Tutorial on the behavioral approach to systems theory

Ivan Markovsky

In the classical approach, a system is an input-output map

$$\text{input} \longrightarrow \boxed{\text{system}} \longrightarrow \text{output}$$

the input *causes* the output

the system is a *signal processor*

the system is defined by *equations*

Premise: familiarity with classical approach

why is a different approach needed?

how is the behavioral approach different?

what new does it bring?

Thesis: behavioral approach has added value

Outline

A 10 minutes introduction to the behavioral approach

Data-driven representation of LTI systems

Showcase: Nonparametric frequency response estimation

Why is a different approach needed?

input/output maps assume zero *initial conditions*

- without input, what is a signal processor processing?
- initial conditions can be added as an afterthought

modeling from first principles leads to *relations*

- *e.g.*, ideal gas law: $PV = cMT$

  (P — pressure, V — volume, M — mass, T — temperature, c — constant)

interconnection of systems is *variables sharing*

- mechanical systems: position and velocity
- electrical systems: potential and current
- hydraulic systems: pressure and flow

The behavioral approach was put forward by Jan C. Willems in the 1980's

3-part, 70-page, Automatica paper:

Part I. Finite dimensional linear time invariant systems
Part II. Exact modelling
Part III. Approximate modelling

From Time Series to Linear System—
Part I. Finite Dimensional Linear Time Invariant Systems*

JAN C. WILLEMS†

Dynamical systems are defined in terms of their behaviour, and input/output systems appear as particular representations. Finite dimensional linear time invariant systems are characterized by the fact that their behaviour is a linear shift invariant complete (equivalently closed) subspace of $(\mathbb{R}^q)^{\mathbb{Z}}$ *or* $(\mathbb{R}^q)^{\mathbb{Z}_+}$.

"Good definition should formalize sensible intuition" Jan C. Willems

"I was not going to use the classical format where a definition is given first, followed by illustrative examples. I wanted this to go the other way around: show how examples lead to definitions."

some of the examples he used:

- Newton's second law
- Maxwell's equations
- the first and second laws of thermodynamics

How is the behavioral approach *different* from the classical one?

dynamical system $\mathscr{B}$ is a set of signals w

$$
\begin{aligned}
w \in \mathscr{B} \quad &\leftrightarrow \quad w \text{ is trajectory of } \mathscr{B} \\
&\leftrightarrow \quad \mathscr{B} \text{ is exact model for } w
\end{aligned}
$$

no inputs and outputs, no causality, no equations

the system is detached from its *representations*

properties and problems are separated from methods

How is the behavioral approach *similar* to the classical one?

input/output partitioning $w = \Pi \begin{bmatrix} u \\ y \end{bmatrix}$ and representations can be derived from $\mathscr{B}$, *e.g.*,

$$\mathscr{B} = \left\{ w = \Pi \begin{bmatrix} u \\ y \end{bmatrix} \in (\mathbb{R}^q)^{\mathbb{N}} \mid \exists\, x \in (\mathbb{R}^n)^{\mathbb{N}},\ \begin{bmatrix} \sigma x \\ y \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} x \\ u \end{bmatrix} \right\}$$

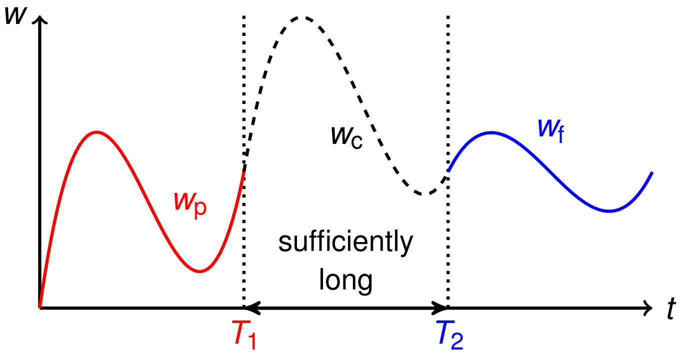
however

- given $\mathscr{B}$, an input/output partitioning is typically not unique
- also, properties and problems are defined in terms of $\mathscr{B}$
- equivalent representations define the same system

Example: what means that $\mathscr{B}$ is controllable?

