Data-driven modeling:
A low-rank approximation problem

Ivan Markovsky

Vrije Universiteit Brussel

Outline

Setup: data-driven modeling

Problems: system identification, machine learning, ...

Behavioral paradigm $\leftrightarrow$ low-rank approximation

Algorithms: optimization, multistage, convex relaxations

Applications: missing data, data-driven simulation

Connections: TLS, EIV, PCA, rank minimization, ...

General setup

$$\underset{\mathscr{D} \subset \mathscr{U}}{\text{data}} \quad \xrightarrow{\text{modeling}} \quad \underset{\mathscr{B} \in \mathscr{M} \in 2^{\mathscr{U}}}{\text{model}}$$

- $\mathscr{D}$ — data, *e.g.*, a vector time series $(\mathbb{R}^q)^{\mathbb{N}}$
- $\mathscr{B}$ — model (behavior): a (sub)set of the data space $\mathscr{U}$
- $\mathscr{M}$ — model class: a set of models

work plan:

1. define a modeling problem
2. find an algorithm that solves the problem
3. implement the algorithm in software
4. use the software in applications

The problem

prior knowledge, assumptions, and/or prejudices

about what the true or desirable model is

- model class — imposes hard constraints

  e.g., bound on the model complexity

- optimization criteria — impose soft constraints

  e.g., small misfit between the model and the data

- real-life problems are vaguely formulated

- often it is not clear what is the “best” problem formulation

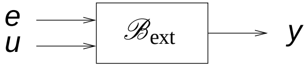
  “A well defined problem is a half solved problem.”

System identification problems

$$\mathscr{U} = (\underbrace{\mathbb{R}^q \times \cdots \times \mathbb{R}^q}_{T_1}) \times \cdots \times (\underbrace{\mathbb{R}^q \times \cdots \times \mathbb{R}^q}_{T_N}) \quad \text{— } N\text{, } q\text{-variable time series}$$

$\mathscr{M}$ is, *e.g.*, bounded complexity (# inputs and lags), LTI systems

- latency (ARMAX):



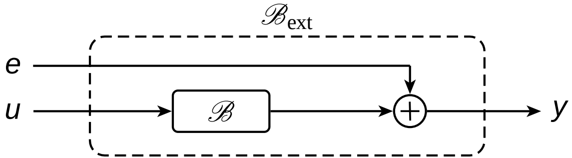
$$\text{minimize} \quad \|e\| \quad \text{subject to} \quad \big((e,u),y\big) \in \widehat{\mathscr{B}}_{\text{ext}} \in \mathscr{M}$$

- misfit (EIV):

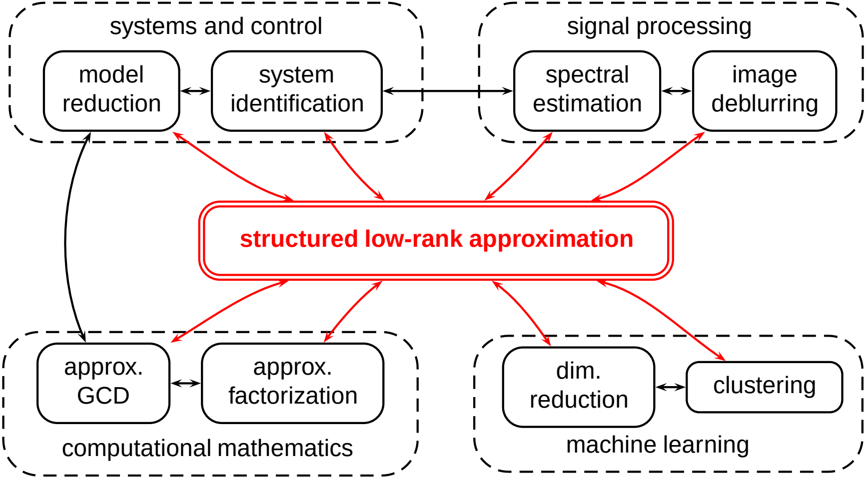
$$\text{minimize} \quad \|(\Delta u, \Delta y)\| \quad \text{subject to} \quad \big(\underbrace{u + \Delta u}_{\widehat{u}}, \underbrace{y + \Delta y}_{\widehat{y}}\big) \in \widehat{\mathscr{B}} \in \mathscr{M}$$

Special cases

- $\mathscr{M}$ with lag $= 0$ $\rightsquigarrow$ static modeling
- $\mathscr{M}$ with # inputs $= 0$ $\rightsquigarrow$ sum-of-damped-exp. modeling
- FIR systems $\rightsquigarrow$ approximate deconvolution
- EIV with $\Delta u = 0$ or special ARMAX $\rightsquigarrow$ output error



