Behavioral approach to system identification and data-driven signal processing

Ivan Markovsky

ICREA

CIMNE

Outline

Direct data-driven design with exact data

Methods for dealing with noise in the data

Showcase: frequency response estimation

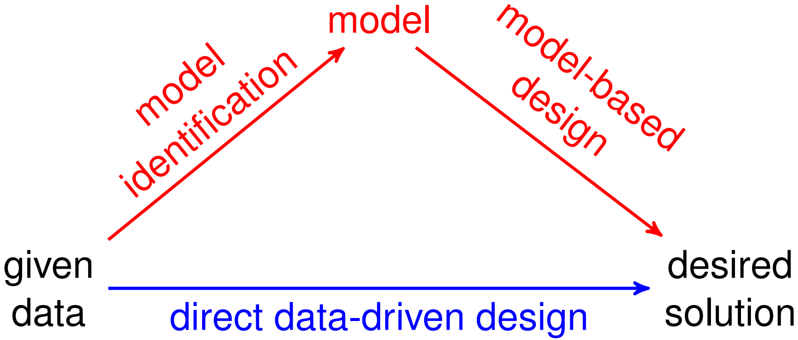
Outline

Direct data-driven design with exact data

Methods for dealing with noise in the data

Showcase: frequency response estimation

Our goal is direct data-driven methods for analysis and design of LTI systems



the classical approach is “indirect data-driven”

Motivation

no separation principle for model-based design

design objective not used in identification

incompatibility of identification and design

Data-driven does not mean model-free

data-driven problems do assume model

however, specific representation is not fixed

the methods we review are non-parametric

We use behavioral approach where dynamical system $\mathscr{B}$ is set of signals

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

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

$(\sigma w)(t) := w(t+1)$ — shift operator

$\sigma\mathscr{B} := \{\sigma w \mid w \in \mathscr{B}\}$

"good definition should formalize sensible intuition"

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

$\mathbf{m}(\mathscr{B})$ — number of inputs

$\mathbf{n}(\mathscr{B})$ — order ($=$ minimal state dimension)

$\boldsymbol{\ell}(\mathscr{B})$ — 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

$\mathscr{B}_1$ less complex than $\mathscr{B}_2 \iff \mathscr{B}_1 \subset \mathscr{B}_2$

in the LTI case, complexity $\leftrightarrow$ dimension

complexity: (# inputs, order, lag)

$$\mathbf{c}(\mathscr{B}) := \big(\mathbf{m}(\mathscr{B}), \mathbf{n}(\mathscr{B}), \boldsymbol{\ell}(\mathscr{B})\big)$$

$\mathscr{L}_c$ — bounded complexity LTI model class

Data-driven representation (infinite horizon)

data: exact infinite trajectory w_{d} of $\mathscr{B} \in \mathscr{L}$

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

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

Data-driven representation (finite horizon)

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

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

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

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

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

Conditions for informativity of the data

$\mathscr{B}|_L = \text{image}\, \mathscr{H}_L(w_\text{d})$ if and only if

$$\text{rank}\, \mathscr{H}_L(w_\text{d}) = L\mathbf{m}(\mathscr{B}) + \mathbf{n}(\mathscr{B}) \qquad \text{(GPE)}$$

I. Markovsky and F. Dörfler, Identifiability in the Behavioral Setting, TAC, 2023

sufficient conditions (input design perspective):

1. $w_\text{d} = \begin{bmatrix} u_\text{d} \\ y_\text{d} \end{bmatrix}$
2. $\mathscr{B}$ controllable
3. $\mathscr{H}_{L+\mathbf{n}(\mathscr{B})}(u_\text{d})$ full row rank (PE)

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

PE — persistency of excitation, GPE — generalized PE

Generic data-driven problem: trajectory interpolation/approximation

given: "data" trajectory $w_{\mathrm{d}} \in \mathscr{B}|_T$

partially specified trajectory $w|_{I_{\text{given}}}$

($w|_{I_{\text{given}}}$ selects the elements of w, specified by I_{given})

aim: minimize over $\widehat{w}$ $\quad \| w|_{I_{\text{given}}} - \widehat{w}|_{I_{\text{given}}} \|$

subject to $\widehat{w} \in \mathscr{B}|_L$

$$\widehat{w} = \mathscr{H}_L(w_{\mathrm{d}}) \big(\mathscr{H}_L(w_{\mathrm{d}})|_{I_{\text{given}}} \big)^{+} w|_{I_{\text{given}}} \qquad \text{(SOL)}$$