controllability is the property of patching
any past trajectory with any future trajectory

$$w_{\mathrm{p}} \wedge w_{\mathrm{c}} \wedge w_{\mathrm{f}} \in \mathscr{B}$$



Compare with the classical definition: transfer from any initial to any terminal state

property of a state-space representation of $\mathscr{B}$

- is lack of controllability due to a bad choice of the state or due to an intrinsic issue with the system?
- in the LTI case, does it make sense to talk about controllability of a transfer function representation?
- how to quantify the distance to uncontrollability?

does not apply to infinite dimensional system

Separating problems from solution methods

different representations $\rightsquigarrow$ different methods

- with different properties (efficiency, robustness, ...)
- their common feature is that they solve the same problem

clarifies links among methods

leads to new methods

Back to the controllability example: how to check controllability of LTI system?

using state-space representation:

1. ensure minimality in the behavioral sense
2. perform rank test for the controllability matrix

using matrix fraction representation:

$$\mathscr{B} = \left\{ w = \Pi \left[\begin{smallmatrix} u \\ y \end{smallmatrix}\right] \in (\mathbb{R}^q)^{\mathbb{N}} \mid N(\sigma)u = D(\sigma)y \right\}$$

- facts: $\mathscr{B}$ is controllable $\iff$ N and D are co-prime
- $\rightsquigarrow$ rank test for the (generalized) Sylvester matrix

The behavioral approach is naturally suited for the data-driven paradigm

1940–1960	classical	SISO transfer function
1960–1980	modern	MIMO state-space
1980–2000	behavioral	the system as a set
2000–now	data-driven	using directly the data

A linear time-invariant system is a shift-invariant subspace

$\mathscr{B}$ is linear system $\quad :\Longleftrightarrow \quad$ $\mathscr{B}$ is subspace

$\mathscr{B}$ is time-invariant $\quad :\Longleftrightarrow \quad \sigma\mathscr{B} = \mathscr{B}$

$$(\sigma w)(t) := w(t+1) \qquad \sigma\mathscr{B} := \{\sigma w \mid w \in \mathscr{B}\}$$

restriction of w and $\mathscr{B}$ to finite interval $[1, L]$

$$w|_L := \big(w(1), \dots, w(L)\big) \qquad \mathscr{B}|_L := \{\, w|_L \mid w \in \mathscr{B}\,\}$$

The set of linear time-invariant systems $\mathscr{L}$ has structure characterized by set of integers

the dimension of $\mathscr{B} \in \mathscr{L}$ is determined by

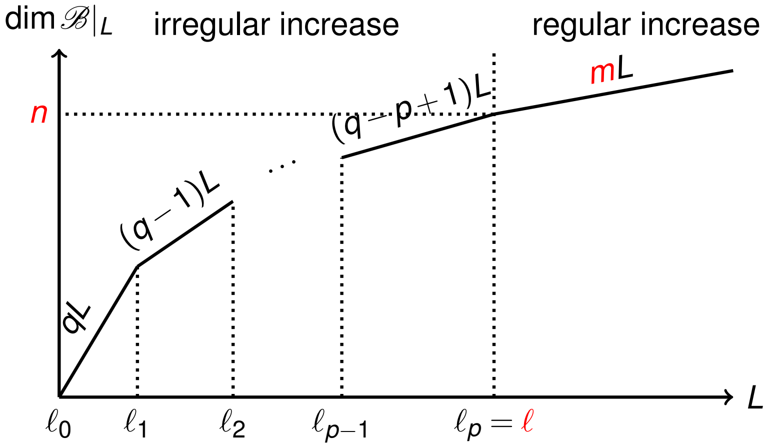
m — # of inputs ($p := q - m$ # of outputs)

n — order ($=$ minimal state dimension)

ℓ — lag ($=$ observability index)

J.C. Willems, From time series to linear systems.
Part I, Finite dimensional linear time invariant systems, Automatica, 22(561–580), 1986

$\dim \mathscr{B}|_L$ is a piecewise affine function of L



in particular, $\dim \mathscr{B}|_L = mL + n$, for all $L \geq \ell$

Identifiability: $w_{\mathrm{d}} \in \mathscr{B}$ specifies $\mathscr{B} \in \mathscr{L}$ (infinite data length case)

define $\quad \widehat{\mathscr{B}} := \operatorname{span}\{\, w_{\mathrm{d}}, \sigma w_{\mathrm{d}}, \sigma^2 w_{\mathrm{d}}, \dots \}$

fact: $\quad \widehat{\mathscr{B}} \in \mathscr{L} \quad$ and $\quad \widehat{\mathscr{B}} \subseteq \mathscr{B}$

identifiability condition: $\quad \widehat{\mathscr{B}} = \mathscr{B}$

J.C. Willems, From time series to linear systems. Part II, Exact modelling, Automatica, 22(675–694), 1986

