

Non-Asymptotic Confidence Regions for Separable Nonlinear Models

Szabolcs Szentpéteri Péter Kovács, *Senior Member, IEEE* Balázs Csanád Csáji, *Senior Member, IEEE*

Abstract—This paper studies robust uncertainty quantification for separable nonlinear least squares (SNLLS) problems. We propose two identification methods, (i) Residual-Perturbed Gradient (RPG) and (ii) Global-Fit-Gap (GFG), to construct joint distribution-free confidence regions for both the linear and nonlinear parameters. By exploiting an assumed group invariance of the noise (such as symmetry or exchangeability), we apply random perturbations (such as sign-changes or permutations) to build the regions based on resampling and ranking. We prove that both methods guarantee non-asymptotically exact coverage (type I error). Finally, we illustrate the constructions through two numerical experiments: (i) identifying generalized finite impulse response (GFIR) systems using Laguerre filters with an adjustable decay rate and (ii) fitting Gaussian radial basis function (RBF) models with variable knots and widths.

I. INTRODUCTION

Many *inverse problems* arising in machine learning, system identification, and signal and image processing can be formulated as *separable nonlinear least-squares* (SNLLS) problems [1]. In these models, the parameters can be naturally partitioned into a *nonlinear* component and a *linear* component. Typically, the nonlinear parameters form a relatively small set with direct physical interpretation, such as absorption coefficients, dominant frequencies, or phases, while the linear parameters correspond to coefficients associated with basis functions or system responses.

A widely used approach for solving SNLLS problems is the *variable projection* (VP) method [2], which reduces the dimensionality of the search space and transforms the original problem into a lower-dimensional nonlinear optimization problem that is often easier to solve and numerically more stable. VP methods are particularly attractive for a wide range of inverse problems [3], including time-series prediction using RBF-ARX models [4], frequency identification [5], blind image deconvolution, and image denoising [1].

Sz. Szentpéteri and B. Cs. Csáji were supported by the “Robust Uncertainty Quantification for Learning and Control” ADVANCED project of the National Research, Development and Innovation Office of Hungary (NKFIH), grant number 153390. P. Kovács was supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences (MTA) and the University Excellence Fund of Eötvös Loránd University (ELTE).

Sz. Szentpéteri is with the Department of Automatic Control, Lund University, Sweden. He is a member of the ELLIIT Strategic Research Area at Lund University. This work was carried out while he was with the Institute for Computer Science and Control (SZTAKI), Hungarian Research Network (HUN-REN), 13–17 Kende utca, 1111, Budapest, Hungary szabolcs.szentpeteri@control.lth.se

P. Kovács is with the Department of Numerical Analysis, Eötvös Loránd University (ELTE), 1117 Budapest, Hungary kovika@inf.elte.hu

B. Cs. Csáji is with the Institute for Computer Science and Control (SZTAKI), Hungarian Research Network (HUN-REN), 13–17 Kende u., 1111, Budapest, Hungary, and with the Department of Probability Theory and Statistics, Institute of Mathematics, Eötvös Loránd University (ELTE), 1/C Pázmány Péter sétány, 1117, Budapest, Hungary csaji@sztaki.hu

Non-smooth [6], [7] and large-scale [8] extensions have only recently been considered, further highlighting the potential of the SNLLS framework and the VP approach.

Golub and Pereyra showed that the *gradient* of the VP objective functional can be computed explicitly [2], which has attracted considerable attention from the machine learning community. In particular, the VP algorithm can be used to train feedforward neural networks [9]. Recently, this approach has been generalized to non-quadratic objectives, allowing VP-based training of neural networks for tasks including segmentation and classification [10]. The VP operator has also been used as an activation mechanism that integrates domain knowledge into network architectures through a layer-wise SNLLS formulation of the target problem [11].

VP methods are very efficient to compute *point estimates*, selecting a single model from a given class. However, for *risk-aware* solutions, these should be augmented with *confidence regions* that quantify the *uncertainty* of likely models.

Despite its *critical importance*, there are very few methods that address the issue of constructing confidence regions around VP estimates. Currently, the prevailing general solution relies on the *asymptotic distribution* of the scaled estimation error [12], [13]. However, ensuring the existence of the limiting distribution is nontrivial for complex nonlinear structures [14]. More importantly, the resulting confidence ellipsoids are only asymptotically guaranteed, providing a false sense of security in the finite-sample regime.

The *main contributions* of this paper are as follows.

- (1) We introduce two methods for building *distribution-free* confidence sets for nonlinear models, centered around VP estimates. The *Residual-Perturbed Gradient* (RPG) method is computationally efficient, but it might capture artifacts. The *Global-Fit-Gap* (GFG) method mitigates this issue, but incurs a higher computational cost.
- (2) We prove that both methods ensure confidence sets with *non-asymptotic* coverage guarantees, under only mild assumptions on the noises affecting the system.
- (3) Finally, we present two case studies: (i) identifying GFIR models for systems characterized by slowly decaying impulse responses, and (ii) estimating Gaussian-based RBF models with adjustable knots and widths.

II. PROBLEM SETTING

In this section, we present the addressed SNLLS problem.

A. Data Generation and Problem Statement

Consider the following *separable nonlinear model*:

$$Y_t \doteq \sum_{i=1}^{d_\gamma} \gamma_i^* \varphi_i(\beta^*, X_t) + W_t, \quad (1)$$

where $\gamma^* \doteq [\gamma_1^*, \dots, \gamma_{d_\gamma}^*]^T \in \mathbb{R}^{d_\gamma}$ is the true *linear* parameter, $\beta^* \in \mathbb{R}^{d_\beta}$ is the true *nonlinear* parameter, $\varphi_i(\beta^*, X_t)$ is a nonlinear basis function, $X_t \in \mathbb{R}^{d_x}$ is the input, $W_t \in \mathbb{R}$ is the random noise, and $Y_t \in \mathbb{R}$ is the output. We are given a sample of size n of inputs and outputs $\{(X_t, Y_t)\}_{t=1}^n$.

