Hidden structures in data-driven representations of dynamical systems

Ivan Markovsky

ICREA

CIMNE
INTERNATIONAL CENTRE
FOR NUMERICAL METHODS
IN ENGINEERING

From STLS to data-driven control, structured matrix algorithms are at the core

Nicola and I shared an office 25 years ago, working on the *structured total least-squares*

Nicola developed fast algorithms for STLS, I used STLS as tool in system identification

STLS $\rightsquigarrow$ SLRA $\rightsquigarrow$ data-driven methods
PhD $\rightsquigarrow$ postdoc $\rightsquigarrow$ current work

Outline

Why?	Generalization from finite data
What?	Analysis, signal processing, and control
How?	Algorithms

Outline

Why? Generalization from finite data

What? Analysis, signal processing, and control

How? Algorithms

Data-driven methods generalize from finite data under specified assumptions

generalization "given data $\mapsto$ all *possible* data" implies a data-generating mechanism (system)

the problem of finding it, however, is ill-posed

assumptions that make the problem well-posed

system: finite-dimensional linear time-invariant (LTI)
data: "informative" for identification

In this talk, the data are trajectories of discrete-time dynamical systems

$w = \big(w(1), \dots, w(T)\big)$ — T-samples trajectory

$w(t) \in \mathbb{R}^q$ — samples are vectors

$\mathscr{B}$ — set of all possible infinite trajectories
$\mathscr{B}|_T$ — set of all possible T-samples trajectories

Key insight: viewing systems as sets of trajectories (the behavioral approach)

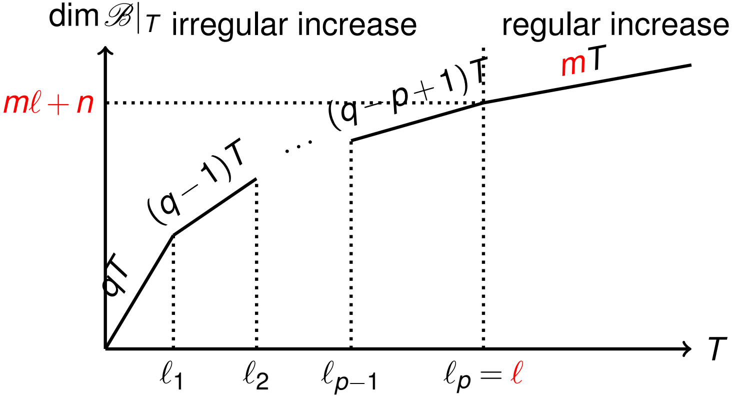
$w \in \mathscr{B}|_T \iff w$ is trajectory of $\mathscr{B}$

$\mathscr{B} \in \mathscr{L}$ — linear time-invariant system

$\mathscr{B}|_T \subset (\mathbb{R}^q)^T$ is subspace — linearity

$\sigma\mathscr{B} = \mathscr{B}$, $\sigma w(t) := w(t+1)$ — time-invariance

Finite-dimensional LTI systems have structure characterized by integers $(\ell_1, \ldots, \ell_p)$



$$\dim \mathscr{B}|_T = mT + n, \quad \text{for } T \geq \ell \qquad (\dim \mathscr{B}|_T)$$

$(\ell_1, \ldots, \ell_p)$ are properties of $\mathscr{B}$ and fully specify its structure

p — number of outputs (dependent variables)
$m = q - p$ — number of inputs (free variables)

$\ell := \max\{\ell_1, \ldots, \ell_p\}$ — lag
$n := \ell_1 + \cdots + \ell_p$ — order

$c := (m, \ell, n)$ — complexity of the system

$\mathscr{L}_{(m,\ell,n)}$ — bounded complexity LTI model class

Informativity of the data $w_\mathrm{d} \in \mathscr{B}|_{T_\mathrm{d}}$ is rank condition of Hankel matrix

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

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

generalized persistence of excitation condition

$$\operatorname{rank} \mathscr{H}_T(w_\mathrm{d}) = mT + n \qquad \text{(GPE)}$$

fact: let $w_\mathrm{d} \in \mathscr{B}|_{T_\mathrm{d}}$ and (GPE), then

$$\widehat{\mathscr{B}}_T := \operatorname{image} \mathscr{H}_T(w_\mathrm{d}) = \mathscr{B}|_T, \quad \text{for } T \geq \ell \quad \text{(DDR)}$$

Proof

$\widehat{\mathscr{B}}_T \subset \mathscr{B}|_T$ — follows by $w_d \in \mathscr{B}|_{T_d}$ and $\mathscr{B} \in \mathscr{L}$

