MTA SZTAKI
Hungarian Academy of Sciences
Institute for Computer Science and Control

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

Algo Carè [1,] Balázs Csáji [2] Marco Campi [3] Erik Weyer [4]

[1] Centrum Wiskunde & Informatica (CWI), Amsterdam, Netherlands

[2] Institute for Computer Science and Control (SZTAKI), Hungarian Academy of Sciences (MTA), Hungary

[3] Department of Information Engineering (DII), University of Brescia, Italy

[4] Department of Electrical and Electronic Engineering, University of Melbourne, Australia

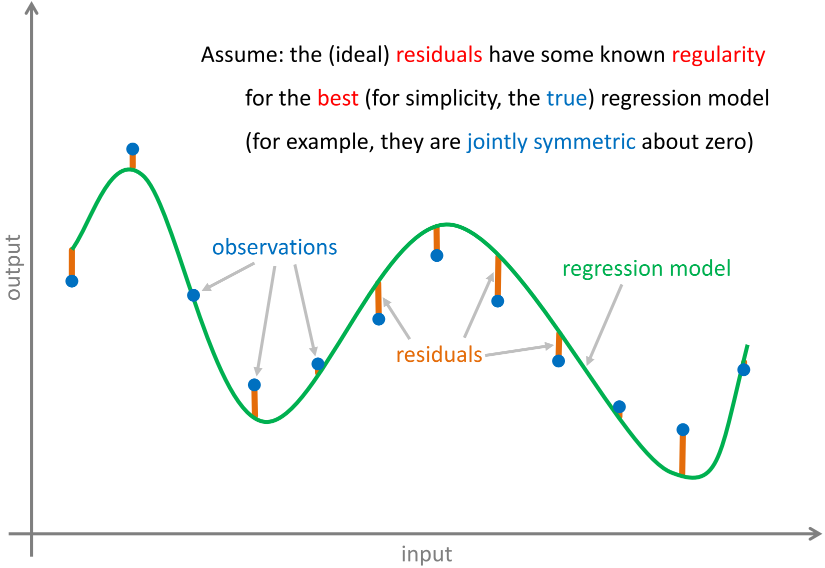
56th IEEE CDC, Melbourne, Australia, December 12-15, 2017

# Regularity Assumption

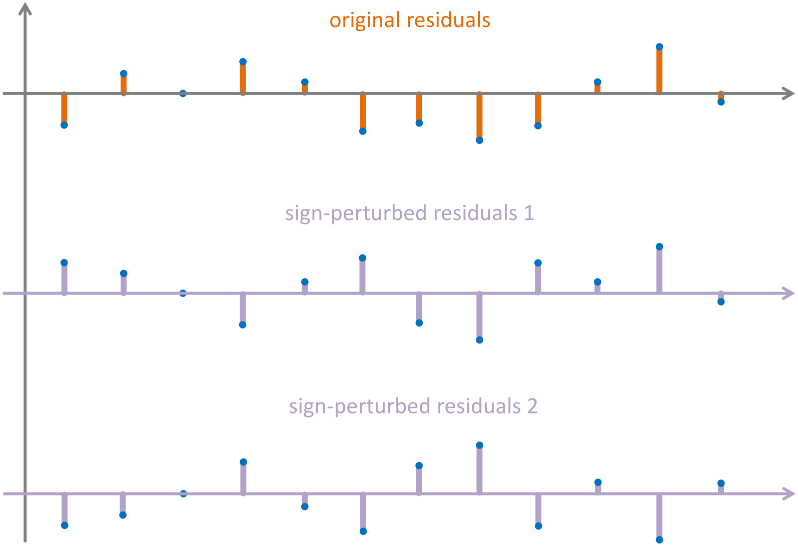
Assume: the (ideal) residuals have some known regularity

for the best (for simplicity, the true) regression model

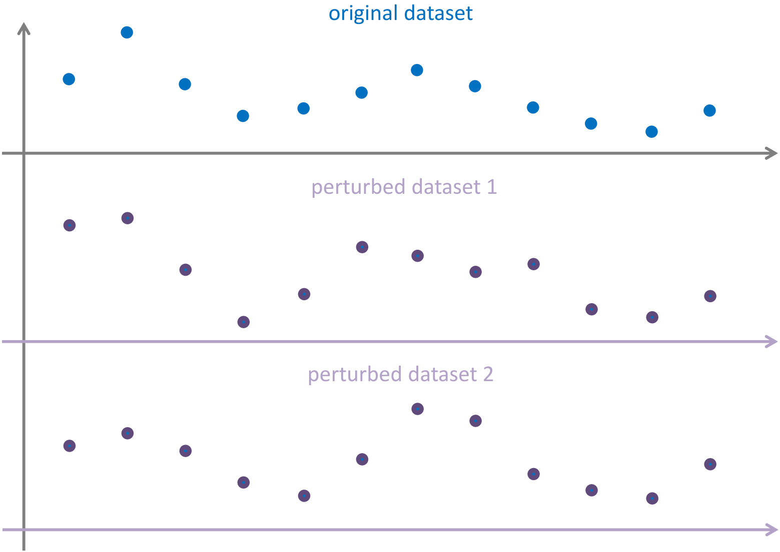
(for example, they are jointly symmetric about zero)



