Finite-Sample Identification of Linear Regression Models With Residual-Permuted Sums

Szabolcs Szentpéteri and Balázs Csanád Csáji, *Member, IEEE*

***Abstract*—This letter studies a distribution-free, finite-sample data perturbation (DP) method, the Residual-Permuted Sums (RPS), which is an alternative of the Sign-Perturbed Sums (SPS) algorithm, to construct confidence regions. While SPS assumes independent (but potentially time-varying) noise terms which are symmetric about zero, RPS gets rid of the symmetricity assumption, but assumes i.i.d. noises. The main idea is that RPS permutes the residuals instead of perturbing their signs. This letter introduces RPS in a flexible way, which allows various design-choices. RPS has exact finite sample coverage probabilities and we provide the first proof that these permutation-based confidence regions are uniformly strongly consistent under general assumptions. This means that the RPS regions almost surely shrink around the true parameters as the sample size increases. The ellipsoidal outer-approximation (EOA) of SPS is also extended to RPS, and the effectiveness of RPS is validated by numerical experiments, as well.**

***Index Terms*—Identification, linear systems, randomized algorithms.**

I. Introduction

Estimating dynamical systems based on empirical data is a fundamental problem in system identification, machine learning and statistics. Classical results in the aforementioned areas, such as prediction error methods, typically provide point estimates with asymptotically guaranteed confidence regions [1]. However, in practical problems, where the robustness of the solution is a crucial aspect, confidence sets with finite sample guarantees are highly desirable. Due to these reasons, in the recent years, significant emphasis was given to the non-asymptotic theory of system identification [2].

Manuscript received 8 March 2024; revised 10 May 2024; accepted 2 June 2024. Date of publication 11 June 2024; date of current version 28 June 2024. This work was supported in part by the European Union within the Framework of the National Laboratory for Autonomous Systems under Grant RRF-2.3.1-21-2022-00002, and in part by the National Research, Development and Innovation Office (NRDIO), Hungary, under Grant TKP2021-NKTA-01. Recommended by Senior Editor T. Oomen. *(Corresponding author: Balázs Csanád Csáji.)*

Szabolcs Szentpéteri is with the Institute for Computer Science and Control (SZTAKI), Hungarian Research Network (HUN-REN), 1111 Budapest, Hungary (e-mail: szabolcs.szentpeteri@sztaki.hu).

Balázs Csanád Csáji is with the Institute for Computer Science and Control (SZTAKI), Hungarian Research Network (HUN-REN), 1111 Budapest, Hungary, and also with the Department of Probability Theory and Statistics, Eötvös Loránd University, 1053 Budapest, Hungary (e-mail: balazs.csaji@sztaki.hu).

Digital Object Identifier 10.1109/LCSYS.2024.3412856

A possible approach to build confidence regions with finite sample guarantees for i.i.d. samples is to utilize randomized hypothesis tests, e.g., Monte Carlo or bootstrap methods [3]. For linear regression problems, a random permutation based test was presented in [4], however, it builds on asymptotic approximations, hence it lacks finite sample guarantees.

An important identification method that uses the ideas behind randomized tests and can build exact confidence regions for finite sample sizes with distribution-free guarantees is Sign-Perturbed Sums (SPS) [5]. Later SPS was generalized, and Data Perturbation (DP) methods were introduced [6].

The core idea of SPS is to perturb the signs of the residuals in the normal equations, assuming that the measurement noises are independent and symmetric about zero. Then, based on the rank statistics computed from these perturbed quantities, SPS builds a confidence region around the least-squares (LS) estimate [5]. DP methods generalize this idea in a way that they allows different types of perturbations depending on the characteristics of the observation noises. As a prime example, a permutation-based DP method was introduced in [6], where the residuals are permuted instead of sign-perturbed. This requires exchangeable noises, but symmetricity is not needed.

Another approach to relax the symmetricity assumption of SPS is LAD-SPS which builds confidence regions around the least-absolute-deviation (LAD) estimate under the assumption that the conditional medians of the noises are zero [7]. In all of the above mentioned algorithms, the coverage probability of the true parameters can be exactly guaranteed for any finite sample size, and the confidence set is given by an indicator function that can be queried at any parameter vector.

For linear regression problems, the (uniform) strong consistency of SPS was proven in [8], which means that the SPS regions almost surely shrink around the true system parameters as the sample size increases. In addition, a compact representation of SPS confidence sets, given by an ellipsoidal outer-approximation (EOA) algorithm, was proposed in [5].

As so far consistency was only proven for SPS, it remained an open question whether other types of DP methods, such as permutation-based constructions, are consistent, as well. In this letter, we propose the Residual-Permuted Sums (RPS) method for linear regression problems, which is a generalization of the original permutation-based approach of [6]. We rigorously prove the (uniform) strong consistency of RPS under general statistical assumptions, and also extend the EOA of SPS to RPS. Finally, simulation experiments are presented that demonstrate the effectiveness of RPS by comparing it with SPS and asymptotic confidence ellipsoids.

II. Problem Setting

This section specifies the addressed linear regression problem and introduces our main assumptions on the system.

A. Data Generation

Consider the following linear regression problem

$$Y_t \doteq \varphi_t^{\mathrm{T}} \theta^* + W_t, \tag{1}$$

for $t \in [n] \doteq \{1, \dots, n\}$, where φ_t is a d-dimensional random regressor, Y_t is the scalar output, W_t is the (random) scalar noise and θ^* is the d-dimensional (constant) true parameter. We are given a sample of size n which consists of regressor vectors (inputs) $\varphi_1, \dots, \varphi_n$ and outputs $Y_1, \dots, Y_n$.

B. Assumptions

Our assumptions on the noises and the regressors are

A1: The noise terms $\{W_t\}$ are independent and identically distributed (i.i.d.) with finite fourth moments: $\mathbb{E}[W_0^4] < \infty$.

Note that these assumptions on the noises are very mild, as most strong consistency results assume independence and require moment conditions from the noise terms. Also note that unlike SPS [5], the noises do not have to be symmetric about zero, nor do they need zero mean; however their i.i.d. nature is essential to ensure exact coverage probabilities.