We use the stacked notation $\theta \doteq [\beta^T, \gamma^T]^T \in \mathbb{R}^{d_\theta}$, where $d_\theta = d_\gamma + d_\beta$. This convention applies equally to parameters, estimators, etc. For $k \in \mathbb{N}$, we also write $[k] \doteq \{1, \dots, k\}$.

Using the input-output data, we address the problem of constructing a confidence region $\mathcal{E}_p$ for the true parameter θ^* such that, for a prescribed confidence probability $p \in (0, 1)$,

$$\mathbb{P}(\theta^* \in \mathcal{E}_p) \geq p. \quad (2)$$

We consider *deterministic* inputs; however, our results can be straightforwardly generalized to *exogenous* inputs, where the input and the noise sequences are independent.

B. Assumptions

Our *main assumptions* on the noises affecting the system and on the nonlinear basis functions are as follows.

A1. The noise $w \doteq [W_1, \dots, W_n]^T$ is distributionally invariant w.r.t. a known compact matrix group of transformations, $(\mathcal{G}, \circ)$, where each $G \in \mathcal{G}$ acts on $\mathbb{R}^n$ and the composition $\circ$ is the group operation, i.e., $\forall G \in \mathcal{G} : G(w) \stackrel{d}{=} w$.

This assumption states that there exist known transformations that leave the *joint distribution* of the noise vector unchanged. Standard examples for which we know such transformations include *exchangeable* sequences, which in particular cover the *independent and identically distributed* (i.i.d.) case, and independent sequences that are *symmetric* about zero. In the exchangeable case, the relevant transformations are multiplications by *permutation* matrices:

$$\mathcal{G} = \{G \in \{0, 1\}^{n \times n} : G\mathbb{1} = \mathbb{1}, G^T\mathbb{1} = \mathbb{1}\}, \quad (3)$$

where $\mathbb{1}$ is the all-one vector. In the independent and symmetric case, the transformations correspond to multiplications by diagonal “*sign*” matrices, i.e., with ± 1 entries:

$$\mathcal{G} = \{\text{diag}(\alpha_1, \dots, \alpha_n) : \forall k \in [n], \alpha_k \in \{-1, 1\}\}, \quad (4)$$

where $\text{diag}(v)$ denotes a diagonal matrix with v on its main diagonal. Both groups are finite, thus compact. In both cases, the group operation is matrix multiplication, the identity is the identity matrix, and the inverse is the matrix inverse.

It is important to note that Assumption A1 is very mild: it imposes no conditions either on the *tail behavior* of the noise distribution or on the existence of its *moments*.

A2. Let $\Upsilon \subseteq \mathbb{R}^{d_\beta}$ be open. For $i \in [d_\gamma]$, the nonlinear basis functions $\varphi_i : \Upsilon \times \mathbb{R}^{d_x} \rightarrow \mathbb{R}$ are continuously differentiable w.r.t. β . Define the regression matrix $\Phi(\beta) \in \mathbb{R}^{n \times d_\gamma}$ for $t \in [n]$ and $i \in [d_\gamma]$ as $\Phi(\beta)_{t,i} \doteq \varphi_i(\beta, X_t)$. Let $J_{\beta'}(\Phi(\beta')\gamma') : \Upsilon \times \mathbb{R}^{d_\gamma} \rightarrow \mathbb{R}^{n \times d_\beta}$ be the Jacobian of $\Phi(\beta')\gamma'$ w.r.t. β' . We assume that for all $\beta \in \Upsilon$, $\gamma \in \mathbb{R}^{d_\gamma}$, and $n \geq d_\gamma + d_\beta$: $\text{rank}([J_{\beta'}(\Phi(\beta')\gamma')(\beta, \gamma), \Phi(\beta)]) = d_\gamma + d_\beta$.

This assumption ensures that the model is locally *identifiable* and that the least-squares problem, detailed in the next

section, is *well-posed*. It also guarantees that the confidence regions generated by our two proposed methods are *bounded* under suitable assumptions on the random transformations.

III. SEPARABLE NONLINEAR LEAST SQUARES

A. Least-Squares Estimate

Given the sample $\{(X_t, Y_t)\}_{t=1}^n$, the parameters γ^*, β^* can be estimated by minimizing the *least-squares* (LS) functional

$$\underset{\beta, \gamma}{\text{argmin}} \frac{1}{2} \sum_{t=1}^n \left(Y_t - \sum_{i=1}^{d_\gamma} \gamma_i \varphi_i(\beta, X_t) \right)^2. \quad (5)$$

This problem is known as the *separable nonlinear least-squares* (SNLLS) problem. By defining the residual as

$$r(\theta) \doteq r(\beta, \gamma) \doteq y - \Phi(\beta)\gamma, \quad (6)$$

where $y \doteq [Y_1, \dots, Y_n]^T$, the LS problem of (5) can be written in a compact matrix form as

$$\hat{\theta} \doteq \underset{\theta \in \mathbb{R}^{d_\theta}}{\text{argmin}} \frac{1}{2} \|r(\theta)\|^2 = \underset{\beta, \gamma}{\text{argmin}} \frac{1}{2} \|r(\beta, \gamma)\|^2. \quad (7)$$