Special cases

simulation

- given data: initial condition and input
- to-be-found: output (exact interpolation)

smoothing

- given data: noisy trajectory
- to-be-found: ℓ_2-optimal approximation

tracking control

- given data: to-be-tracked trajectory
- to-be-found: ℓ_2-optimal approximation

Generalizations

multiple data trajectories $w_d^1, \ldots, w_d^N$

$$\mathscr{B} = \text{image} \begin{bmatrix} \mathscr{H}_L(w_d^1) & \cdots & \mathscr{H}_L(w_d^N) \end{bmatrix}$$

w_d not exact / noisy

maximum-likelihood estimation
$\rightsquigarrow$ Hankel structured low-rank approximation/completion
nuclear norm and ℓ_1-norm relaxations
$\rightsquigarrow$ nonparametric, convex optimization problems

nonlinear systems

results for special classes of nonlinear systems:
Volterra, Wiener-Hammerstein, bilinear, ...

Summary: data-driven signal processing

data-driven representation

leads to general, simple, practical methods

interpolation/approximation of trajectories

simulation, filtering and control are special cases
assumes only LTI dynamics; no hyper parameters

dealing with noise and nonlinearities

nonlinear optimization
convex relaxations

Outline

Direct data-driven design with exact data

Methods for dealing with noise in the data

Showcase: frequency response estimation

The data w_d being exact vs inexact / "noisy"

w_d exact and satisfying (GPE)

- "system theory" problems
- image $\mathscr{H}_L(w_\text{d})$ is nonparametric finite-horizon model
- data-driven solution $=$ model-based solution

w_d inexact, due to noise and/or nonlinearities

- naive approach: apply the solution (SOL) for exact data
- rigorous: assume noise model $\rightsquigarrow$ ML estimation problem
- heuristics: convex relaxations of the ML estimator

The maximum-likelihood estimation problem in the errors-in-variables setup is nonconvex

errors-in-variables setup: $w_\mathrm{d} = \overline{w}_\mathrm{d} + \widetilde{w}_\mathrm{d}$

- $\overline{w}_\mathrm{d}$ — true data, $\overline{w}_\mathrm{d} \in \mathscr{B}|_T$, $\mathscr{B} \in \mathscr{L}_c^q$
- $\widetilde{w}_\mathrm{d}$ — zero mean, white, Gaussian measurement noise

ML problem: given w_d, c, and $w|_{I_\text{given}}$

$$\begin{aligned} \underset{g}{\text{minimize}} \quad & \| w|_{I_\text{given}} - \mathscr{H}_L(\widehat{w}_\mathrm{d}^*)|_{I_\text{given}} g \| \\ \text{subject to} \quad & \widehat{w}_\mathrm{d}^* = \arg\min_{\widehat{w}_\mathrm{d}, \widehat{\mathscr{B}}} \quad \| w_\mathrm{d} - \widehat{w}_\mathrm{d} \| \\ & \qquad \text{subject to} \quad \widehat{w}_\mathrm{d} \in \widehat{\mathscr{B}}|_T \text{ and } \widehat{\mathscr{B}} \in \mathscr{L}_c^q \end{aligned}$$

The ML estimation problem is equivalent to Hankel structured low-rank approximation

$$\begin{aligned}
&\underset{g}{\text{minimize}} \quad \| w|_{I_\text{given}} - \mathscr{H}_L(\widehat{w}_\text{d}^*)|_{I_\text{given}} g \| \\
&\text{subject to} \quad \widehat{w}_\text{d}^* = \arg\min_{\widehat{w}_\text{d}, \widehat{\mathscr{B}}} \quad \| w_\text{d} - \widehat{w}_\text{d} \| \\
&\qquad\qquad\qquad \text{subject to} \quad \widehat{w}_\text{d} \in \widehat{\mathscr{B}}|_T \text{ and } \widehat{\mathscr{B}} \in \mathscr{L}_c^q
\end{aligned}$$

$$\Updownarrow$$

$$\begin{aligned}
&\underset{g}{\text{minimize}} \quad \| w|_{I_\text{given}} - \mathscr{H}_L(\widehat{w}_\text{d}^*)|_{I_\text{given}} g \| \\
&\text{subject to} \quad \widehat{w}_\text{d}^* = \arg\min_{\widehat{w}_\text{d}} \quad \| w_\text{d} - \widehat{w}_\text{d} \| \\
&\qquad\qquad\qquad \text{subject to} \quad \operatorname{rank} \mathscr{H}_{\ell+1}(\widehat{w}_\text{d}) \leq (\ell+1)m + n
\end{aligned}$$