In the finite data length case, shifting and cutting w_d leads to the Hankel matrix

for $w_\mathrm{d} = \big(w_\mathrm{d}(1), \dots, w_\mathrm{d}(T)\big)$ and $1 \le L \le T$

$$\mathscr{H}_L(w_\mathrm{d}) := \Big[(\sigma^0 w_\mathrm{d})|_L \quad (\sigma^1 w_\mathrm{d})|_L \quad \cdots \quad (\sigma^{T-L} w_\mathrm{d})|_L \Big]$$

define $\quad \widehat{\mathscr{B}}_L := \operatorname{image} \mathscr{H}_L(w_\mathrm{d})$

fact: $\quad \widehat{\mathscr{B}}_L \subseteq \mathscr{B}|_L$

Identifiability condition verifiable from $w_{\mathrm{d}} \in \mathscr{B}|_T$ and (m, ℓ, n)

fact: $\mathscr{B} = \mathscr{B}' \iff \mathscr{B}|_{\ell+1} = \mathscr{B}'|_{\ell+1}$, then

$$
\begin{aligned}
\widehat{\mathscr{B}} = \mathscr{B} \quad &\iff \quad \widehat{\mathscr{B}}|_{\ell+1} = \mathscr{B}|_{\ell+1} \\
&\iff \quad \dim \widehat{\mathscr{B}}|_{\ell+1} = \dim \mathscr{B}|_{\ell+1}
\end{aligned}
$$

$\mathscr{B}$ is identifiable from $w_{\mathrm{d}} \in \mathscr{B}|_T$ if and only if

$$
\operatorname{rank} \mathscr{H}_{\ell+1}(w_{\mathrm{d}}) = (\ell+1)m + n
$$

Nonparametric repr. $\mathscr{B}|_L = \text{image}\, \mathscr{H}_L(w_d)$

$\widehat{\mathscr{B}}_L \subseteq \mathscr{B}|_L$, $L \geq \ell$, equality holds if and only if

$$\text{rank}\, \mathscr{H}_L(w_d) = Lm + n \qquad \text{(GPE)}$$

sufficient conditions (the "fundamental lemma"):

1. $w_d = \begin{bmatrix} u_d \\ y_d \end{bmatrix}$
2. $\mathscr{B}$ controllable
3. $\mathscr{H}_{L+n}(u_d)$ full row rank

J.C. Willems et al., A note on persistency of excitation
Systems & Control Letters, (54)325–329, 2005

Problem formulation

given: "data" trajectory $(u_\mathrm{d}, y_\mathrm{d}) \in \mathscr{B}|_{T_\mathrm{d}}$ and $z \in \mathbb{C}$

find: $H(z)$, where H is the transfer function of $\mathscr{B}$

Direct data-driven solution

we are interested in trajectory

$$w = \begin{bmatrix} u \\ y \end{bmatrix} = \begin{bmatrix} \exp_z \\ \widehat{H} \exp_z \end{bmatrix} \in \mathscr{B}|_L, \quad \text{where } \exp_z(t) := z^t$$

using the data-driven representation, we have

$$\begin{bmatrix} \mathscr{H}_L(u_\mathrm{d}) \\ \mathscr{H}_L(y_\mathrm{d}) \end{bmatrix} g = \begin{bmatrix} \mathbf{z} \\ \widehat{H}\mathbf{z} \end{bmatrix}, \quad \text{where } \mathbf{z} := \begin{bmatrix} z^1 \\ \vdots \\ z^L \end{bmatrix}$$

which leads to the system

$$\begin{bmatrix} 0 & \mathscr{H}_L(u_\mathrm{d}) \\ -\mathbf{z} & \mathscr{H}_L(y_\mathrm{d}) \end{bmatrix} \begin{bmatrix} \widehat{H} \\ g \end{bmatrix} = \begin{bmatrix} \mathbf{z} \\ 0 \end{bmatrix} \tag{SYS}$$

Solution method: solve (SYS) for $\widehat{H}$

under (GPE) with $L \geq \ell + 1$, $\widehat{H} = H(z)$

without prior knowledge of ℓ

$$L = L_{\max} := \lfloor (T_{\mathrm{d}} + 1)/3 \rfloor$$

trivial generalization to

- multivariable systems
- multiple data trajectories $\{ w_{\mathrm{d}}^1, \dots, w_{\mathrm{d}}^N \}$
- evaluation of $H(z)$ at multiple points in $\{ z_1, \dots, z_K \} \in \mathbb{C}^K$

Comparison with classical nonparametric frequency response estimation methods

ignored initial/terminal conditions $\rightsquigarrow$ *leakage*

DFT grid $\rightsquigarrow$ limited *frequency resolution*

improvements by windowing and interpolation

- the leakage is not eliminated
- the methods involve *hyper-parameters*

Summary

why is the behavioral approach needed?