B. Variable Projection

A well-known approach to solve the SNLLS problem is the *variable projection* (VP) algorithm [2], [3]. This method first solves the linear LS problem for a fixed β as

$$\hat{\gamma}(\beta) = \underset{\gamma}{\text{argmin}} \frac{1}{2} \|r(\beta, \gamma)\|^2 = \Phi(\beta)^\dagger y, \quad (8)$$

where $(\cdot)^\dagger$ denotes pseudoinverse (Moore–Penrose inverse). Then, it minimizes the following VP functional

$$\underset{\beta}{\text{argmin}} \frac{1}{2} \|r_2(\beta)\|^2 = \underset{\beta}{\text{argmin}} \frac{1}{2} \|(I - \Phi(\beta)\Phi(\beta)^\dagger)y\|^2, \quad (9)$$

which can be obtained by substituting $\hat{\gamma}(\beta)$ to the original functional in (7). Golub and Pereyra [2] showed that the analytical expression for the k -th column of the *Jacobian* of the reduced functional $J_{\beta'}(r_2(\beta'))(\beta)$ is given as

$$\begin{aligned} J_{\beta'}^{(k)}(r_2(\beta'))(\beta) = & - \left[P_{\Phi(\beta)}^\perp \frac{\partial \Phi(\beta)}{\partial \beta_k} \Phi(\beta)^\dagger \right. \\ & \left. + (P_{\Phi(\beta)}^\perp \frac{\partial \Phi(\beta)}{\partial \beta_k} \Phi(\beta)^\dagger)^T \right] y, \end{aligned} \quad (10)$$

where $P_{\Phi(\beta)}^\perp \doteq I - \Phi(\beta)\Phi(\beta)^\dagger$. The Jacobian matrix of the reduced residual $J_{\beta'}(r_2(\beta'))(\beta)$ plays a key role in computing the gradient and the Hessian of the VP functional, solving the optimization problem of (9) and obtaining estimate $\hat{\beta}$.

C. Gradients of the Residual

In [2], the *gradients* of $\frac{1}{2} \|r(\theta)\|^2$ are also given as:

$$\begin{aligned} \nabla_\beta \frac{1}{2} \|r(\theta)\|^2 &= \nabla_\beta \frac{1}{2} \|r(\beta, \gamma)\|^2 \\ &= -J_{\beta'}^T(\Phi(\beta')\gamma')(\beta, \gamma)r(\beta, \gamma), \end{aligned} \quad (11)$$

$$\nabla_\gamma \frac{1}{2} \|r(\theta)\|^2 = \nabla_\gamma \frac{1}{2} \|r(\beta, \gamma)\|^2 = -\Phi(\beta)^T r(\beta, \gamma), \quad (12)$$

therefore, we have

$$\nabla_\beta \frac{1}{2} \|r(\beta, \gamma)\|^2 \Big|_{(\beta^*, \gamma^*)} = -J_{\beta'}^T(\Phi(\beta')\gamma')(\beta^*, \gamma^*)w, \quad (13)$$

$$\nabla_{\gamma} \frac{1}{2} \|r(\beta, \gamma)\|^2 \Big|_{(\beta^*, \gamma^*)} = -\Phi(\beta^*)^T w, \quad (14)$$

and the Jacobian of $r(\theta) = r(\beta, \gamma)$ is

$$\begin{aligned} J(\theta) &\doteq [J_{\beta'}(r(\beta', \gamma'))(\beta, \gamma), J_{\gamma'}(r(\beta', \gamma'))(\beta, \gamma)] \\ &= -[J_{\beta'}(\Phi(\beta')\gamma')(\beta, \gamma), \Phi(\beta)]. \end{aligned} \quad (15)$$

We also introduce the notation

$$R_n(\theta) \doteq R_n(\beta, \gamma) = \frac{1}{n} J(\theta)^T J(\theta), \quad (16)$$

which matrix plays an important role in the asymptotic theory of nonlinear least-squares problems presented later. Note that under Assumption A2, $\det(R_n(\theta)) \neq 0$ for every $n \geq d_\theta$.

D. Asymptotic Confidence Regions

Under standard regularity assumptions, the nonlinear least-squares estimator is *asymptotically normal* [14]. More specifically, assume that $\theta^* = [(\beta^*)^T, (\gamma^*)^T]^T$ lies in the interior of the compact set $\Theta \subset \mathbb{R}^{d_\theta}$, the regression function $f(\theta) \doteq \Phi(\beta)\gamma$ is twice continuously differentiable in a neighborhood of θ^* , and that the noise sequence $\{W_t\}$ is i.i.d. with $\mathbb{E}[W_t] = 0$ and $\mathbb{E}[W_t^2] = \sigma^2$. Assume further that the model is *identifiable*. Then, any sequence of estimators $\hat{\theta}_n$ globally minimizing the empirical least-squares objective over Θ is *consistent*, i.e., $\hat{\theta}_n \xrightarrow{p} \theta^*$ as $n \rightarrow \infty$. Finally, assuming that $R = \lim_{n \rightarrow \infty} R_n(\theta^*)$ is non-singular, the estimators satisfy

$$\sqrt{n}(\hat{\theta}_n - \theta^*) \xrightarrow{d} \mathcal{N}(0, \sigma^2 R^{-1}). \quad (17)$$