A2: Regressors $\{\varphi_t\}$ have uniformly bounded fourth moments, $\forall t : \mathbb{E}[\|\varphi_t\|^4] \leq \kappa < \infty$, and they are ℓ-independent: $\forall t : \varphi_t \perp\!\!\!\perp \{\varphi_k\}_{|t-k| \geq \ell}$, where "$\perp\!\!\!\perp$" denotes independence.

A consequence of ℓ-independent regressors is that our analysis covers FIR and Generalised FIR models [1]. Also note that the independence of $\{\varphi_t\}$ and $\{W_t\}$ is not assumed.

C. Co-Regressor Construction

We introduce a co-regressor based construction that is used by the proposed RPS algorithm. The motivation for using co-regressors is to cover various design-choices. We denote the d-dimensional random co-regressor by ψ_t, and assume that

A3: The co-regressor vectors $\{\psi_t\}$ are ℓ-independent with uniformly bounded fourth moments, i.e., for every t, we have $\mathbb{E}[\|\psi_t\|^4] \leq \kappa < \infty$, furthermore $\{\psi_t\}$ is independent of the noise sequence $\{W_t\}$, and for every $t : \mathbb{E}[\psi_t W_t] = 0$.

A4: We have

$$\forall i, j : |i - j| \geq \ell : \psi_i \perp\!\!\!\perp \varphi_j \quad \text{and} \quad \mathbb{E}\big[\psi_i \varphi_j^{\mathrm{T}}\big] = 0, \tag{2}$$

$$\forall k, l : \forall i, j : |i - j| < \ell : \mathbb{E}\big[\psi_{i,k}^4 \varphi_{j,l}^4\big] \leq \kappa < \infty, \tag{3}$$

furthermore, the condition below holds almost surely

$$\lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \psi_t \varphi_t^{\mathrm{T}} \doteq \lim_{n\to\infty} V_n = V \succ 0, \tag{4}$$

where "$\succ 0$" denotes positive definiteness.

A5: There are user-chosen (random) matrices $\{R_n\}$, such that $\{R_n\} \perp\!\!\!\perp \{W_t\}$, $\{R_n\}$ are positive semidefinite, and there is a positive definite R, such that almost surely

$$\lim_{n\to\infty} R_n = R \succ 0, \tag{5}$$

Matrices $\{R_n\}$ are only introduced to allow the reshaping of RPS regions, however, $R_n = I_n$ is also a valid choice.

There are several possible design choices for co-regressors that can satisfy A3 and A4, here we list some of them. In case $\mathbb{E}[\varphi_t]$ is known, and the regressors are independent of the noises, one can simply use $\psi_t \doteq \varphi_t - \mathbb{E}[\varphi_t]$ as a co-regressor. In case $\mathbb{E}[\varphi_t]$ is unknown, one can replace it with an estimate, $\psi_t \doteq \varphi_t - \zeta_t$, where $\mathbb{E}[\zeta_t] = \mathbb{E}[\varphi_t]$, for example, an independent copy of φ_t can be used as ζ_t. Finally, one can also apply $\psi_t \doteq f(\varphi_t)$, where f is a suitable function.

Let us illustrate this latter option with signed-regressors often used in adaptive filtering [9]. Let $d = 1$ and $Y_t \doteq \varphi_t \theta^* + |\varphi_t| N_t$, where $\varphi_t \perp\!\!\!\perp N_t$ and $\varphi_t, N_t \sim \mathcal{N}(0, 1)$. Then, $W_t \doteq |\varphi_t| N_t$ is not independent of φ_t. However, as $\text{sign}(\varphi_t)$ and $|\varphi_t|$ are independent, we can use $\psi_t \doteq \text{sign}(\varphi_t)$ as a co-regressor, since this ensures the independence of ψ_t and W_t.

Algorithm 1 Pseudocode: RPS-Initialization (p)

1: Given a (rational) confidence probability $p \in (0, 1)$, set integers $m > q > 0$ such that $p = 1 - q/m$.
2: Choose a positive semidefinite matrix R_n and find its principal square root $R_n^{1/2}$, i.e., the p.s.d. matrix with
$$R_n^{1/2} R_n^{1/2\mathrm{T}} = R_n.$$
3: For all $i \in [n]$, generate (independent) uniform random permutations $\sigma_{i,n}$ of $[n]$, that is, each of the $n!$ possible permutations has probability $1/(n!)$ to be selected.
4: Generate a random permutation π of $\{0, \dots, m-1\}$ uniformly, i.e., each permutation has probability $1/(m!)$

III. The Residual-Permuted Sums Algorithm

In this section, we introduce the Residual-Permuted Sums algorithm. The method is a generalization of the permutation-based hypothesis test proposed in [6]. It consists of two parts, in the first part the main parameters and the random permutations are computed, while the second part decides whether a given parameter θ is included in the confidence region. The first part is given by Algorithm 1 and the second is presented by Algorithm 2. Using this construction, the p-level RPS confidence region can be defined as follows

$$\mathcal{C}_{p,n} \doteq \{\theta \in \mathbb{R}^d \text{: RPS-Indicator } (\theta) = 1\}. \tag{6}$$

A. Exact Coverage of RPS Confidence Regions

The exact coverage probability of the permutation-based variant of SPS was proved in [6], for the case of deterministic regressors. This result can be extended to cover the exact confidence of RPS regions under our assumptions:

Theorem 1: Assuming $\{W_t\}$ are i.i.d., and R_n, $\{\psi_t\}$ are independent of $\{W_t\}$, the coverage probability of the constructed confidence region $\mathcal{C}_{p,n}$ is exactly p, that is

$$\mathbb{P}\big(\theta^* \in \mathcal{C}_{p,n}\big) = 1 - \frac{q}{m} = p. \tag{7}$$