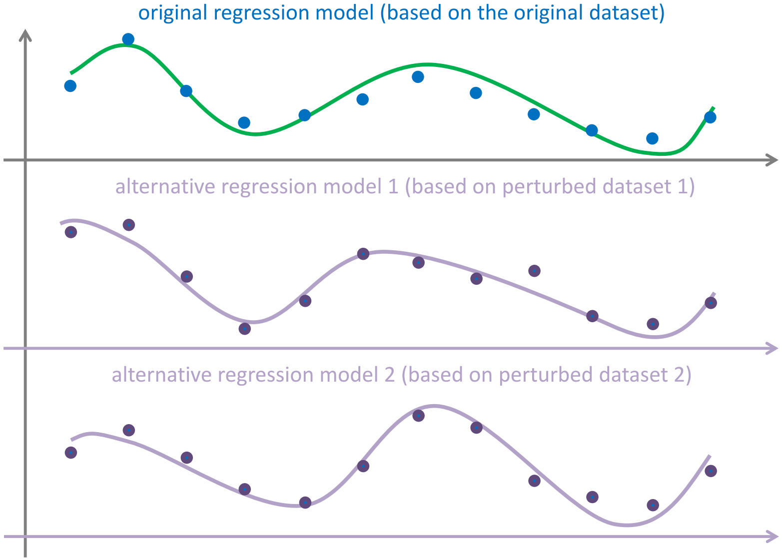
# Perturbed Residuals



# Perturbed Datasets



# Alternative Regression Models



# Data Generation

Let us consider the following data generating system

**System Structure**

$$\mathbf{Y}_n \triangleq \mathbb{F}(\mathbf{U}_n, \mathbf{W}_n, \mathcal{I})$$

where

$\mathcal{I}$ — initial conditions

$\mathbf{U}_n \triangleq (U_1, \dots, U_n)^{\mathrm{T}}$ — inputs

$\mathbf{W}_n \triangleq (W_1, \dots, W_n)^{\mathrm{T}}$ — noises

$\mathbf{Y}_n \triangleq (Y_1, \dots, Y_n)^{\mathrm{T}}$ — outputs

$\mathbb{F}$ — true data generating function

# Point Estimation

Consider the parametric estimation problem of the system

$$\mathbf{Y}_n \triangleq \mathbb{F}_{\theta^*}(\mathbf{U}_n, \mathbf{W}_n, \mathcal{I})$$

parametrized with $\theta^* \in \Theta \subseteq \mathbb{R}^d$ (true parameter)

Given: finite sample of data, $\mathcal{Z} \triangleq (\mathbf{U}_n, \mathbf{Y}_n, \mathcal{I})$
We typically search for a model that best fit the data, that is

## Point Estimate (Parametric)

$$\widehat{\theta}_{\mathcal{Z}} \triangleq \operatorname*{arg\,min}_{\theta \in \Theta} \mathcal{V}(\theta \mid \mathcal{Z})$$

where $\mathcal{V}$ is a criterion function

# Confidence Regions

In practice often some quality tag is needed to judge the estimate.

Safety, stability, or quality requirements? $\Rightarrow$ confidence regions

### Confidence Region (Level $\mu$)

$$\mathbb{P}\big(\theta^* \in \widehat{\Theta}_{\mathcal{Z},\mu}\big) \geq \mu$$

for some $\mu \in (0,1)$, where $\theta^*$ is the "true" parameter, $\widehat{\Theta}_{\mathcal{Z},\mu} \subseteq \Theta$.

Typically the level sets of the (scaled) limiting distribution is used.

Issues: only approximately correct for finite samples,
requires the existence of a (known) limiting distribution.