Consequently, as $n \rightarrow \infty$, we have

$$\frac{n}{\sigma^2} (\hat{\theta}_n - \theta^*)^T R (\hat{\theta}_n - \theta^*) \xrightarrow{d} \chi^2(d_\theta), \quad (18)$$

where $\chi^2(d_\theta)$ is the chi-squared distribution with d_θ degrees of freedom. For a prescribed confidence level p , let $\mu_p \doteq F_{\chi^2(d_\theta)}^{-1}(p)$ denote the p -quantile of the $\chi^2(d_\theta)$ distribution. Using the finite sample estimates of $R_n(\hat{\theta}_n) = \frac{1}{n} J(\hat{\theta}_n)^T J(\hat{\theta}_n)$ and $\hat{\sigma}_n^2 \doteq \frac{1}{n-d_\theta} \|y - \Phi(\hat{\beta}_n)\hat{\gamma}_n\|^2$, an asymptotic confidence ellipsoid can be constructed as

$$\mathcal{A}_n^{(p)} \doteq \left\{ \theta \in \mathbb{R}^{d_\theta} : (\theta - \hat{\theta}_n)^T R_n(\hat{\theta}_n) (\theta - \hat{\theta}_n) \leq \frac{\mu_p \hat{\sigma}_n^2}{n} \right\}. \quad (19)$$

This construction guarantees $\lim_{n \rightarrow \infty} \mathbb{P}(\theta^* \in \mathcal{A}_n^{(p)}) = p$.

IV. NON-ASYMPTOTIC CONFIDENCE REGIONS

In this section, we propose two non-asymptotic confidence region constructions for SNLLS problems. They are motivated by the *Sign-Perturbed Sums* (SPS) method [15] and its generalizations, the *Data Perturbation* (DP) methods [16]. DP is a family of hypothesis tests used to construct non-asymptotic confidence regions with *exact* coverage. Although DP methods are related to the *wild bootstrap*, in that they typically resample residuals, they do not necessarily require *refitting* the model, yet they still guarantee *finite-sample coverage* (see [15, Section V.D] for a comparison). Operating under very mild statistical assumptions, they require only a distributional invariance from the noise. The core idea is to generate alternative datasets via resampling, rank them using a performance measure, and reject the candidate model if the rank of the original dataset exceeds a given threshold.

Algorithm 1 Pseudocode: Initialization (p)

- 1: Given the (rational) confidence probability $p \in (0, 1)$, set integers $m > q > 0$ such that $p = 1 - q/m$.
 - 2: Generate $G_1, \dots, G_{m-1} \sim \text{Unif}(\mathcal{G})$ i.i.d. (uniform) random perturbations. See (3) and (4) for examples.
 - 3: Generate (uniformly) a random permutation π of the set $\{0, \dots, m-1\}$, where each of the $m!$ possible permutations has probability $1/(m!)$ to be selected.
-

Algorithm 2 Pseudocode: RPG-Indicator (θ)

- 1: Compute the residual vector for θ

$$r(\theta) \doteq y - \Phi(\beta)\gamma.$$
 - 2: Evaluate the reference and the perturbed functions:
$$\begin{aligned} S_0(\theta) &\doteq \frac{1}{n} J(\theta)^T r(\theta), \\ S_i(\theta) &\doteq \frac{1}{n} J(\theta)^T G_i(r(\theta)), \quad i \in [m-1]. \end{aligned}$$
 - 3: Compute the inverse $R_n^{-1}(\theta)$ and the normalized rank of $\|S_0(\theta)\|_{R_n^{-1}(\theta)}^2$ among $\{\|S_i(\theta)\|_{R_n^{-1}(\theta)}^2\}_{i=1}^{m-1}$:
$$\mathcal{R}(\theta) \doteq \frac{1}{m} \left[1 + \sum_{i=1}^{m-1} \mathbb{I} \left(\|S_0(\theta)\|_{R_n^{-1}(\theta)}^2 \succ_{\pi} \|S_i(\theta)\|_{R_n^{-1}(\theta)}^2 \right) \right],$$

where “ $\succ_{\pi}$ ” is “ $>$ ” with random tie-breaking, more precisely, $\|S_k(\theta)\|_{R_n^{-1}(\theta)}^2 \succ_{\pi} \|S_j(\theta)\|_{R_n^{-1}(\theta)}^2$ if and only if $(\|S_k(\theta)\|_{R_n^{-1}(\theta)}^2 > \|S_j(\theta)\|_{R_n^{-1}(\theta)}^2) \vee (\|S_k(\theta)\|_{R_n^{-1}(\theta)}^2 = \|S_j(\theta)\|_{R_n^{-1}(\theta)}^2 \wedge \pi(k) > \pi(j))$.
 - 4: Return 1 if $\mathcal{R}(\theta) \leq p$, otherwise return 0.
-

A. The Residual-Perturbed Gradient Algorithm

Our first proposed DP method, called *Residual-Perturbed Gradient (RPG)*, consists of two parts: initialization, presented in Algorithm 1, and the indicator function, given in Algorithm 2. The initialization part computes the main parameters and the i.i.d. random perturbations given the desired confidence probability p as input. The indicator function first computes the residual for the given candidate parameter vector, then forms a collection of perturbed residuals using the random perturbations. For each residual (original and perturbed) it computes a score vector using the Jacobian. Then, the original score vector is ranked among the randomized ones with performance measure $\|\cdot\|_{R_n^{-1}(\theta)}$, a scale-adjusted metric that defines the classical asymptotic confidence region, see (19). Using this indicator function, the p -level RPG confidence region can be defined as

$$\mathcal{B}_p \doteq \{ \theta \in \mathbb{R}^{d_\theta} : \text{RPG-Indicator}(\theta) = 1 \}. \quad (20)$$

Notice that, as a consequence of (11) and (15), the quantity $\|S_0(\theta)\|_{R_n^{-1}(\theta)}^2$ measures how strongly the first-order optimality conditions are violated at θ . In particular, at any *stationary point* of the least-squares objective (7), we have $S_0(\theta) = 0$, and therefore such points are always in $\mathcal{B}_p$.