Proof: The exact coverage of RPS regions can be proven very similarly to the proofs in [5], [6], [10]: by showing that $\{\|S_i(\theta^*)\|^2\}_{i=0}^{m-1}$ are exchangeable. First, for deterministic R_n and $\{\psi_t\}$, it can be shown that $\{\|S_i(\theta^*)\|^2\}_{i=0}^{m-1}$ are conditionally i.i.d. (thus also exchangeable) w.r.t. the ordered noises, i.e., the σ-algebra generated by $(W_{(1)}, \dots, W_{(n)})$. This can be generalized for random R_n and $\{\psi_t\}$ by using that they are independent of $\{W_t\}$ and the law of total expectation, i.e., by conditioning on R_n and $\{\psi_t\}$, as well. ■

Algorithm 2 Pseudocode: RPS-Indicator (θ)

1: Compute the prediction errors for θ: for $t \in [n]$ let
$$\varepsilon_t(\theta) \doteq Y_t - \varphi_t^{\mathrm{T}}\theta.$$
2: Evaluate for $i \in [m-1]$ the following functions:
$$S_0(\theta) \doteq R_n^{-\frac{1}{2}} \frac{1}{n} \sum_{t=1}^{n} \psi_t \varepsilon_t(\theta),$$
$$S_i(\theta) \doteq R_n^{-\frac{1}{2}} \frac{1}{n} \sum_{t=1}^{n} \psi_t \varepsilon_{\sigma_{i,n}(t)}(\theta).$$
3: Compute the rank of $\|S_0(\theta)\|^2$ among $\{\|S_i(\theta)\|^2\}$:
$$\mathcal{R}(\theta) \doteq \left[1 + \sum_{i=1}^{m-1} \mathbb{I}\Big(\|S_0(\theta)\|^2 \succ_\pi \|S_i(\theta)\|^2 \Big) \right],$$
where "$\succ_\pi$" is "$>$" with random tie-breaking, i.e., $\|S_k(\theta)\|^2 \succ_\pi \|S_j(\theta)\|^2$ if and only if $(\|S_k(\theta)\|^2 > \|S_j(\theta)\|^2) \vee (\|S_k(\theta)\|^2 = \|S_j(\theta)\|^2 \wedge \pi(k) > \pi(j))$.
4: Return 1 if $\mathcal{R}(\theta) \leq m - q$, otherwise return 0.

IV. Strong Consistency of RPS Regions

In this section, we present one of the main contributions of this letter, the proof that the confidence regions generated by RPS are strongly consistent. First, we prove a lemma that plays a key part in the proof of the main theorem.

A. Permutation Lemma

The next lemma is a strong law of large numbers (SLLN) for randomly permuted sequences. The main idea behind its proof is to extend Cantelli's SLLN [11]. Note that SLLN type theorems for permuted sequences in the literature mainly focus on a single exchangeable sequence, however, we have a new permutation for every n, i.e., a double-indexed sequence.

Lemma 1: Let $\{X_i\}$ and $\{Y_i\}$ be sequences of ℓ-independent random variables with $\mathbb{E}[X_i^a Y_j^b] \leq \kappa_0 < \infty$ for all $|j-i| < \ell$ and $a, b \in \mathbb{N}_0$ satisfying $0 \leq a, b \leq 4$ and ($a = b$ or $a+b \leq 4$). Furthermore, for $|j-i| \geq \ell$ let X_i and Y_j be independent and $\mathbb{E}[X_i Y_j] = 0$. Let $\{\sigma_n\}$ be independent, where σ_n is a uniform random permutation of $[n]$. Then, we have
$$\frac{1}{n}\sum_{i=1}^{n} X_i Y_{\sigma_n(i)} \xrightarrow{\text{a.s.}} 0 \qquad (\text{as } n \to \infty). \tag{8}$$

Proof: For every $s \in \mathbb{N}$, let $J_s \doteq \{s, s+\ell, s+2\ell, \dots\}$. Then, for each s, $\{X_j\}_{j \in J_s}$ is an independent sequence. Let $I_s \doteq I_s^n \doteq J_s \cap [n]$, hence if $s \leq \ell$, $\lfloor n/\ell \rfloor \leq |I_s| \leq \lceil n/\ell \rceil$.

By the (first) Borell-Cantelli lemma [12], (8) holds if
$$\sum_{n=1}^{\infty} \mathbb{P}\left\{ \left| \frac{1}{n}\sum_{i=1}^{n} X_i Y_{\sigma_n(i)} \right| \geq \varepsilon \right\}$$
$$\leq \sum_{n=1}^{\infty} \mathbb{P}\left\{ \left| \frac{1}{n}\sum_{i \in I_1} X_i Y_{\sigma_n(i)} \right| + \dots + \left| \frac{1}{n}\sum_{i \in I_\ell} X_i Y_{\sigma_n(i)} \right| \geq \varepsilon \right\}$$
$$\leq \sum_{n=1}^{\infty} \sum_{s=1}^{\ell} \mathbb{P}\left\{ \left| \frac{1}{n}\sum_{i \in I_s} X_i Y_{\sigma_n(i)} \right| \geq \frac{\varepsilon}{\ell} \right\} < \infty, \tag{9}$$
for any $\varepsilon > 0$, where we applied the triangular inequality and the union bound. By using the (generalized) Chebyshev inequality, the convergence of the series in (9) follows from
$$\frac{\ell^4}{\varepsilon^4} \sum_{n=1}^{\infty} \sum_{s=1}^{\ell} \mathbb{E}\left| \frac{1}{n} \sum_{i \in I_s} X_i Y_{\sigma_n(i)} \right|^4 < \infty. \tag{10}$$