(DDR) then follows by ($\dim \mathscr{B}|_T$) and (GPE)

$$\dim \widehat{\mathscr{B}}_T = \operatorname{rank} \mathscr{H}_T(w_d) = \dim \mathscr{B}|_T$$

notes

- in order to check (GPE) one needs to know (m, ℓ, n)
- however, for using (DDR), (m, ℓ, n) are not needed
- (GPE) implies a minimum number of samples $T_{\min}$

$$T_d \geq T_{\min} := (m+1)T + n - 1$$

What about noise in the data?

with noisy data $\mathscr{H}_T(w_\mathrm{d})$ is full row rank

denoising procedure:

$$\begin{array}{ll} \text{minimize} \quad \text{over } \widehat{w}_\mathrm{d} & \|w_\mathrm{d} - \widehat{w}_\mathrm{d}\| \\ \text{subject to} & \operatorname{rank} \mathscr{H}_T(\widehat{w}_\mathrm{d}) = mT + n \end{array}$$

$\rightsquigarrow$ Hankel structured low-rank approximation

in this case, m and n should be a priori given

Outline

Why? Generalization from finite data

What? Analysis, signal processing, and control

How? Algorithms

(DDR) leads to computational methods for:

identification (find representations from data)

e.g., subspace methods aim at (A, B, C, D)

system analysis (direct from data)

e.g., controllability and distance to uncontrollability

signal processing and control (direct from data)

e.g., simulation, smoothing, missing data recovery

Subspace identification methods use $\mathscr{H}_T(w_\mathrm{d})$ to find state-space representation

$$\widehat{\mathscr{B}}_T := \operatorname{image} \mathscr{H}_T(w_\mathrm{d}), \text{ assuming (GPE)}$$

MOESP-type methods

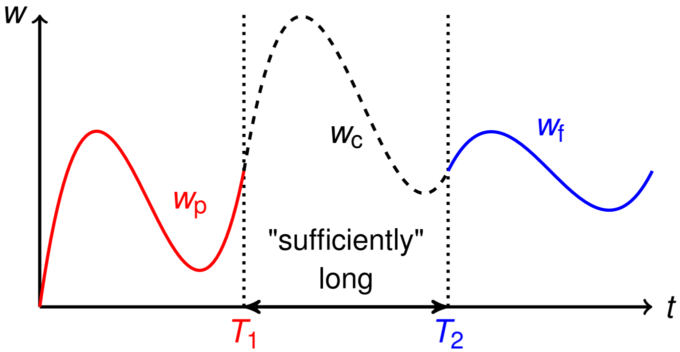
1. $\widehat{\mathscr{B}}_T \mapsto \widehat{\mathscr{B}}_{\mathrm{aut},T}$ — find autonomous sub-behavior (projection)
2. $\widehat{\mathscr{B}}_{\mathrm{aut},T} \mapsto (A, C)$ — solving system of linear equations
3. $(\widehat{\mathscr{B}}_T, A, C) \mapsto (B, D)$ — solving system of linear equations

N4SID-type methods

1. $\widehat{\mathscr{B}}_T \mapsto X$ — find state sequence (past/future intersection)
2. $(w_\mathrm{d}, X) \mapsto (A, B, C, D)$ — solving system of linear equations

Controllability is the property of “patching” any past trajectory with any future trajectory

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



Compare with the classical definition:

$$\mathscr{C} := \begin{bmatrix} B & AB & \cdots & A^{n-1}B \end{bmatrix} \text{ is full row rank}$$

property of the parameters (A, B)

lack of controllability could be due to “bad” state choice and/or deficiency of $\mathscr{B}$ to be controlled

how to quantify “distance” to uncontrollability?
two options: $\|[A\ B] - [\widehat{A}\ \widehat{B}]\|_{\mathrm{F}}$ and $\operatorname{dist}(\mathscr{B}, \widehat{\mathscr{B}})$

Smoothing and tracking are projections

the errors-in-variables Kalman smoother solves

$$\widehat{w}^* := \arg \min_{\widehat{w} \in \mathscr{B}|_T} \|w - \widehat{w}\|$$

LQ tracking is mathematically the same problem

Interpolation with LTI trajectories (missing data recovery)

given:

$$\begin{array}{lcl} w_\mathrm{d} \in \mathscr{B}|_{T_\mathrm{d}} & — & \text{“data” trajectory} \\ w|_{I_\mathrm{given}} & — & \text{partially specified trajectory} \end{array}$$

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

find:

$$\widehat{w} \in \mathscr{B}|_T, \quad \text{such that} \quad \widehat{w}|_{I_\mathrm{given}} = w|_{I_\mathrm{given}}$$