Algorithm 3 Pseudocode: GFG-Indicator (θ)

1: Determine the residual vector for θ ,

$$r(\theta) \doteq y - \Phi(\beta)\gamma.$$

2: Evaluate alternative outputs, for $i \in [m-1]$,

$$\tilde{y}_i(\theta) \doteq \Phi(\beta)\gamma + G_i(r(\theta)),$$

and set $\tilde{y}_0(\theta) \doteq y$.

3: Calculate global-refit estimates, for $i \in \{0, \dots, m-1\}$,

$$(\hat{\beta}_i, \hat{\gamma}_i) \doteq \underset{\beta', \gamma'}{\operatorname{argmin}} \|\tilde{y}_i(\theta) - \Phi(\beta')\gamma'\|^2.$$

4: Determine global-fit-gap scores, for $i \in \{0, \dots, m-1\}$:

$$F_i(\theta) \doteq \|\tilde{y}_i(\theta) - \Phi(\beta)\gamma\|^2 - \|\tilde{y}_i(\theta) - \Phi(\hat{\beta}_i)\hat{\gamma}_i\|^2.$$

5: Compute the normalized rank of $F_0(\theta)$ in $\{F_i(\theta)\}_{i=1}^{m-1}$,

$$\mathcal{R}(\theta) \doteq \frac{1}{m} \left[1 + \sum_{i=1}^{m-1} \mathbb{I}(F_0(\theta) \succ_{\pi} F_i(\theta)) \right],$$

where “ $\succ_{\pi}$ ” is “ $>$ ” with random tie-breaking, as before.

6: Return 1 if $\mathcal{R}(\theta) \leq p$, otherwise return 0.

B. The Global-Fit-Gap Algorithm

While the RPG-Indicator is a natural extension of gradient-perturbation based DP methods to SNLLS, it may be undesirable in highly *nonconvex* problems where several local minima can occur. To address this issue, we introduce an alternative construction, the *Global-Fit-Gap* (GFG), presented in Algorithm 3, which evaluates candidate parameters based on *global optimality* rather than first-order conditions.

The GFG-Indicator generates alternative datasets by applying random perturbations to the residual, and for each dataset a nonlinear LS fit is performed. Then, it evaluates the input parameter by comparing its least-squares fit to that obtained after re-optimizing the model on the randomly perturbed residual data. Hence, the resulting statistic quantifies the *suboptimality* of the input parameter for the corresponding dataset. Using this algorithm, the GFG confidence region is

$$\mathcal{C}_p \doteq \{\theta \in \mathbb{R}^{d_{\theta}} : \text{GFG-Indicator}(\theta) = 1\}. \quad (21)$$

In the optimization step, any suitable nonlinear least-squares routine can be applied to obtain a global optimum. Besides the found solution, no other stationary points will be automatically included in the resulting confidence set. Compared to RPG, the GFG-Indicator requires repeated refitting of the model, and hence it is computationally more demanding, but it provides a stronger, more reliable acceptance criterion.

C. Illustration of RPG and GFG

We illustrate the difference between RPG and GFG using a scalar sinc function setup:

$$Y_t = \gamma^* \frac{\sin(\beta^* X_t)}{\beta^* X_t} + W_t, \quad (22)$$

where $\gamma^* = 1$, $\beta^* = 2.2$, X_t is deterministic and equidistant on $[-2, 2]$, $W_t \sim \mathcal{N}(0, 0.55^2)$, and $n = 72$. Since the

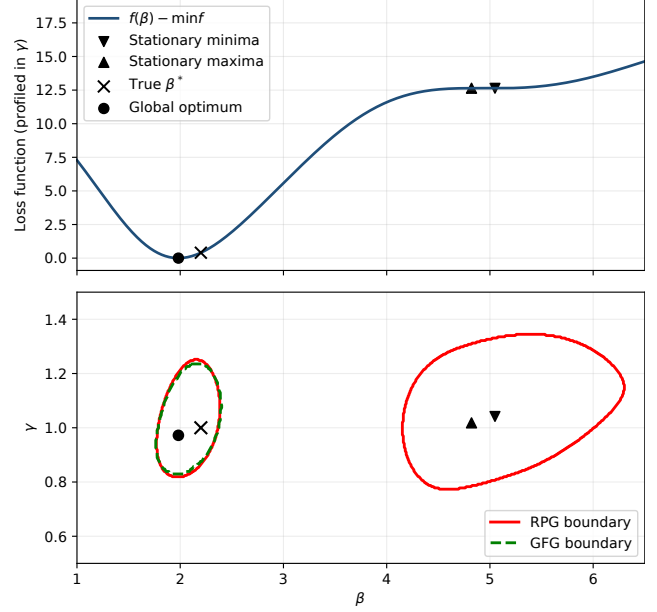


Fig. 1. Comparison of the 0.5-level ($m = 2$, $q = 1$) RPG and GFG indicators for $n = 72$, considering a $\text{sinc}(x)$ function with Gaussian noise.

noise is symmetric, we used sign-perturbations to construct the confidence regions. Fig. 1 shows the profiled loss in β , obtained after minimizing over γ , together with the 0.5-level ($m = 2$, $q = 1$) acceptance boundaries of RPG and GFG.

It can be observed that RPG and GFG constructed confidence regions that included the true parameter and the global optimum of the loss function, while RPG also constructed a separate region that included non-global stationary points. In contrast, GFG did not provide additional acceptance regions. This illustrates that GFG can produce sharper confidence regions at the price of higher computational cost.