A unifying setting for data modeling



Desirable features of a paradigm

simple:	can be introduced in 1 slide
flexible:	applies to a rich class of problems
practical:	leads to solution methods and algorithms
optimal:	in theory, finds the "best" solution
effective:	in practice, can “solve” real-life problems
automatic:	hyper param. correspond to prior knowledge
compact:	software implementation requires short code

Structured low-rank approximation

- structure specification $\mathscr{S} : \mathbb{R}^{n_p} \to \mathbb{R}^{m\times n}$
- vector of structure parameters $p \in \mathbb{R}^{n_p}$
- weighted 2-norm $\|p\|_W^2 := p^\top W p$
- rank specification r

$$\begin{array}{lll} \text{minimize} & \text{over } \widehat{p} \in \mathbb{R}^{n_p} & \|p - \widehat{p}\|_W^2 \\ \text{subject to} & \operatorname{rank}\big(\mathscr{S}(\widehat{p})\big) \le r & \end{array} \qquad \text{(SLRA)}$$

Structure $\mathscr{S}$ $\leftrightarrow$ Model class $\mathscr{M}$

unstructured	$\leftrightarrow$	linear static
Hankel	$\leftrightarrow$	scalar LTI
$q \times 1$ Hankel	$\leftrightarrow$	q-variate LTI
$q \times N$ Hankel	$\leftrightarrow$	N equal length traj.
mosaic Hankel	$\leftrightarrow$	N general traj.
$\begin{bmatrix}\text{Hankel} & \text{unstructured}\end{bmatrix}$	$\leftrightarrow$	finite impulse response
block-Hankel Hankel-block	$\leftrightarrow$	2D linear shift-invariant

(SLRA) $\leftrightarrow$ approximate data modeling

- $p \leftrightarrow \text{vec}(\mathscr{D})$
- $r \leftrightarrow$ model complexity
- $W \leftrightarrow$ prior knowledge about the data accuracy

(SLRA) is a maximum likelihood estimator in the EIV setting

Singular weight matrix $\leftrightarrow$ fixed and missing values

- consider the special case of element-wise weights

$$\|p - \widehat{p}\|_w = \sqrt{\textstyle\sum_{i=1}^{n_p} w_i (p_i - \widehat{p}_i)^2}$$

specified by a vector $w \in \mathbb{R}^{n_p}$

- $w_i = \infty$ imposes equality constraint $\widehat{p}_i = p_i$ on (SLRA)

$$w_i = \infty \qquad \Longrightarrow \qquad \widehat{p}_i = p_i$$

- $w_i = 0$ makes the problem (SLRA) independent of p_i

$$w_i = 0 \qquad \Longrightarrow \qquad p_i \text{ is ignored}$$

alternatively, problem (SLRA) is solved with p_i missing

Solution methods

- global solution methods
  - SDP relaxations of rational function minimization problem
  - systems of polynomial equations (computer algebra)
    - resultant-based methods
    - Stetter-Moller methods
    - subdivision methods
    - homotopy continuation
- local optimization methods
  - variable projections
  - alternating projections
  - variations

parameterization

$+$

optimization method

$=$

method

- heuristics
  - multistage methods
  - nuclear norm heuristic

VARPRO-like solution method

- using the kernel parameterization

$$\operatorname{rank}\big(\mathscr{S}(\widehat{p})\big) \leq r \quad \iff \quad R\mathscr{S}(\widehat{p}) = 0, \quad \operatorname{rank}(R) = m - r$$

- (SLRA) becomes

$$\begin{aligned} &\text{minimize} \quad \text{over } \widehat{p} \text{ and } R \quad \|p - \widehat{p}\|_w^2 \\ &\text{subject to} \quad R\mathscr{S}(\widehat{p}) = 0, \ \operatorname{rank}(R) = m - r \end{aligned} \qquad (\text{SLRA}_R)$$

- (SLRA_R) is separable in $\widehat{p}$ and R, *i.e.*, it is equivalent to

$$\begin{aligned} &\text{minimize} \quad \text{over } R \quad f(R) \\ &\text{subject to} \quad \operatorname{rank}(R) = m - r \end{aligned} \qquad (\text{OUTER})$$

where

$$f(R) := \min_{\widehat{p}} \quad \|p - \widehat{p}\|_w^2 \quad \text{subject to} \quad R\mathscr{S}(\widehat{p}) = 0 \quad (\text{INNER})$$

- $\widehat{p}$ is eliminated (projected out) of (SLRA_R)

- evaluation of $f(R)$, *i.e.*, solving (INNER), is least norm prob.
- in SYSID, evaluation of $f(R)$ is a data smoothing operation
- in a stochastic setting, it is the likelihood evaluation
- efficient computation using Riccati recursion (Kalman smoothing)
- in other applications, $f(R)$ can also be evaluate efficiently, by exploiting the matrix structure
- software implementation for mosaic Hankel-like matrices, with fixed and missing data, and linearly structured kernel