Therefore, our goal will be to show (10). Let us expand
$$\begin{aligned}\left| \sum_{i \in I_s} X_i Y_{\sigma_n(i)} \right|^4 &= \sum_{i \in I_s} X_i^4 Y_{\sigma_n(i)}^4 \\ &+ 6 \sum_{i,j \in I_s, i<j} X_i^2 Y_{\sigma_n(i)}^2 X_j^2 Y_{\sigma_n(j)}^2 \\ &+ 12 \sum_{i,j,k \in I_s, i<j<k} X_i^2 Y_{\sigma_n(i)}^2 X_j Y_{\sigma_n(j)} X_k Y_{\sigma_n(k)} \\ &+ 24 \sum_{i,j,k,l \in I_s, i<j<k<l} X_i Y_{\sigma_n(i)} X_j Y_{\sigma_n(j)} X_k Y_{\sigma_n(k)} X_l Y_{\sigma_n(l)} \\ &+ 8 \sum_{i,j \in I_s, i \neq j} X_i^3 Y_{\sigma_n(i)}^3 X_j Y_{\sigma_n(j)}.\end{aligned} \tag{11}$$

Now, we will take the expectation of (11), term by term. In the first two terms, for any permutation, none of the summed expected values are zero, and each one of them can be upper bounded by a corresponding power of κ_0 using our independence and moment assumptions. Note that $a = b = 0$ is also allowed which ensures that $\kappa_0 \geq 1$. Then,
$$\begin{aligned}&\sum_{s=1}^{\ell} \sum_{i \in I_s} \mathbb{E}\Big[X_i^4 Y_{\sigma_n(i)}^4\Big] \leq \ell \lceil n/\ell \rceil \kappa_0^2 = \mathcal{O}(n), \\ &\sum_{s=1}^{\ell} \sum_{i,j \in I_s, i<j} \mathbb{E}\Big[X_i^2 Y_{\sigma_n(i)}^2 X_j^2 Y_{\sigma_n(j)}^2\Big] \leq \ell \binom{\lceil \frac{n}{\ell} \rceil}{2} \kappa_0^4 = \mathcal{O}\Big(n^2\Big).\end{aligned} \tag{12}$$

In order to upper bound the expectation of the third term, we introduce the ℓ-neighbourhood of index i as $N(i) \doteq \{j : |i-j| < \ell\}$. If $j \notin N(i)$, then $\mathbb{E}[X_i Y_j] = 0$ and $X_i \perp\!\!\!\perp Y_j$, and consequently, the summed expectations in the third term can be nonzero only: *a)* if the ℓ-neighbourhoods of j and k each contains at least one of $\sigma_n(i), \sigma_n(j), \sigma_n(k)$ or *b)* $\sigma_n(j)$ and $\sigma_n(k)$ are in the ℓ-neighbourhood of i or $\sigma_n(i)$, or if they in the ℓ-neighbourhood of each other. More precisely, by using the law of total probability and introducing the events $A \doteq A_i(j,k) \doteq \{\sigma_n(i) \in N(j) \cap N(k)\}$ and $B_j(k) \doteq B_j(i,k) \doteq \{\sigma_n(j) \in N(i) \cup N(\sigma_n(i)) \cup N(\sigma_n(k))\}$,
$$\begin{aligned}&\sum_{s=1}^{\ell} \sum_{\substack{i,j,k \in I_s \\ i<j<k}} \mathbb{E}\big[X_i^2 Y_{\sigma_n(i)}^2 X_j Y_{\sigma_n(j)} X_k Y_{\sigma_n(k)} \,\big|A \vee (B_j(k) \wedge \\ &B_k(j))\big]\mathbb{P}\big(A \vee (B_j(k) \wedge B_k(j))\big) \leq \ell \binom{\lceil \frac{n}{\ell} \rceil}{3} \kappa_0^4 \cdot \\ &\cdot \left[\left(\frac{\ell\,(n-1)\,(n-2)}{n(n-1)(n-2)} \right) + \left(\frac{4\ell\, 6\ell\,(n-2)}{n(n-1)(n-2)} \right) \right] = \mathcal{O}\Big(n^2\Big),\end{aligned} \tag{13}$$
where the upper bound κ_0^4 on the conditional expectation follows from our moment assumptions and the repeated application of the Cauchy–Schwarz inequality. The argument

behind the definition of event A is to cover the possibilities given in the description of point a). Note that it can be that $N(j) \cap N(k) \neq \emptyset$ due to the construction of I_s, and our definition of A gives the least constraints on σ_n to ensure that the corresponding summand has a nonzero expectation. As event A has the highest probability among those events that ensure this, $\mathbb{P}(A) = \mathcal{O}(1/n)$ can be used as an upper bound for the probabilities of all events described in a).

The expectation of the fourth term can be upper bounded very similarly to the third one, using case separation. For the sake of brevity, we only provide the final bound

$$\sum_{s=1}^{\ell} \sum_{\substack{i,j,k,l \in I_s \\ i<j<k<l}} \mathbb{E}\big[X_i Y_{\sigma_n(i)} X_j Y_{\sigma_n(j)} X_k Y_{\sigma_n(k)} X_l Y_{\sigma_n(l)}\big] \leq \ell \binom{\lceil \frac{n}{\ell} \rceil}{4} \kappa_0^5 \Bigg[\left(\frac{\ell^2 (n-2)(n-3)}{n(n-1)(n-2)(n-3)} \right) + \left(\frac{(n-5\ell)(n-7\ell) 2\ell \, 4\ell}{n(n-1)(n-2)(n-3)} \right) \Bigg] = \mathcal{O}\left(n^2\right), \quad (14)$$

where we can repeatedly use Hölder's inequality to upper bound the expectation terms in (14) by κ_0^5.

The expectation of the fifth term of (15) can be upper bounded similarly to the second term, that is by using that at most n^2 expectations are summed, therefore

$$\sum_{s=1}^{\ell} \sum_{i,j \in I_s, i<j} \mathbb{E}\Big[X_i^3 Y_{\sigma_n(i)}^3 X_j Y_{\sigma_n(j)}\Big] = \mathcal{O}\left(n^2\right). \quad (15)$$

Putting the five expectations together, we get

$$\sum_{s=1}^{\ell} \mathbb{E}\left(\sum_{i=1}^{n} X_i Y_{\sigma_n(i)} \right)^4 = \mathcal{O}\left(n^2\right). \quad (16)$$

Consequently, we can conclude that:

$$\frac{\ell^4}{\varepsilon^4} \sum_{n=1}^{\infty} \sum_{s=1}^{\ell} \mathbb{E} \left| \frac{1}{n} \sum_{i=1}^{n} X_i Y_{\sigma_n(i)} \right|^4 \leq c \sum_{n=1}^{\infty} \frac{1}{n^2} < \infty, \quad (17)$$

for some constant c, which completes the proof. ■

B. Strong Consistency

In the following, we state and prove our main theorem about the strong consistency of RPS. Our proof takes several ideas from the strong consistency proof of IV-SPS [10]. A major difference is that in case of sign-perturbations (such as in IV-SPS), standard SLLN type results can be applied, while for the case of RPS, we must use Lemma 1.