D. Theoretical Guarantees

The constructed confidence regions $\mathcal{B}_p$ and $\mathcal{C}_p$ have exact coverage probability; therefore, the following theorem holds:

Theorem 1. Under Assumptions A1 and A2 the coverage probability of the confidence regions $\mathcal{B}_p$ and $\mathcal{C}_p$ is exactly p :

$$\mathbb{P}(\theta^* \in \mathcal{B}_p) = p, \quad \text{and} \quad \mathbb{P}(\theta^* \in \mathcal{C}_p) = p. \quad (23)$$

Proof. From (11)–(14), it follows that in the RPG and GFG indicator functions, for $\theta = \theta^*$:

$$S_i(\theta^*) = \frac{1}{n} J(\theta^*)^T G_i(w), \quad (24)$$

and

$$\tilde{y}_i(\theta^*) = \Phi(\beta^*)\gamma^* + G_i(w) \quad (25)$$

for all $i \in \{0, \dots, m-1\}$, where G_0 is the identity element of $\mathcal{G}$, i.e., the identity transformation. Hence, for all $i \in \{0, \dots, m-1\}$, the quantities $S_i(\theta^*)$ and $\tilde{y}_i(\theta^*)$ depend on the perturbed noise sequences through the same functions, respectively, which can be defined as $\psi_s(G_i(w)) \doteq S_i(\theta^*)$ and $\psi_y(G_i(w)) \doteq \tilde{y}_i(\theta^*)$. Then, the application of [16, Theorem 2.17] guarantees that $\{\|S_i(\theta^*)\|_{R_n^{-1}(\theta^*)}^2\}_{i=0}^{m-1}$ and

TABLE I

MONTE CARLO COVERAGE IN THE GFIR MODEL WITH 95% INTERVALS.

m	$p = 1 - 1/m$	RPG empirical coverage	GFG empirical coverage
2	0.500	0.500 [0.490, 0.510]	0.497 [0.487, 0.507]
5	0.800	0.799 [0.791, 0.807]	0.797 [0.789, 0.805]
10	0.900	0.899 [0.893, 0.905]	0.896 [0.890, 0.902]
20	0.950	0.946 [0.942, 0.951]	0.951 [0.946, 0.955]
50	0.980	0.979 [0.976, 0.982]	0.982 [0.979, 0.985]

$\{F_i(\theta^*)\}_{i=0}^{m-1}$ are *uniformly ordered*. In the theorem above, our two suggested methods have performance measures:

$$Z(\theta, x, y') = \|\frac{1}{n}J(\theta)^T(y' - \Phi(\beta, x)\gamma)\|_{R_n^{-1}(\theta)}^2 \quad (26)$$

in the RPG case, and

$$Z(\theta, x, y') = \|y' - \Phi(\beta, x)\gamma\|^2 - \|y' - \Phi(\hat{\beta}, x)\hat{\gamma}\|^2 \quad (27)$$

in the GFG case, where we have written $x \doteq [X_1, \dots, X_n]^T$ for the sake of completeness and $(\hat{\beta}, \hat{\gamma}) \doteq \operatorname{argmin}_{\beta', \gamma'} \|y' - \Phi(\beta', x)\gamma'\|^2$. Consequently, $\|S_0(\theta^*)\|_{R_n^{-1}(\theta^*)}^2$ and $F_0(\theta^*)$ take each position in their respective ordering with the same probability $1/m$; therefore, their normalized rank is at most $p = 1 - q/m$ with probability exactly p . $\square$

V. SIMULATION EXPERIMENTS

In this section, we illustrate the RPG and GFG algorithms through two numerical experiments. In the first one we consider the identification of a *Generalized Finite Impulse Response* (GFIR) system with Laguerre basis functions that can be used to model systems with *slowly decaying impulse responses*. In the second experiment we consider a curve fitting problem with *Gaussian basis functions*, a classical model in kernel methods and function approximation.

A. Generalized FIR Models with Adjustable Decay Rate

We consider a 2-dimensional GFIR model using *Laguerre basis functions* with *adjustable decay rate*:

$$Y_t = \gamma_1^* \varphi_1(\beta^*, U_t) + \gamma_2^* \varphi_2(\beta^*, U_t) + W_t, \quad (28)$$

where for $i \in \{1, 2\}$: $\varphi_i(\beta, U_t) = L_i(z, \beta)U_t$, z denotes the shift operator, and the filter is defined as

$$L_i(z, \beta) = \frac{1}{z - \beta} \left(\frac{1 - \beta z}{z - \beta} \right)^{i-1}. \quad (29)$$

In this experiment, the true parameters are $\beta^* = 0.55$, $\gamma^* = [1.2, -0.8]^T$, $\{W_t\}$ are i.i.d. heavy-tailed random variables sampled from a Pareto distribution with scale parameter 1 and shape parameter 1.5, the input $\{U_t\}$ is sampled from a uniform distribution with boundaries $[-2, 2]$, and the sample size is $n = 250$. Notice that the resulting parameter vector is $\theta^* = [\beta^*, (\gamma^*)^T]^T$, and the model is linear in γ^* and nonlinear in β^* through the Laguerre filters $L_i(z, \beta)$, hence, the identification problem can be formulated as an SNLLS.

For the construction of the confidence regions we used permutations, see (3). The parameter estimate $(\hat{\beta}, \hat{\gamma})$ was obtained by VP with Levenberg–Marquardt (LM) refinement.