Solution methods

local optimization

- choose a parametric representation of $\widehat{\mathscr{B}}(\theta)$
- optimize over $\widehat{w}$, $\widehat{w_\mathrm{d}}$, and θ
- depends on the initial guess

convex relaxation based on the nuclear norm

$$\text{minimize} \quad \text{over } \widehat{w}_\mathrm{d} \text{ and } \widehat{w} \quad \|w|_{I_\mathrm{given}} - \widehat{w}|_{I_\mathrm{given}}\| + \|w_\mathrm{d} - \widehat{w}_\mathrm{d}\| + \gamma \cdot \left\| \begin{bmatrix} \mathscr{H}_\Delta(\widehat{w}_\mathrm{d}) & \mathscr{H}_\Delta(\widehat{w}) \end{bmatrix} \right\|_*$$

convex relaxation based on ℓ_1-norm (LASSO)

$$\text{minimize} \quad \text{over } g \quad \|w|_{I_\mathrm{given}} - \mathscr{H}_L(w_\mathrm{d})|_{I_\mathrm{given}} g\| + \lambda \|g\|_1$$

Empirical validation on real-life datasets

	data set name	T	m	p
1	Air passengers data	144	0	1
2	Distillation column	90	5	3
3	pH process	2001	2	1
4	Hair dryer	1000	1	1
5	Heat flow density	1680	2	1
6	Heating system	801	1	1

G. Box, and G. Jenkins. Time Series Analysis: Forecasting and Control, Holden-Day, 1976

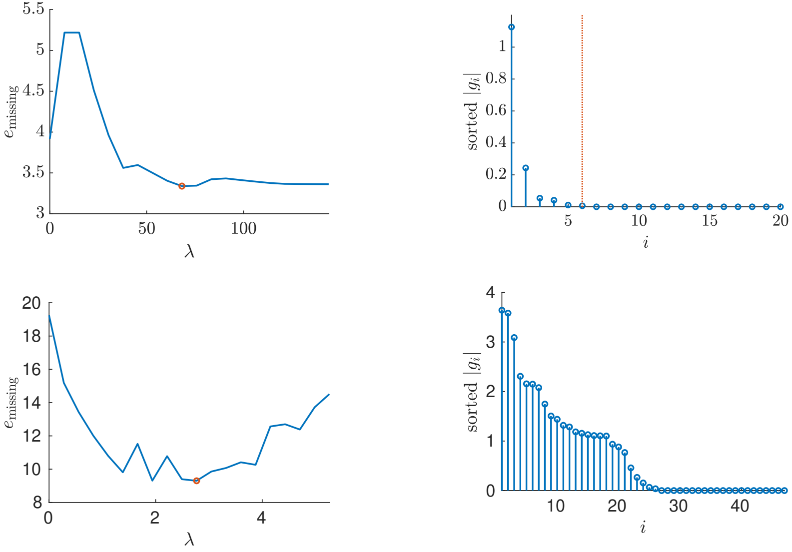
B. De Moor, et al.*DAISY: A database for identification of systems. Journal A, 38:4–5, 1997*

ℓ_1-norm regularization with optimized λ achieves the best performance

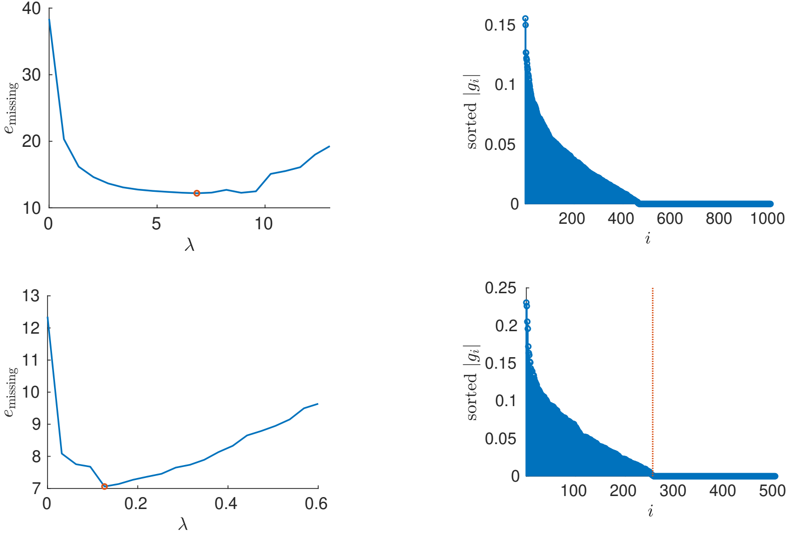
$$e_{\text{missing}} := \frac{\left\| w|_{I_{\text{missing}}} - \widehat{w}|_{I_{\text{missing}}} \right\|}{\left\| w|_{I_{\text{missing}}} \right\|} \, 100\%$$

	data set name	naive	ML	LASSO
1	Air passengers data	3.9	fail	3.3
2	Distillation column	19.24	17.44	9.30
3	pH process	38.38	85.71	12.19
4	Hair dryer	12.35	8.96	7.06
5	Heat flow density	7.16	44.10	3.98
6	Heating system	0.92	1.35	0.36