Theorem 2: Assuming A1-A5, $\forall\, \varepsilon > 0$, we have

$$\mathbb{P}\left(\bigcup_{k=1}^{\infty} \bigcap_{n=k}^{\infty} \left\{ \mathcal{C}_{p,n} \subseteq \mathcal{B}_\varepsilon\left(\theta^*\right) \right\} \right) = 1, \quad (18)$$

where $\mathcal{B}_\varepsilon(\theta^*) \doteq \{\theta \in \mathbb{R}^d : \|\theta - \theta^*\| \leq \varepsilon\}$.

Proof: In the first part of the proof we are going to prove that for any "false" parameter vector $\theta' \neq \theta^*$, we have

$$\left\| S_0\left(\theta'\right) \right\|^2 \xrightarrow{a.s.} \left\| R^{-\frac{1}{2}} V \left(\theta^* - \theta'\right) \right\|^2 > 0, \quad (19)$$

while, for $i \neq 0$, we have

$$\left\| S_i\left(\theta'\right) \right\|^2 \xrightarrow{a.s.} 0, \quad (20)$$

as $n \to \infty$. Recall the definitions of V and R from A4, A5. As a consequence of (19) and (20), as n grows, the rank $\mathcal{R}(\theta')$ of $\left\| S_0(\theta') \right\|^2$ will be eventually m, therefore θ' will be (a.s.) excluded from the confidence region, as $n \to \infty$.

As a first step, we reformulate $S_0(\theta')$ and $S_i(\theta')$,

$$\begin{aligned} S_0\left(\theta'\right) &= R_n^{-\frac{1}{2}} \frac{1}{n} \sum_{t=1}^{n} \psi_t \varepsilon_t\left(\theta'\right) \\ &= R_n^{-\frac{1}{2}} \frac{1}{n} \sum_{t=1}^{n} \psi_t \left(\varphi_t^{\mathrm{T}} \theta^* + W_t - \varphi_t^{\mathrm{T}} \theta'\right) \\ &= R_n^{-\frac{1}{2}} \frac{1}{n} \sum_{t=1}^{n} \psi_t \varphi_t^{\mathrm{T}} \tilde{\theta} + \psi_t W_t, \end{aligned} \quad (21)$$

$$\begin{aligned} S_i\left(\theta'\right) &= R_n^{-\frac{1}{2}} \frac{1}{n} \sum_{t=1}^{n} \psi_t \varepsilon_{\sigma_{i,n}(t)}\left(\theta'\right) \\ &= R_n^{-\frac{1}{2}} \frac{1}{n} \sum_{t=1}^{n} \psi_t \left(\varphi_{\sigma_{i,n}(t)}^{\mathrm{T}} \theta^* + W_{\sigma_{i,n}(t)} - \varphi_{\sigma_{i,n}(t)}^{\mathrm{T}} \theta'\right) \\ &= R_n^{-\frac{1}{2}} \frac{1}{n} \sum_{t=1}^{n} \psi_t \varphi_{\sigma_{i,n}(t)}^{\mathrm{T}} \tilde{\theta} + \psi_t W_{\sigma_{i,n}(t)}, \end{aligned} \quad (22)$$

where $\tilde{\theta} \doteq \theta^* - \theta'$. We will examine the four terms from (21) and (22) separately. In case of the reference sum, we first assume that $\ell = 1$, then generalize our result to arbitrary ℓ.

i) Reference sum first term: Using A4 and A5 it holds that

$$\lim_{n \to \infty} R_n^{-\frac{1}{2}} \frac{1}{n} \sum_{t=1}^{n} \psi_t \varphi_t^{\mathrm{T}} \tilde{\theta} = R^{-\frac{1}{2}} V \tilde{\theta} \quad \text{(a.s.)}. \quad (23)$$

In the following, we will prove almost sure convergence to zero for every sum, therefore the term $R_n^{-\frac{1}{2}}$ can be omitted from the sums as $R_n^{-\frac{1}{2}} \xrightarrow{a.s.} R^{-\frac{1}{2}}$ (A5).

ii) Reference sum second term: Using Cantelli's SLLN element-wise it can be proven that

$$\lim_{n \to \infty} \frac{1}{n} \sum_{t=1}^{n} \psi_t W_t = 0 \quad \text{(a.s.)}, \quad (24)$$

since $\{\psi_{t,j} W_t\}$ is an independent sequence, $\mathbb{E}[\psi_{t,j} W_t] = 0$ and $\mathbb{E}[\psi_{t,j}^4 W_t^4] < \infty$ (A1, A3).