To empirically check the finite-sample coverage claim and the dependence on m , we ran 10^4 independent Monte

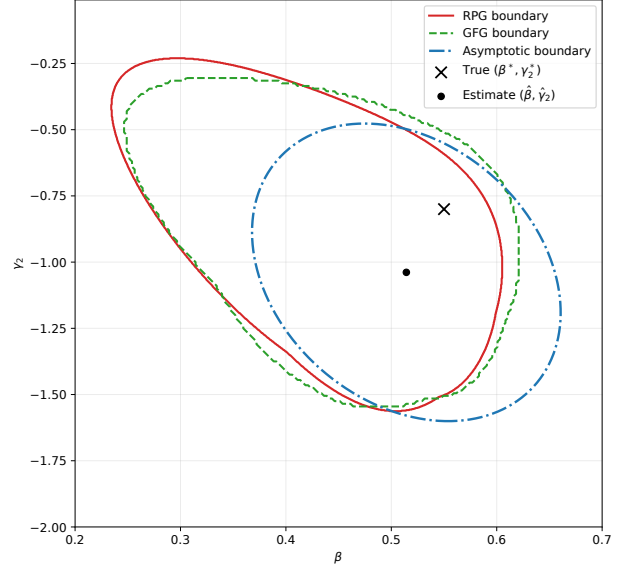


Fig. 2. Comparison of 0.9-level RPG, GFG, and asymptotic confidence regions considering a GFIR model with heavy-tailed Pareto noise and fixed $\gamma_1 = \hat{\gamma}_1$. $m = 10$, $q = 1$, $n = 250$.

Carlo trials, regenerating both the data and the randomized permutation transformations in each one. Table I reports the fraction of trials for which $\theta^* \in \mathcal{B}_p$ and $\theta^* \in \mathcal{C}_p$ with exact 95% Clopper–Pearson (CP) intervals. We used $q = 1$, hence $p = 1 - q/m$, so increasing m increases the nominal confidence level and gives more conservative rank regions. The empirical coverages closely follow the nominal levels for both methods. Moreover, in all cases, the nominal confidence level lies inside the corresponding exact 95% CP interval.

The 0.9-level RPG and GFG confidence regions with $m = 10$ and $q = 1$, together with the asymptotic region, see (19), are illustrated in Fig. 2. In this figure we consider the (β, γ_2) slice with $\gamma_1 = \hat{\gamma}_1$. The contours of the indicator regions were obtained by evaluating them on a grid. It can be seen that the RPG and GFG regions have similar shapes and sizes, and both form a single connected component including the true parameter and the estimate. Here, the RPG does not create additional disconnected regions in the displayed slice. The “asymptotic ellipsoid” of (19) is included only as a heuristic reference. It provides no finite-sample guarantees, and given that the noise terms have *infinite variance*, it lacks asymptotic guarantees as well, since the classical central limit theorem does not apply to justify asymptotic normality.

B. Gaussian Models with Adjustable Knots and Widths

In this experiment, we consider the following separable nonlinear *Gaussian* model with *adjustable knots and widths*:

$$Y_t = \sum_{k=1}^d \gamma_k^* \exp \left(-\frac{(X_t - \mu_k^*)^2}{2(\sigma_k^*)^2} \right) + W_t, \quad (30)$$

for $t \in [n]$, where $d = 3$, the nonlinear parameter vector is

$$\beta^* = [\mu_1^* \quad \mu_2^* \quad \mu_3^* \quad \sigma_1^* \quad \sigma_2^* \quad \sigma_3^*]^T$$

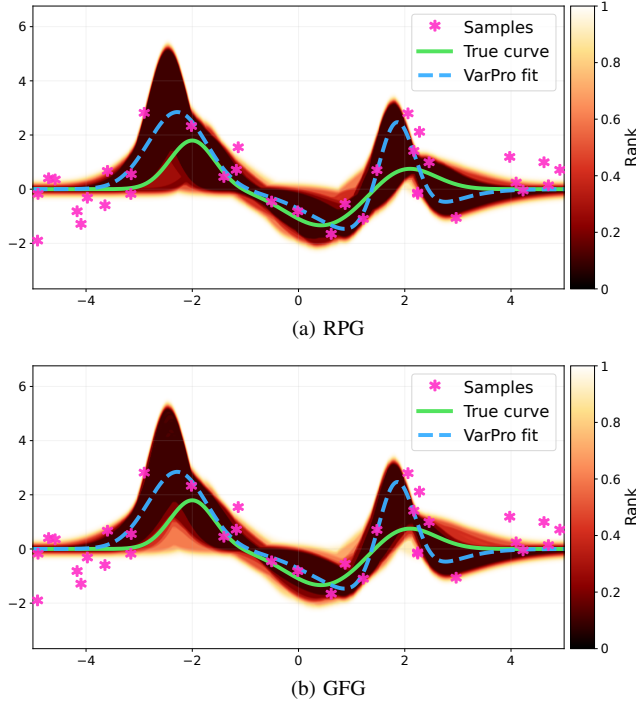


Fig. 3. Comparison of RPG and GFG rank maps considering a Gaussian model with adjustable knots, widths, and Laplace noise. $n = 35$, $m = 10$. For visual clarity, the displayed rank maps use the same post-processing.

$$= [-2.0 \quad 0.45 \quad 2.0 \quad 0.42 \quad 0.70 \quad 0.55]^T, \quad (31)$$

and true linear parameters are $\gamma^* = [1.8, -1.35, 0.85]^T$. The input points $\{X_t\}$ are generated by uniform random sampling on $[-5, 5]$, the noises $\{W_t\}$ are i.i.d. Laplace distributed with zero mean and variance 1.5, and $n = 35$.