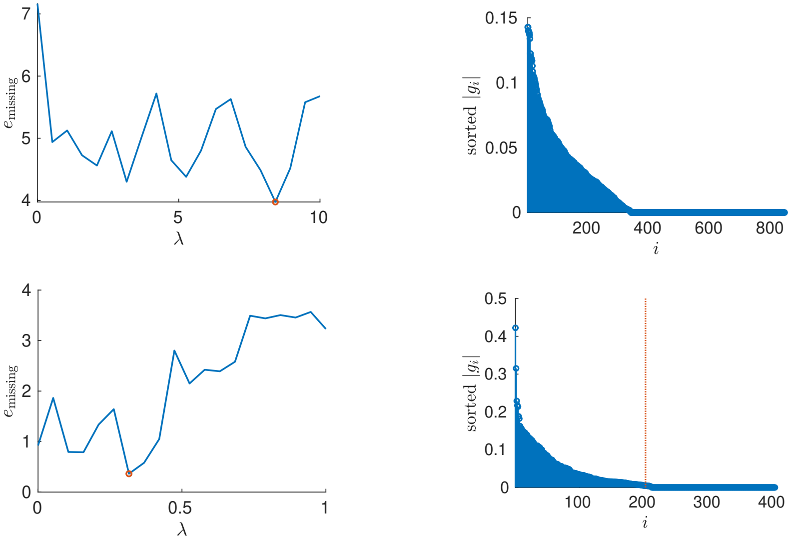
Tuning of λ and sparsity of g (datasets 1, 2)



Tuning of λ and sparsity of g (datasets 3, 4)



Tuning of λ and sparsity of g (datasets 5, 6)



Summary: convex relaxations

w_{d} exact $\rightsquigarrow$ system theory

- exact analytical solution
- current work: efficient real-time algorithms

w_{d} inexact $\rightsquigarrow$ nonconvex optimization

- subspace methods
- local optimization
- convex relaxations

empirical validation

- the naive approach works (surprisingly) well
- parametric local optimization is not robust
- ℓ_1-norm regularization gives the best results

Outline

Direct data-driven design with exact data

Methods for dealing with noise in the data

Showcase: frequency response estimation

Problem formulation

given: "data" trajectory $(u_\text{d}, y_\text{d}) \in \mathscr{B}|_{T_\text{d}}$ 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

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

which leads to the system

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

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*

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

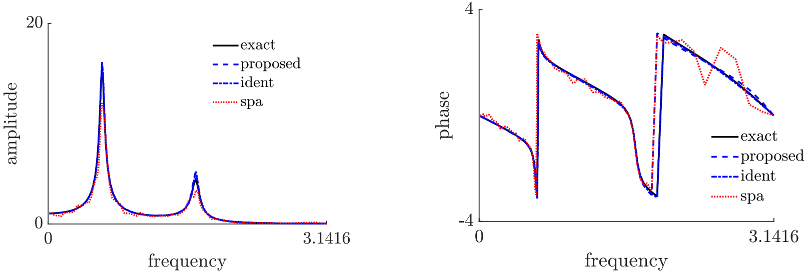
Matlab implementation

```
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))]
  hg = A \ [kron(z(k) .^ t, eye(m)); zeros(p * L, m
  Hh(:, :, k) = hg(1:p, :);
end
```

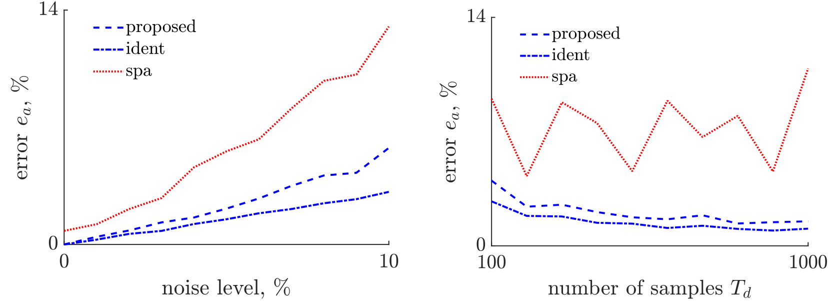
- 5 lines of essential code
- MIMO case, multiple evaluation points
- $L = n + 1$ in order to have a single hyper-parameter

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



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

Conclusions

detach the system from its representations

- define properties and problems in terms of the behavior
- lead to new, more general, definitions and problems
- avoid inconsistencies of the classical approach

separate problem from solution methods

- different representations lead to different methods
- show links among different methods
- lead to new solutions

naturally suited for the “data-driven paradigm”