We conclude that, as $n \to \infty$, we have

$$\left\| S_0\left(\theta'\right) \right\|^2 \xrightarrow{a.s.} \left\| R^{-\frac{1}{2}} V \tilde{\theta} \right\|^2 > 0. \quad (25)$$

For an arbitrary ℓ, the same construction can be used as in the proof of Lemma 1, to decompose the sequence into subsequences $\{\psi_k \varphi_k\}_{k \in J_s}$ of independent variables, e.g.,

$$\frac{1}{n} \sum_{t=1}^{n} \psi_t \varphi_t^{\mathrm{T}} \tilde{\theta} = \frac{\lceil \frac{n}{\ell} \rceil}{n} \sum_{s=1}^{\ell} \left(\frac{1}{\lceil \frac{n}{\ell} \rceil} \sum_{t \in I_s^n} \psi_t \varphi_t^{\mathrm{T}} \tilde{\theta} \right), \quad (26)$$

where I_s^n and J_s are defined above (26), hence $\lfloor n/\ell \rfloor \leq \left| I_s^n \right| \leq \lceil n/\ell \rceil$. As we decomposed $\{\psi_k \varphi_k\}$ into the sum of ℓ subseries, and each such subseries converges (a.s.) based on our previous arguments, the original series converges (a.s.) to the sum of the corresponding limits. Thus, (25) is ensured for any ℓ.

iii) Permuted sum first term: Notice that the summed elements in the first term of $S_i(\theta')$ (22) do not form an independent sequence, therefore well-known SLLN results cannot be applied to prove a.s. convergence. To prove that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \psi_t \varphi_{\sigma_{i,n}(t)}^{\mathrm{T}} \tilde{\theta} = 0 \quad \text{(a.s.)}, \tag{27}$$

Lemma 1 can be applied element-wise. Note that the conditions of the lemma follows from A2, A3 and A4 using that $\mathbb{E}|X|^p < \infty \Rightarrow \mathbb{E}|X|^q < \infty$ if $q \le p$, and one can use the Cauchy–Schwarz inequality to show that the expectations of the cross-products are also bounded. Finally, the maximum of the obtained bounds can serve as κ_0.

iv) Permuted sum second term: The summed terms form an ℓ-independent sequence in this case, however for every sample size n, a new sum is generated. Nevertheless, we can apply Lemma 1 element-wise again to prove that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \psi_t W_{\sigma_{i,n}(t)} = 0 \quad \text{(a.s.)}, \tag{28}$$

since A1 and A3 satisfies the assumptions of Lemma 1, for the same reasons as in the previous sum (*iii*).

From the previous derivations, we can conclude that

$$\left\| S_i(\theta') \right\|^2 \xrightarrow{a.s.} 0, \tag{29}$$

as $n \to \infty$, for each $i \in \{1, \dots, m-1\}$.

Now, we prove that the confidence region converges to θ^* uniformly, not just pointwise. Let us introduce

$$\Phi_n \doteq \begin{bmatrix} \varphi_1^{\mathrm{T}} \\ \varphi_2^{\mathrm{T}} \\ \vdots \\ \varphi_n^{\mathrm{T}} \end{bmatrix}, \qquad \Psi_n \doteq \begin{bmatrix} \psi_1^{\mathrm{T}} \\ \psi_2^{\mathrm{T}} \\ \vdots \\ \psi_n^{\mathrm{T}} \end{bmatrix}, \qquad w_n \doteq \begin{bmatrix} W_1 \\ W_2 \\ \vdots \\ W_n \end{bmatrix},$$

$$Q_{i,n} \doteq \frac{1}{n} \sum_{t=1}^{n} \psi_t \varphi_{\sigma_{i,n}(t)}^{\mathrm{T}} = \frac{1}{n} \Psi_n^{\mathrm{T}} P_{i,n} \Phi_n, \tag{30}$$

where $P_{i,n}$ is a permutation matrix corresponding to $\sigma_{i,n}$. Using the definition from (4), it holds that $V_n = \frac{1}{n} \Psi_n^{\mathrm{T}} \Phi_n$.

Our previous results showed that for every i, $\|S_i(\theta')\|^2$ converges (a.s.). As a consequence, for each realization $\omega \in \Omega$ (from an event with probability one, where $(\Omega, \mathcal{F}, \mathbb{P})$ is the underlying probability space), and for each $\delta > 0$, there exists a $N(\omega) > 0$ such that for $n \ge N$ and $i \ne 0$,

$$\left\| R_n^{-\frac{1}{2}} V_n - R^{-\frac{1}{2}} V \right\| \le \delta, \; \left\| \tfrac{1}{n} R_n^{-\frac{1}{2}} \Psi_n^{\mathrm{T}} w_n \right\| \le \delta, \tag{31}$$

$$\left\| R_n^{-\frac{1}{2}} Q_{i,n} \right\| \le \delta, \; \left\| \tfrac{1}{n} R_n^{-\frac{1}{2}} \Psi_n^{\mathrm{T}} P_{i,n} w_n \right\| \le \delta. \tag{32}$$

Then, using similar reformulations as in the proof of [10, Th. 2], for all $n \ge N$, we have

$$\left\| S_0(\theta') \right\| \ge \sigma_{\min}\left(R^{-\frac{1}{2}} V\right) \|\tilde{\theta}\| - \delta \|\tilde{\theta}\| - \delta, \tag{33}$$

where $U_\sigma \Sigma V_\sigma^{\mathrm{T}}$ is the SVD decomposition of $R^{-\frac{1}{2}} V$ and $\sigma_{\min}(\cdot)$ denotes the smallest singular value. We also have

$$\begin{aligned} \left\| S_i(\theta') \right\| &= \left\| R_n^{-\frac{1}{2}} Q_{i,n} \tilde{\theta} + \tfrac{1}{n} R_n^{-\frac{1}{2}} \Psi_n^{\mathrm{T}} P_{i,n} w_n \right\| \le \\ &\left\| R_n^{-\frac{1}{2}} Q_{i,n} \right\| \|\tilde{\theta}\| + \left\| \tfrac{1}{n} R_n^{-\frac{1}{2}} \Psi_n^{\mathrm{T}} P_{i,n} w_n \right\| \le \delta \|\tilde{\theta}\| + \delta. \end{aligned} \tag{34}$$