# Main Assumptions

## Assumption 1

For any value of $\theta^* \in \Theta$, the relation $\mathbf{Y}_n \triangleq \mathbb{F}_{\theta^*}(\mathbf{U}_n, \mathbf{W}_n, \mathcal{I},)$ is noise invertible in the sense that, given the values of $\mathbf{Y}_n$, $\mathbf{U}_n$, $\mathcal{I}$, we can recover the noise $\mathbf{W}_n$.

## Assumption 2

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

# Residuals and Sign-Perturbations

Given a $\theta \in \Theta$ and dataset $\mathcal{Z}$, the estimated noise is $\widehat{\mathbf{W}}_n(\theta)$.

Note that we have $\widehat{\mathbf{W}}_n(\theta^*) = \mathbf{W}_n$ (Assumption 1).

Given vector $\mathbf{v}_n = (v_1, \dots, v_n)$ and signs $\mathbf{s}_n = (\sigma_1, \dots, \sigma_n)$ $\in \{+1, -1\}^n$, we denote the sign-perturbed vector by

$$\mathbf{s}_n[\mathbf{v}_n] \triangleq (\sigma_1 v_1, \dots, \sigma_n v_n).$$

Note that $\mathbf{W}_n \stackrel{d}{=} \mathbf{s}_n[\mathbf{W}_n]$, for all $\mathbf{s}_n \in \{+1, -1\}^n$ (Assumption 2)

where "$\stackrel{d}{=}$" denotes equal in distribution.

# Evaluation Functions

A core concept is the evaluation function (test statistic),

$$Z : \mathbb{R}^n \times \mathbb{R}^n \times \Theta \to \mathbb{R},$$

to evaluate the parameter based on ideas discussed before.
(Note that $Z$ can also depend on the initial conditions.)

Using $Z$ we define a reference and $m-1$ sign-perturbed functions,

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

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

for $i = 1, \dots, m-1$, where $\mathbf{s}_n^{(1)}, \dots, \mathbf{s}_n^{(m-1)}$ are $m-1$ user-generated vectors containing i.i.d. symmetric random signs.

# Evaluating Parameters

It can be shown that $Z_0(\theta^*), \ldots, Z_{m-1}(\theta^*)$ are conditionally i.i.d.

Consider the ordering $Z_{(0)}(\theta^*) < \cdots < Z_{(m-1)}(\theta^*)$,
where we apply random tie-breaking, if needed.

Then $\boxed{\text{All orderings are equally probable!}}$

We want to design $Z$ to such that as $\theta$ gets "far away" from $\theta^*$,

$$Z_0(\theta) < Z_i(\theta)$$

with "high probability" for all $i = 1, \ldots, m-1$; or

$$Z_i(\theta) < Z_0(\theta)$$

with "high probability" for all $i = 1, \ldots, m-1$.

# Non-Asymptotic Confidence Regions

The rank of $Z_0(\theta)$ in the ascending ordering of $\{Z_i(\theta)\}_{i=0}^{m-1}$ is

$$\mathcal{R}(\theta) = 1 + \sum\nolimits_{i=1}^{m-1} \mathbb{I}(Z_i(\theta) < Z_0(\theta)),$$

where $\mathbb{I}(\cdot)$ is an indicator function.

## Exact Confidence

The confidence region defined as

$$\widehat{\Theta}_n \triangleq \left\{ \theta \in \mathbb{R}^d \,:\, h \leq \mathcal{R}(\,\theta\,) \leq k \right\}$$

is such that $\mathbb{P}\{\theta^* \in \widehat{\Theta}_n\} = (k - h + 1)/m$, where $h, k$ and $m$ are user-chosen integers (design parameters).

# Construction Ideas

Typical construcions of the evaluation function $Z$ are based on

- Correlations: we use the fact that, for the true parameter, the residuals (noises) are uncorrelated, also with the inputs
  E.g.: LSCR (Leave-out Sign-dominant Correlation Regions)
- Gradients: based on the gradient (w.r.t. the parameter) of the criterion function of a given point estimate; we perturb the residuals in the gradient and scalarize it with a norm
  E.g.: SPS (Sign-Perturbed Sums)
- Models: new models are estimated based on the alternative (perturbed) datasets and then they are compared to the original (unperturbed) estimate (bootstrap style approach)
  E.g.: DP (Data Perturbation)

# A New Correlation Approach: Combining LSCR and SPS

What are the advantages and disadvantages of LSCR and SPS?

LSCR uses correlations (and subsampling).
It is a flexible and easy to implement algorithm.
It is computationally light, does not require perturbed datasets.
However, it is conservative for high dimensinal parameters.