  `http://github.com/slra/slra`

Structured kernel

- (OUTER) is a nonlinear least-squares problem
- it can be solved with additional constraints
- *e.g.*, linear structure of the kernel

$$R = \mathscr{R}(\theta) := \text{vec}^{-1}(\theta\Psi)$$

- applications requiring structured kernel:
  - harmonic retrieval $\rightsquigarrow$ R palindromic
  - SYSID with fixed poles $\rightsquigarrow$ $R = R_{\text{fixed}} \star R_{\text{free}}$
  - SYSID with fixed observ. indices $\rightsquigarrow$ $R = \begin{bmatrix} \times & \cdots & \times & 1 & 0 & 0 \\ \vdots & \ddots & & \ddots & \ddots & 0 \\ \times & \cdots & \times & \cdots & \times & 1 \end{bmatrix}$
  - common dynamics estimation $\rightsquigarrow$ $\mathscr{R}$ nonlinear

Autonomous system identification with missing data

- $\mathscr{M} = \mathscr{L}_{0,\ell}$ — LTI systems with 0 inputs and lag $\leq \ell$
- data $y \in \underbrace{\mathbb{R}^{\mathtt{p}}_{\text{ext}} \times \cdots \times \mathbb{R}^{\mathtt{p}}_{\text{ext}}}_{T}$, where $\mathbb{R}_{\text{ext}} = \mathbb{R} \cup \mathtt{NaN}$
- problem: given y and ℓ,

$$\begin{aligned} &\text{minimize} \quad \text{over } \widehat{y} \in (\mathbb{R}^{\mathtt{p}})^T \text{ and } \widehat{\mathscr{B}} \quad \|y - \widehat{y}\|_w^2 \\ &\text{subject to} \quad \widehat{y} \in \widehat{\mathscr{B}} \in \mathscr{L}_{0,\ell} \end{aligned}$$

- w assigns zeros to the missing data ($y_i(t) = \mathtt{NaN}$)
- $\exists \widehat{\mathscr{B}}$, such that $\widehat{y} \in \widehat{\mathscr{B}} \in \mathscr{L}_{0,\ell} \iff \operatorname{rank}\big(\mathscr{H}_{\ell+1}(\widehat{y})\big) \leq \ell\mathtt{p}$
- the problem is Hankel structured low-rank approximation

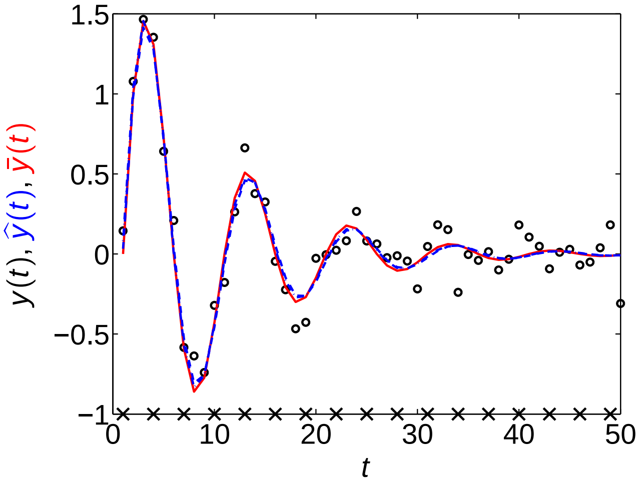
Simulation example

- $p = 1$, $\ell = 2$, $T = 50$, $y = \bar{y} +$ white noise, where

  $$\bar{y}(t) = 1.456\bar{y}(t-1) - 0.81\bar{y}(t-2), \qquad \bar{y}(0) = 0, \quad \bar{y}(1) = 1$$

- missing values distributed periodically with period 3
- solved with the algorithm based on the VARPRO approach

System identification with periodically missing data



true — solid line optimal approximation — dashed blue

circles — data points crosses — location of missing data

Classical simulation problem

given

- LTI system $\mathscr{B}$ (specified by some representation)
- initial condition w_{ini} (specified by trajectory of $\mathscr{B}$)
- input u

find the output y of $\mathscr{B}$, corresponding to w_{ini} and u

- there are many ways to solve the problem
- the algorithms depend on the model representation (state-space, transfer function, impulse response, ...)