- respects physics
- suited for the data-driven paradigm

how is the behavioral approach different?

- a system is a set of trajectories — the behavior
- properties/problems are defined in terms of the behavior

what does it bring?

- broad framework where new questions can be asked
- data-driven representation $\mathscr{B}|_L = \text{image}\, \mathscr{H}_L(w_\text{d})$

"Telling people something they didn't know doesn't always mean surprising them. Sometimes it means telling them something they knew unconsciously but had never put into words. In fact those may be the more valuable insights, because they tend to be more fundamental."

P. Graham

What about noise in the data w_d? Solving (SYS) with noisy data

preprocessing: rank-$mL+n$ approx. of $\mathscr{H}_L(w_d)$

- hyper-parameters L and n $\quad (L \geq \ell + 1)$
- if the approximation preserves the Hankel structure, the method is maximum-likelihood in the EIV setting

regularization with $\|g\|_1$

- hyper-parameter: the 1-norm regularization parameter

regularization with the nuclear norm of $\mathscr{H}_L(\widehat{w_d})$

- hyper-parameters: L and the regularization parameter

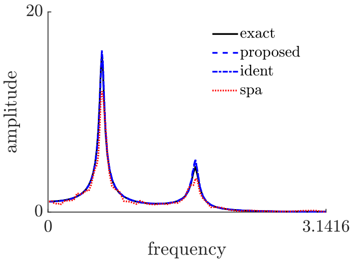
Matlab implementation

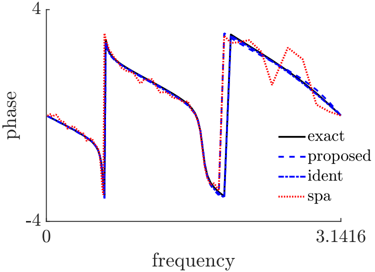
```
function Hh = dd_frest(ud, yd, z, n)
L = n + 1; t = (1:L)';
m = size(ud, 2); p = size(yd, 2);

%% preprocessing by low-rank approximation
H = [moshank(ud, L); moshank(yd, L)];
[U, ~, ~] = svd(H); P = U(:, 1:m * L + n);

%% form and solve the system of equations
for k = 1:length(z)
  A  = [[zeros(m * L, p); -kron(z(k) .^ t, eye(p))] P];
  hg = A \ [kron(z(k) .^ t, eye(m)); zeros(p * L, m)];
  Hh(:, :, k) = hg(1:p, :);
end
```

- 5 lines of essential code
- MIMO case, multiple evaluation points
- $L = n + 1$ in order to have a single hyper-parameter

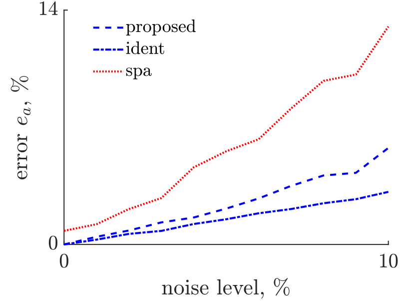
Empirical validation: 4th order system in the errors-in-variables setup

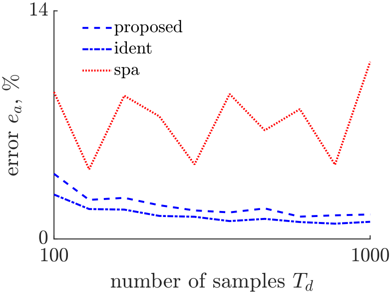
- `dd_frest` — proposed method
- `ident` — parametric maximum-likelihood estimator
- `spa` — nonparameteric estimator with Welch filter





Monte-Carlo simulation over different noise levels and number of samples





$$e_a := 100\% \cdot |(|\overline{H}_z| - |\widehat{H}_z|)| \; / \; |\overline{H}_z|$$