Therefore, we have $\left\| S_i(\theta') \right\| < \left\| S_0(\theta') \right\|, \forall \theta'$ that satisfy

$$\delta \|\tilde{\theta}\| + \delta < \sigma_{\min}\left(R^{-\frac{1}{2}} V\right) \|\tilde{\theta}\| - \delta \|\tilde{\theta}\| - \delta, \tag{35}$$

which can be reformulated as

$$\mu(\delta) \doteq \frac{2\delta}{\sigma_{\min}\left(R^{-\frac{1}{2}} V\right) - 2\delta} < \|\tilde{\theta}\|, \tag{36}$$

therefore, those θ' for which $\mu(\delta) < \|\theta^* - \theta'\|$ are not in the confidence region $\mathcal{C}_{p,n}$, for $n \ge N$. For any $\varepsilon > 0$, by setting $\delta = (\varepsilon \sigma_{\min}(R^{-\frac{1}{2}} V))/(2 + 2\varepsilon)$, we have $\mathcal{C}_{p,n} \subseteq \mathcal{B}_\varepsilon(\theta^*)$, therefore, the claim of the theorem follows. ■

V. Ellipsoidal Outer-Approximation

The RPS-Indicator function can decide whether a given parameter is included in the confidence region. In order to give a compact representation of the whole region, we introduce a permutation-based version of the ellipsoidal outer-approximation (EOA) method [5], [10]. The main motivation behind such constructions is that evaluating every parameter, even on a grid, to build the RPS region is computationally demanding, especially in higher dimensions. The ellipsoids are constructed in a way that they have the same shape, orientation and center as the asymptotic confidence ellipsoids, see (38), only their sizes (determined by the radius parameters) are different. However, they have finite sample guarantees. The radius computation, which is based on the ordering of the residual-permuted sums, and the construction of the ellipsoid can be derived similarly as for IV-SPS [10], since the sign-perturbations can be simply replaced by permutations.

First, we introduce a correlation approach based estimate [1, Sec. VII.5], that will be the center of the region, as

$$\hat{\theta}_n \doteq \left(\sum_{t=1}^{n} \psi_t \varphi_t^{\mathrm{T}} \right)^{-1} \sum_{t=1}^{n} \psi_t Y_t \tag{37}$$

Then, the RPS outer-approximation can be given as

$$\mathcal{O}_{n,p} \doteq \left\{ \theta \in \mathbb{R}^d : \left\| R_n^{-\frac{1}{2}} V_n \left(\theta - \hat{\theta}_n \right) \right\|^2 \le r \right\}, \tag{38}$$

where r is the qth largest solution of the following convex semi-definite programming problems, for $i \in \{1, \dots, m-1\}$

$$\begin{aligned} \min \quad & \gamma \\ \text{s.t.} \quad & \lambda \ge 0 \\ & \begin{bmatrix} -I + \lambda A_i & \lambda b_i \\ \lambda b_i^{\mathrm{T}} & \lambda c_i + \gamma \end{bmatrix} \succeq 0. \end{aligned} \tag{39}$$

In (39), "$\succeq 0$" denotes positive semidefinitness and

$$\begin{aligned} A_i &\doteq I - R_n^{\frac{1}{2}\mathrm{T}} V_n^{-\mathrm{T}} Q_{i,n}^{\mathrm{T}} R_n^{-1} Q_{i,n} V_n^{-1} R_n^{\frac{1}{2}} \\ b_i &\doteq R_n^{\frac{1}{2}\mathrm{T}} V_n^{-\mathrm{T}} Q_{i,n}^{\mathrm{T}} R_n^{-1} \left(\xi_i - Q_{i,n} \hat{\theta}_n \right) \\ c_i &\doteq -\xi_i^{\mathrm{T}} R_n^{-1} \xi_i + 2 \hat{\theta}_n^{\mathrm{T}} Q_{i,n}^{\mathrm{T}} R_n^{-1} \xi_i - \hat{\theta}_n^{\mathrm{T}} Q_{i,n}^{\mathrm{T}} R_n^{-1} Q_{i,n} \hat{\theta}_n \\ \xi_i &\doteq \frac{1}{n} \sum_{t=1}^{n} \psi_t Y_{\sigma_{i,n}(t)}. \end{aligned} \tag{40}$$

As $\mathcal{O}_{n,p}$ is an outer approximation of $\mathcal{C}_{n,p}$ it follows that

$$\mathbb{P}\left(\theta^* \in \mathcal{O}_{n,p}\right) \ge 1 - \frac{q}{m} = p, \tag{41}$$

hence $\mathcal{O}_{n,p}$ is a guaranteed confidence ellipsoid for any n.

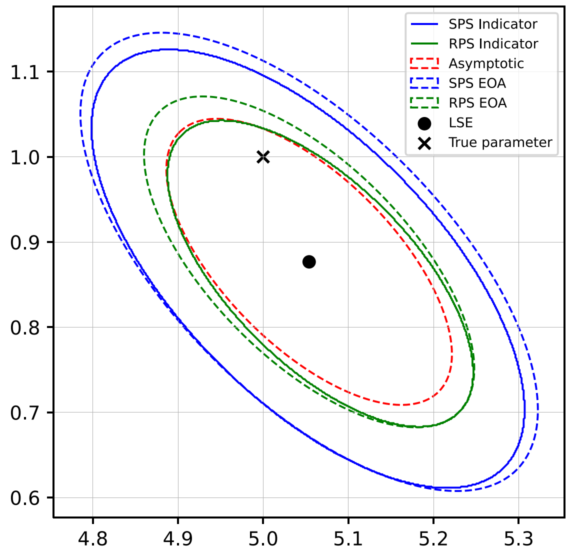

Fig. 1. Comparison of 0.9-level RPS indicator, RPS EOA, SPS indicator, SPS EOA and asymptotic confidence regions for $n = 250$.

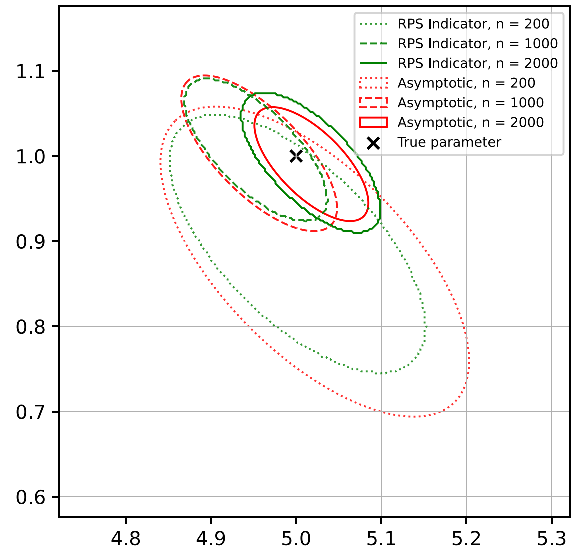

Fig. 2. Comparison of 0.9-level RPS indicator and asymptotic confidence regions for $n = 200$, $n = 1000$ and $n = 2000$.