Data-driven simulation

given

- trajectory w' of LTI system $\mathscr{B}$ and the lag ℓ of $\mathscr{B}$
- initial condition $w''_{\mathrm{p}} = \big(w''(1),\dots,w''(\ell)\big)$
- input $u''_{\mathrm{f}} = \big(u''(\ell+1),\dots,u''(T_2)\big)$

find the output y''_{f} of $\mathscr{B}$, corresponding to w''_{p} and u''

$$y''_{\mathrm{f}} = \big(y''(\ell+1),\dots,y''(T_2)\big)$$

$$\Updownarrow$$

$$\begin{array}{llll} \text{find} & y''_{\mathrm{f}} & \text{and} & \mathscr{B} \in \mathscr{L}_{\mathtt{m},\ell} \\ \text{such that} & w' \in \widehat{\mathscr{B}} & \text{and} & \underbrace{w''_{\mathrm{p}} \wedge (u''_{\mathrm{f}}, y''_{\mathrm{f}})}_{w''} \in \mathscr{B} \end{array}$$

- there is $\widehat{\mathscr{B}} \in \mathscr{L}_{\mathtt{m},\ell}$, such that $w' \in \widehat{\mathscr{B}}$ and $w'' \in \widehat{\mathscr{B}}$

$$\Updownarrow$$

$$\operatorname{rank}\left(\begin{bmatrix} \mathscr{H}_{\ell+1}(w') & \mathscr{H}_{\ell+1}(w'') \end{bmatrix}\right) \leq 2\ell + 1$$

mosaic Hankel matrix completion

- with noisy w', the problem is

$$\begin{array}{ll} \text{minimize} & \text{over } \widehat{w}', \widehat{w}'', \widehat{\mathscr{B}} \in \mathscr{L}_{\mathtt{m},\ell} \quad \|w' - \widehat{w}'\|_2^2 \\ \text{subject to} & \widehat{w}', \widehat{w}'' \in \widehat{\mathscr{B}}, \quad \widehat{w}''_{\mathrm{p}} = w''_{\mathrm{p}}, \quad \widehat{u}''_{\mathrm{f}} = u''_{\mathrm{f}} \end{array}$$

mosaic Hankel low-rank approximation
with exact and missing data

Simulation example

- second order SISO system, defined by difference equation

$$\bar{y}(t) = 1.456\bar{y}(t-1) - 0.81\bar{y}(t-2) + \bar{u}(t) - \bar{u}(t-1)$$

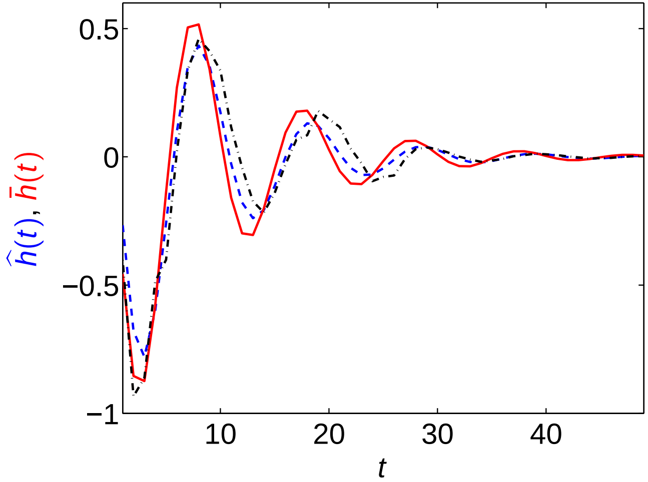
- w' is noisy trajectory generated from random input

- y''_{f} is the impulse response $\bar{h}$, *i.e.*,

$$u'' = (\underbrace{0,\ldots,0}_{\ell},\underbrace{1,0,\ldots,0}_{\text{pulse input}})$$

$$y'' = (\underbrace{0,\ldots,0}_{\ell},\underbrace{\widehat{h}(0),\widehat{h}(1),\ldots,\widehat{h}(T_2-\ell-1)}_{\text{impulse response}})$$

Data-driven simulation of impulse response



true — solid line   optimal approximation — dashed blue

Related frameworks

- behavioral approach: representation free modeling
- total least squares: (SLRA) with I/O representation

$$R\mathscr{S}(\widehat{p}) = \begin{bmatrix} X^\top & -I \end{bmatrix} \begin{bmatrix} \widehat{A}^\top \\ \widehat{B}^\top \end{bmatrix} = 0 \quad \iff \quad \widehat{A}X = \widehat{B} \quad \text{(TLS)}$$

- errors-in-variables: statistical setup for (TLS)
- principal component analysis: another statistical setup
- rank minimization: “dual” to (SLRA)
  (soft constraint on complexity, hard constraint on accuracy)

Current/future work

- comparison of different optimization methods for

$$\text{minimize} \quad \text{over } R \quad M(R) \quad \text{subject to} \quad RR^\top = I$$

- fast misfit computation $\leftrightarrow$ Kalman smoothing

$$M(R) = \text{vec}^\top(w)\Gamma^{-1}(R)\,\text{vec}(w)$$

  efficient computation in the case of missing data

- singularity of Γ (poles/zeros on the unit circle)
- solution of ARMAX identification problems (latency min.)
- static nonlinear modeling is nonlinear SLRA (kernel PCA)
- nD system identification (block-Hankel Hankel-block SLRA)