Outline

Why? Generalization from finite data

What? Analysis, signal processing, and control

How? Algorithms

The "main player" is orthonormal basis for the restricted behavior

$$\mathscr{B}|_T = \text{image}\, B_T, \quad B_T^\top B_T = I_r \qquad (B_T)$$

$\mathscr{B} \in \mathscr{L}_{(m,\ell,n)}$ and $T \geq \ell$ implies $r = mT + n$

$\mathscr{B}|_{\ell+1}$ defines $\mathscr{B} \implies$ representation of $\mathscr{B}$

nonparameteric representation of $\mathscr{B}|_T$

$$w \in \mathscr{B}|_T \iff w = B_T g,\ g \in \mathbb{R}^r$$

Core sub-problem is obtaining B_T from other representations and data

$$\mathscr{W}_\mathrm{d} = \{\, w_\mathrm{d}^1, \ldots, w_\mathrm{d}^N \,\} \;\mapsto\; B_T = \mathtt{orth}\, \mathscr{H}_T(w_\mathrm{d})$$

$$(A,B,C,D) \;\mapsto\; B_T = \mathtt{orth}\, \Pi \begin{bmatrix} 0_{mT\times n} & I_{mT} \\ \mathscr{O}_T(A,C) & \mathscr{T}_T(H) \end{bmatrix}$$

alternative indirect approach:

$$(A,B,C,D) \xmapsto{\text{simulation}} \mathscr{W}_\mathrm{d} \xmapsto{\mathtt{orth}\, \mathscr{H}_T(w_\mathrm{d})} B_T$$

Dual of (DDR) is the kernel representation

$$\mathscr{B} = \ker R(\sigma), \quad \begin{bmatrix} R^1(z) \\ \vdots \\ R^g(z) \end{bmatrix} = \begin{bmatrix} R_0^1 + R_1^1 z + \cdots + R_{\ell_1}^1 z^{\ell_1} \\ \vdots \\ R_0^g + R_1^g z + \cdots + R_{\ell_g}^g z^{\ell_g} \end{bmatrix} \tag{KER}$$

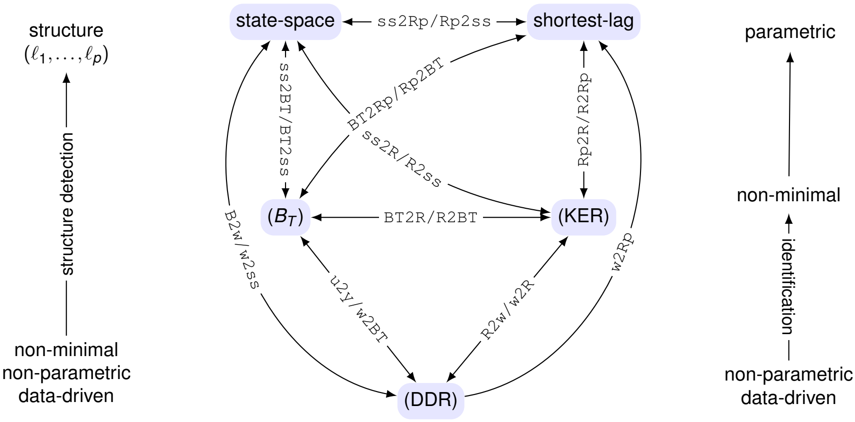
multiplication matrix (generalized Sylvester)

$$\mathscr{M}_T(R) := \begin{bmatrix} \mathscr{M}_T(R^1) \\ \vdots \\ \mathscr{M}_T(R^g) \end{bmatrix}, \quad \mathscr{M}_T(R^i) := \begin{bmatrix} R_0^i & R_1^i & \cdots & R_{\ell_i}^i & & \\ & \ddots & \ddots & & \ddots & \\ & & R_0^i & R_1^i & \cdots & R_{\ell_i}^i \end{bmatrix}$$

assuming $\operatorname{rank} R = \mathbf{p}(\mathscr{B})T - \mathbf{n}(\mathscr{B})$,

$$\mathscr{B}|_T = \ker \mathscr{M}_T(R), \quad \text{for } T > \ell \tag{FHK}$$

Transitions among state-space, kernel, and data-driven representations



Core “low-level” linear algebra operations

rank and orthonormal basis for image/kernel

for Hankel/multiplication structured matrices

in case of noisy data, the problem is SLRA

Summary

well-posedness assumptions:
LTI structure + data informativity

data-driven representation: $\widehat{\mathscr{B}} := \text{image}\, \mathscr{H}_T(w_\text{d})$
Hankel SLRA as denoising procedure

leads to new problems, methods, and insights

- distance to uncontrollability
- interpolation of LTI trajectories
- decoupling problem statements from solution methods