VI. Simulation Experiments

In this section, we illustrate RPS through two numerical experiments. In the first experiment, we compared the RPS indicator and outer-approximation regions with the SPS indicator and outer-approximation regions, furthermore, with the confidence ellipsoid based on the classical asymptotic theory [1]. We consider a 2-dimensional FIR system

$$Y_t = b_1^* U_{t-1} + b_2^* U_{t-2} + W_t, \tag{42}$$

where $b_1^* = 5$, $b_2^* = 1$, $\{W_t\}$ are i.i.d. Laplacian random variables with mean 0 and variance 1, and the input is $U_t = \sum_{i=1}^{5} c_i V_{t-i+1}$, with $V_t \sim \mathcal{N}(0, 1)$ and $c_1 = 1$, $c_2 = 0.775$, $c_3 = 0.55$, $c_4 = 0.325$, $c_5 = 0.1$. From (42), the linear regression problem can be constructed as $\theta^* = [b_1^*, b_2^*]^{\mathrm{T}}$ and $\varphi_t = [U_{t-1}, U_{t-2}]^{\mathrm{T}}$. The sample size was $n = 250$. We chose $\psi_t = \varphi_t$ and $R_n = \frac{1}{n}\Phi_n^{\mathrm{T}}\Phi_n$ to make sure that the RPS and SPS regions have the same shape and orientation as the asymptotic confidence ellipsoids. The 0.9-level confidence regions, with $m = 10$ and $q = 1$ for the RPS and SPS methods, are presented in Fig. 1. It can be observed that the RPS regions are smaller than the SPS sets, and that they are about the same size as the asymptotic confidence regions. This experiment indicates that RPS can outperform SPS sample complexity wise, while having an advantage over the asymptotic region that it has finite sample coverage guarantees.

In the second experiment, we investigated the sizes of RPS regions for different sample sizes. We used the same system setting as in the first experiment, with the exception that $\{W_t\}$ was a sequence of i.i.d exponential random variables with parameter 0.5, i.e., not a symmetric distribution about zero. Fig. 2 illustrates the RPS indicator and asymptotic confidence regions for $n = 200$, $n = 1000$ and $n = 2000$. It shows that for smaller sample sizes, RPS regions have smaller sizes than asymptotic ellipsoids, but this size difference decreases as the sample size increases. Nonetheless, RPS has exact finite sample coverage guarantees, unlike the asymptotic region.

VII. Conclusion

In this letter, we have introduced the Residual-Permuted Sums (RPS) algorithm, motivated by a permutation-based DP method, as an alternative to SPS, in which the symmetricity and independence assumptions on the noises are replaced by an i.i.d. condition. RPS can construct exact, non-asymptotic, distribution-free confidence regions for the true parameters of linear regression problems. One of the main contributions of this letter is that we proved the (uniform) strong consistency of the RPS construction under general assumptions. We also demonstrated the effectiveness of RPS empirically.

References

[1] L. Ljung, *System Identification: Theory for the User*, 2nd ed. Prentice Hall, NJ, USA: Upper Saddle River, 1999.
[2] I. Ziemann, A. Tsiamis, B. Lee, Y. Jedra, N. Matni, and G. J. Pappas, “A tutorial on the non-asymptotic theory of system identification,” in *Proc. 62nd IEEE Conf. Decision Control (CDC)*, 2023, pp. 8921–8939.
[3] E. L. Lehmann and J. P. Romano, *Testing Statistical Hypotheses*, 4th ed. Cham, Switzerland: Springer, 2022.
[4] J. Nyblom, “Permutation tests in linear regression,” in *Modern Nonparametric, Robust and Multivariate Methods: Festschrift in Honour of Hannu Oja*, K. Nordhausen and S. Taskinen, Eds. Cham, Switzerland: Springer, 2015, pp. 69–90.
[5] B. Cs. Csáji, M. C. Campi, and E. Weyer, “Sign-Perturbed Sums: A new system identification approach for constructing exact non-asymptotic confidence regions in linear regression models,” *IEEE Trans. Signal Process.*, vol. 63, no. 1, pp. 169–181, Jan. 2015.
[6] S. Kolumbán, I. Vajk, and J. Schoukens, “Perturbed datasets methods for hypothesis testing and structure of corresponding confidence sets,” *Automatica*, vol. 51, pp. 326–331, Jan. 2015.
[7] A. Carè, B. Cs. Csáji, M. C. Campi, and E. Weyer, “Sign-Perturbed Sums (SPS) with asymmetric noise: Robustness analysis and robustification techniques,” in *Proc. 55th IEEE Conf. Decision Control (CDC)*, Las Vegas, NV, USA, 2016, pp. 262–267.
[8] E. Weyer, M. C. Campi, and B. Cs. Csáji, “Asymptotic properties of SPS confidence regions,” *Automatica*, vol. 81, pp. 287–294, Aug. 2017.
[9] A. H. Sayed, *Fundamentals of Adaptive Filtering*. Hoboken, NJ, USA: Wiley, 2003.
[10] V. Volpe, B. Cs. Csáji, A. Carè, E. Weyer, and M. C. Campi, “Sign-Perturbed Sums (SPS) with instrumental variables for the identification of ARX systems,” in *Proc. 54th IEEE Conf. Decis. Control (CDC)*, Osaka, Japan, 2015, pp. 2115–2120.
[11] A. Shiryaev, *Probability-2*. New York, NY, USA: Springer, 2019.
[12] A. Shiryaev, *Probability-1*. New York, NY, USA: Springer, 2016.