COMPARISON OF IDENTIFICATION METHODS ON DATA SETS FROM DAISY

Ivan Markovsky, Jan C. Willems, and Bart De Moor
ESAT/SCD, Katholieke Universiteit Leuven, Belgium

Data sets from DAISY:

#	Data set name	T	m	p	l
1	Lake Erie	57	5	2	1
2	Distillation column	90	5	3	1
3	Heating system	801	1	1	2
4	Industrial dryer	867	3	3	1
5	Hair dryer	1000	1	1	5
6	Ball-and-beam setup in SISTA	1000	1	1	2
7	Wing flutter data	1024	1	1	5
8	Flexible robot arm	1024	1	1	4
9	Glass furnace (Philips)	1247	3	6	1
10	Heat flow density	1680	2	1	2
11	pH neutralization process	2001	2	1	6
12	CD-player arm	2048	2	2	1
13	Industrial winding process	2500	5	2	2
14	Heat exchanger	4000	1	1	2
15	Industrial evaporator	6305	3	3	1
16	Tank reactor	7500	1	2	1
17	Steam generator	9600	4	4	1

m — number of inputs T — number of data points
p — number of outputs l — lag of the identified model

In all examples, the data $w = (u,y)$ is split into

w_{idt} — identification part, and
w_{val} — validation part.

The model class is LTI systems with a bound $n = lp$ on the order.

Compared methods:

Name	Description
`subid`	robust combined subspace algorithm
`uy2ssbal`	balanced subspace identification
`w2x2ss`	deterministic subspace algorithm
`cva`	`n4sid` with `N4Weight` = `CVA`
`moesp`	`n4sid` with `N4Weight` = `MOESP`
`pem`	OE identification in the PEM setting
`gtls`	OE identification using STLS

The validation criterion corresponds to the "simulation fit" computed by the function `compare` of the System Identification Toolbox.

Given $w = (u,y)$ and $\mathscr{B}$, define the approximation $\hat{y}$ of y in $\mathscr{B}$

$$\hat{y}\big((u,y),\mathscr{B}\big) := \min_{\hat{y}} \|y-\hat{y}\| \quad \text{subject to} \quad \operatorname{col}(u,\hat{y}) \in \mathscr{B}.$$

Let $\bar{y} := \sum_{t=1}^{T} y(t)/T$. The fit of w by $\mathscr{B}$ is defined as

$$F(w,\mathscr{B}) := 100 \max\big(0, 1 - \|y-\hat{y}(w,\mathscr{B})\|/\|y-\bar{y}\|\big).$$

`pem` is called with options:

- `'dist','none'`, which chooses output error model structure,
- `'nk',0`, which requires a feedthrough term to be estimated, and
- `'LimitError',0` which disables the default robustification of the cost function.

We list $F(w_{\text{val}}, \hat{\mathscr{B}})$ for the models produced by the compared identification methods.

Average fit in % on all datasets:

Experiment		`subid`	`uy2ssbal`	`w2x2ss`	`moesp`	`cva`	`pem`	`gtls`
70i/30v	idt	51.18	49.27	46.39	**55.52**	49.79	57.43	**68.46**
	val	32.14	31.57	32.34	**38.97**	33.38	37.77	**48.40**
30v/70i	idt	46.34	47.46	48.83	**53.86**	50.78	59.13	**68.87**
	val	36.96	37.69	38.15	**40.43**	37.10	45.17	**53.72**
80i/20v	idt	49.14	46.82	45.56	**55.13**	50.88	56.84	**68.36**
	val	30.01	28.20	29.75	**33.01**	31.75	36.17	**44.14**
20v/80i	idt	49.47	48.20	48.07	**54.48**	51.90	58.93	**68.48**
	val	**46.09**	37.30	40.81	39.79	39.81	45.28	**56.88**
90i/10v	idt	50.92	47.61	48.59	**54.79**	51.25	58.39	**68.95**
	val	**40.47**	32.89	31.46	37.06	35.07	39.48	**48.55**
10v/90i	idt	48.16	48.46	47.34	**53.93**	50.71	58.78	**69.06**
	val	**45.58**	43.71	45.13	44.12	39.71	43.62	**56.28**
Execution time		0.11	0.95	**0.05**	4.45	5.03	**14.79**	25.14

"70i/30v" is a short notation for "first 70% of the data is used for identification and the remaining 30% for validation"

The best fits and smallest execution times obtained by subspace and optimization methods are marked with **bold face**.

Discussion points:

- Data preprocessing
  - detrending
  - scaling
  - ??
- Imposing stability
- Fitting criteria
  - Determinant vs. trace
  - output error
  - errors-in-variables
  - ??