As the noises are symmetric about zero, sign-perturbations are used, i.e., $G_i = \text{diag}(\alpha_{i,1}, \dots, \alpha_{i,n})$, where for $i \in [m-1]$ and $t \in [n]$, $\mathbb{P}(\alpha_{i,t} = 1) = \mathbb{P}(\alpha_{i,t} = -1) = 1/2$. Throughout this experiment we use $m = 10$ and $q = 1$.

Fig. 3 shows the noisy samples together with the true and estimated curve, obtained by VP, as well as the corresponding rank maps of RPG and GFG. To compute these rank maps, we evaluated both methods on a structured set of candidate nonlinear parameter vectors constructed around the true and VP-estimated parameters by interpolating between them. For each candidate, the linear coefficients are recomputed by LS and the resulting curve is mapped into the input-output plane.

It can be observed that the GFG rank map is locally more concentrated around the fitted curve. This can be seen near $x \approx -2.2$ and on the right shoulder around $x \approx 3$, where RPG spreads confidence over a broader family of curves. Hence, GFG yields a sharper local characterization of uncertainty, at the cost of repeated refits.

VI. CONCLUSIONS

SNLLS problems are widely used in control theory and machine learning, with the VP method as their archetypal solution. Despite their popularity, there are very few results on uncertainty quantification (UQ) for SNLLS. In this paper, we introduced two new methods, RPG and GFG, for constructing distribution-free confidence regions in SNLLS

problems. By leveraging the group invariance of the noise, such as symmetry or exchangeability, we have shown that it is possible to move beyond traditional asymptotic ellipsoids, which often provide unreliable guarantees in finite-sample regimes. We proved that both methods guarantee non-asymptotically exact coverage, requiring only mild assumptions on the noise distribution. Numerical experiments further highlighted the practical trade-offs between the two approaches: while RPG is more computationally efficient, it may capture local extrema in highly non-convex landscapes. GFG provides more selective confidence regions by focusing on global optimality, albeit at a higher computational cost.

ACKNOWLEDGMENTS

The authors would like to thank Sándor Kolumbán for the valuable discussions on data perturbation methods.

REFERENCES

- [1] G.-Y. Chen, M. Gan, S. Wang, and C. L. P. Chen, "Insights into algorithms for separable nonlinear least squares problems," *IEEE Transactions on Image Processing*, vol. 30, pp. 1207–1218, 2021.
- [2] G. H. Golub and V. Pereyra, "The differentiation of pseudo-inverses and nonlinear least squares problems whose variables separate," *SIAM Journal on Numerical Analysis*, vol. 10, no. 2, pp. 413–432, 1973.
- [3] G. Golub and V. Pereyra, "Separable nonlinear least squares: the variable projection method and its applications," *Inverse problems*, vol. 19, no. 2, p. R1, 2003.
- [4] M. Gan, H.-X. Li, and H. Peng, "A variable projection approach for efficient estimation of RBF-ARX model," *IEEE Transactions on Cybernetics*, vol. 45, no. 3, pp. 462–471, 2015.
- [5] T. Dózsa, M. Szabari, A. Soumelidis, and P. Kovács, "Pole identification using discrete laguerre expansion and variable projection," in *22nd IFAC World Congress*, vol. 56, no. 2, 2023, pp. 3994–3999.
- [6] G.-Y. Chen, X.-X. Su, M. Gan, W. Guo, and C. L. P. Chen, "Robust variable projection algorithm for the identification of separable nonlinear models," *IEEE Transactions on Automatic Control*, vol. 69, no. 9, pp. 6293–6300, 2024.
- [7] T. van Leeuwen and A. Y. Aravkin, "Variable projection for nonsmooth problems," *SIAM Journal on Scientific Computing*, vol. 43, no. 5, pp. S249–S268, 2021.
- [8] G. Chen, P. Xue, M. Gan, J. Chen, W. Guo, and C. P. Chen, "Variable projection algorithms: Theoretical insights and a novel approach for problems with large residual," *Automatica*, vol. 177, p. 112300, 2025.
- [9] M. Gan, Y. Guan, G.-Y. Chen, and C. L. P. Chen, "Recursive variable projection algorithm for a class of separable nonlinear models," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 11, pp. 4971–4982, 2021.
- [10] E. Newman, L. Ruthotto, J. Hart, and B. van Bloemen Waanders, "Train like a (Var)Pro: Efficient training of neural networks with variable projection," *SIAM Journal on Mathematics of Data Science*, vol. 3, no. 4, pp. 1041–1066, 2021.
- [11] T. Dózsa, C. Böck, J. Meier, and P. Kovács, "Weighted Hermite variable projection networks for classifying visually evoked potentials," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 7, pp. 12 415–12 428, 2025.
- [12] D. Kundu and A. Mitra, "On asymptotic behavior of least squares estimators and the confidence intervals of the superimposed exponential signals," *Signal Processing*, vol. 72, no. 2, pp. 129–139, 1999.
- [13] N. Bertozzi and C. Bonanno, "Asymptotic behaviour of the confidence region in orbit determination for hyperbolic maps with a parameter," *Physica D: Nonlinear Phenomena*, vol. 470, p. 134403, 2024.
- [14] A. W. van der Vaart, *Asymptotic Statistics*. Cambridge University Press, 1998.
- [15] 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 Transactions on Signal Processing*, vol. 63, no. 1, pp. 169–181, 2015.
- [16] S. Kolumbán, "System identification in highly non-informative environment," Ph.D. dissertation, Budapest University of Technology and Economics, Hungary, and Vrije Universiteit Brussels, Belgium, 2015.