SPS uses gradients (and sign-perturbations).
It evaluates the errors in all parameters simultaneously (norm).
It always constructs confidence regions having exact confidence.
However, it needs perturbed datasets, it is computationally heavy.

Let us try to combine the advantages of these two approaches!

# A New Correlation Approach: SPCR

New method: SPCR (Sign-Perturbed Correlation Regions).

For concreteness, let us consider an ARX($n_a$, $n_b$) model

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

### Stacked Correlations

For a generic $\mathbf{U}'_n$ and $\mathbf{W}'_n$, we introduce the correlation vectors

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

for $t = 1, \ldots, n$, where $k$ and $l$ are user-chosen parameters.

(Typically $k + l \geq n_a + n_b$, and we may need terms from $\mathcal{I}$.)

# A New Correlation Approach: SPCR

## Evaluation Function for SPCR

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

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

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

which is assumed to be invertible, for convenience.

# A New Correlation Approach: SPCR

## Confidence Regions for SPCR

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

And we have exact confidence for parameter vectors, as well

$$\mathbb{P}\{\,\theta^* \in \widehat{\Theta}_n\,\} = (k+1)/m.$$

Note that SPCR is a class of methods where different constructions correspond to different choices of $(k, l)$.

# Simulation Example for SPCR

Consider a bilinear system generated by

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

for $t = 1, \dots, n$, with $a^* = 0.7$, $b^* = 1$, with zero initial conditions.

The input sequence $\{U_t\}$ is generated by $U_t \triangleq 0.5\, U_{t-1} + V_t$, with zero initial conditions, where $\{V_t\}$ is i.i.d. standard normal.

The noise sequence $\{N_t\}$ is i.i.d. Laplacian with zero mean and unit variance, independent of $\{U_t\}$.

Our model class is ARX(1, 1), that is

$$\widehat{Y}_t(\theta) \triangleq a\, Y_{t-1} + b\, U_t.$$

# Simulation Example for SPCR

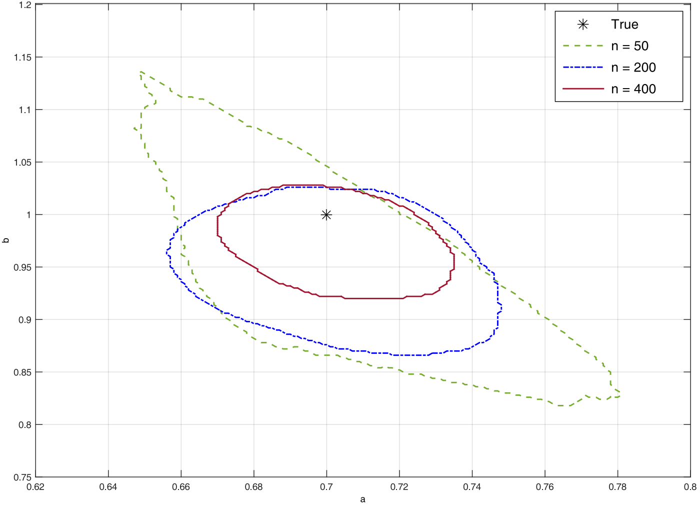



Figure: 95% confidence regions built by SPCR with $k = 2$ and $l = 2$.

# Desirable Properties of Finite Sample Sys.Id. Methods

- Inclusion of a point estimate: the confidence region should be centered around a given point estimate (e.g., PEM, QML).
- Consistency: for any false parameter, $\theta' \neq \theta^*$, the probability of $\theta' \in \widehat{\Theta}_n$ should decrease as the sample size, $n$, increases.
- Favorable topology: region $\widehat{\Theta}_n$ should have good topological properties, e.g., it should be bounded, connected, star convex.
- Weak computability: deciding whether a candidate parameter value $\theta$ belongs to $\widehat{\Theta}_n$ should be computationally easy.
- Strong computability: calculating a representation of $\widehat{\Theta}_n$ or an approximation of it should be computationally feasible.

# Conclusions

- A general, unifying overview on finite-sample system identification (FSID) methods was provided.
- The core ideas behind bulding exact, non-asymptotic, quasi distribution-free confidence regions were highlighted.
- A new method, SPCR (Sign-Perturbed Correlation Regions) was suggested as the combination of LSCR and SPS.
- SPCR combines the computational advantages of LSCR with the exactness of SPS by using stacked correlation vectors.
- A numerical experiment on a bilinear system was presented.
- Finally, desirable properties of FSID methods were highlighted and discussed based on the LSCR, SPS and SPCR methods.

# Thank you for your attention!

✉ balazs.csaji@sztaki.mta.hu