handbook of Econometrics*. London, U.K.: Palgrave Macmillan, 2006, vol. 1, pp. 812–38.

[25] L. Godfrey, *Bootstrap Tests for Regression Models*. New York, NY, USA: Palgrave Macmillan, 2009.

[26] M. Kieffer and E. Walter, “Guaranteed characterization of exact non-asymptotic confidence regions as defined by LSCR and SPS,” *Automatica*, vol. 49, pp. 507–512, 2013.

[27] M. Kieffer and E. Walter, “Guaranteed characterization of exact non-asymptotic confidence regions in nonlinear parameter estimation nonlinear control systems,” in *Proc. 9th IFAC Symp. Nonlinear Contr. Syst.*, 2013, pp. 56–61.

[28] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2009.

[29] Löfberg, "Yalmip: A toolbox for modeling and optimization in MATLAB," in *Proc. CACSD Conf.*, Taipei, Taiwan, 2004.

[30] O. Kallenberg, *Foundations of Modern Probability*, 2nd ed. New York, NY, USA: Springer-Verlag, 2001.

[31] F. Zhang, *The Schur Complement and its Applications*. Berlin, Germany: Springer-Verlag, 2005.

[32] F. Zhang, *Matrix Theory: Basic Results and Techniques*. Berlin, Germany: Springer-Verlag, 2011.

Balázs Csanád Csáji (M'13) is a Senior Research Fellow at MTA SZTAKI, The Institute for Computer Science and Control of the Hungarian Academy of Sciences, Budapest, Hungary. He received a Ph.D. degree in computer science (2008) from the Eötvös Loránd University (ELTE), Budapest, Hungary. Previously, he received Master's degrees in computer science combined with mathematics (2001) as well as in philosophy (2006), also from ELTE.

During his studies he spent semesters and internships at the Eindhoven University of Technology, Netherlands (2001), the British Telecom, UK (2002), and the Johannes Kepler University, Linz, Austria (2003). He was a Postdoctoral Researcher at the Université catholique de Louvain, Louvain-la-Neuve, Belgium (2008–2009), and a Research Fellow at the University of Melbourne, Australia (2009–2012).

Dr. Csáji has received a number of awards for his achievements including the Best Ph.D. Student and Institute Awards of MTA SZTAKI, the Junior Award for Research Excellence from the Hungarian Academy of Sciences, the Best Paper Award at the 6th International Workshop on Emergent Synthesis (Tokyo, Japan), and a Discovery Early Career Researcher Award (DECRA) from the Australian Research Council. His main research interests revolve around stochastic models and related statistical problems, especially in the fields of machine learning and system identification.

Marco Claudio Campi (A'98–SM'08–F'12) is Professor of Automatic Control at the University of Brescia, Italy.

In 1988, he received the Doctor degree in electronic engineering from the Politecnico di Milano, Milano, Italy. From 1988 to 1989, he was a Lecturer at the Department of Electrical Engineering of the Politecnico di Milano. From 1989 to 1992, he was a Research Fellow at the Centro di Teoria dei Sistemi of the National Research Council (CNR) in Milano and, in 1992, he joined the University of Brescia, Brescia, Italy. He has held visiting and teaching appointments at the Australian National University, Canberra, Australia; the University of Illinois at Urbana-Champaign, USA; the Centre for Artificial Intelligence and Robotics, Bangalore, India; the University of Melbourne, Australia; the Kyoto University, Japan. The research interests of Marco Campi include: system identification, inductive methods, randomized algorithms, robust optimization, and learning theory.

Since 2011, Marco Campi is the chair of the Technical Committee IFAC on Modeling, Identification and Signal Processing (MISP). He has been in various capacities on the Editorial Board of Automatica, Systems and Control Letters and the European Journal of Control. Marco Campi is a recipient of the "Giorgio Quazza" prize, and, in 2008, he received the IEEE CSS George S. Axelby outstanding paper award for the article The Scenario Approach to Robust Control Design. He has delivered plenary and semi-plenary addresses at major conferences including SYSID, MTNS, and CDC, moreover he has been a distinguished lecturer of the Control Systems Society. Marco Campi is a Fellow of IEEE, a member of IFAC, and a member of SIDRA.

Erik Weyer (S'92–M'93) received the Siv. Ing. degree in 1988 and the Ph.D. in 1993, both from the Norwegian Institute of Technology, Trondheim, Norway.

From 1994 to 1996 he was a Research Fellow at the University of Queensland, and since 1997 he has been with the Department of Electrical and Electronic Engineering, the University of Melbourne, where he is currently an Associate Professor. He has held visiting positions at the University of Brescia, Italy, the Technical University of Vienna, Austria, and Politecnico di Milano, Italy. His research interests are in the areas of system identification and control, with particular emphasis on finite sample properties of system identification methods, and modelling and control of irrigation channels and rivers.

From 2010 to 2012 he was an Associate Editor of IEEE Transactions of Automatic Control.