# Applications of structured low-rank approximation

Ivan Markovsky

School of Electronics and Computer Science
University of Southampton

## Two well known examples

- System realisation
- Discrete deconvolution

## System realisation

The sequence

$$h := \big(h(0), h(1), \dots\big), \qquad h(t) \in \mathbb{R}^{\mathtt{p}\times\mathtt{m}}$$

is realisable by a finite dim. LTI system, if and only if

$$\mathscr{H}(h) := \begin{bmatrix} h(1) & h(2) & h(3) & \cdots \\ h(2) & h(3) & ⋰ & \\ h(3) & ⋰ & & \\ \vdots & & & \end{bmatrix}$$

has finite rank. Moreover,

$$\begin{aligned} \operatorname{rank}\big(\mathscr{H}(h)\big) &= \text{state dim. of a minimal realisation of } h \\ &= \text{complexity of an exact LTI model for } h. \end{aligned}$$

## Approximate realisation $=$ Model reduction

However, rank deficiency is a nongeneric property (in $\mathbb{Z}_+ \to \mathbb{R}^{\mathtt{p}\times\mathtt{m}}$).

Rank is computed numerically most reliably by the SVD.

From a system theoretic point of view

the SVD does model reduction (Kung's algorithm).

The truncated SVD gives (2-norm) optimal unstructured approx.

Instead, we are aiming at a

structured rank-n approximation of $\mathscr{H}(h)$:

Find $\widehat{h}$, such that $\|h - \widehat{h}\|$ is minimized and $\operatorname{rank}\big(\mathscr{H}(\widehat{h})\big) = \mathtt{n}$.

## Approximate realisation (model reduction) $\Updownarrow$ Hankel structured low-rank approximation

The approximate realisation (model reduction) problem is

> Given $h := \big(h(0), h(1), \dots\big)$ and $\mathtt{n} \in \mathbb{N}$, find
>
> $$\min_{\widehat{h}} \|h - \widehat{h}\| \quad \text{subject to} \quad \operatorname{rank}\big(\mathscr{H}(\widehat{h})\big) \le \mathtt{n}$$

a Hankel structured low-rank approximation (SLRA) problem.

Unfortunately, this problem is NP-complete.

## Deconvolution

Consider the finite sequences

$$h := \big(h(0), h(1), \dots, h(\mathtt{n})\big), \quad \text{where} \quad h \in \mathbb{R}^{\mathtt{p}\times\mathtt{m}}$$

$$u := \big(u(-\mathtt{n}), \dots, u(0), u(1) \dots, u(T)\big) \quad \text{and} \quad y := \big(y(1), \dots, y(T)\big).$$

Define $\operatorname{row}(y) := \begin{bmatrix} y(1) & \cdots & y(T) \end{bmatrix}$ and the Toeplitz matrix

$$\mathscr{T}_{\mathtt{n}+1}(u) := \begin{bmatrix} u(1) & u(2) & u(3) & \dots & u(T) \\ u(0) & u(1) & u(2) & \dots & u(T-1) \\ \vdots & \vdots & \vdots & & \vdots \\ u(-\mathtt{n}) & u(1-\mathtt{n}) & u(2-\mathtt{n}) & \cdots & u(T-\mathtt{n}) \end{bmatrix}$$

With this notation,

$$\underset{\text{(convolution)}}{y = h \star u} \quad \Longleftrightarrow \quad \underset{\text{(linear algebra)}}{\operatorname{row}(y) = \operatorname{row}(h)\mathscr{T}_{\mathtt{n}+1}(u)}$$

## Exact and approximate deconvolution

Exact deconv. problem: Given $u$ and $y$, find $h$, such that $y = h \star u$.

Solution exists if and only if the system of equations

$$\operatorname{row}(y) = \operatorname{row}(h)\mathscr{T}_{\mathtt{n}+1}(u)$$

is solvable for $h$. However with $T > (\mathtt{n}+1)\mathtt{m}$, generically solution does not exist $\leadsto$ approximate deconvolution problem:

> Given $u$, $y$, and $\mathtt{n} \in \mathbb{N}$, find
>
> $$\min_{\widehat{u},\, \widehat{y},\, \widehat{h}} \|\operatorname{col}(u, y) - \operatorname{col}(\widehat{u}, \widehat{y})\| \quad \text{subject to}$$
>
> $$\operatorname{row}(\widehat{y}) = \operatorname{row}(\widehat{h})\mathscr{T}_{\mathtt{n}+1}(\widehat{u})$$

## Deconvolution $=$ FIR system identification

We can interpret

$$y = h \star u$$

as the response of an FIR system with impulse response $h$ to

- initial conditions $\big(u(-\mathtt{n}), \dots, u(0)\big)$, and
- input $\big(u(1) \dots, u(T)\big)$.

Then the deconvolution problem has the meaning of an FIR system identification problem:

> Given initial condition, input, and output, find an FIR model.

- exact deconvolution $\implies$ exact FIR fitting model
- approx. deconvolution $\implies$ approx. FIR fitting model

The parameter $\mathtt{n}$ bounds the FIR model complexity.

## Approximate deconvolution ⇝ SLRA

Assuming that $\mathscr{T}_{n+1}(\widehat{u})$ is full rank (persistency of excitation),

$$\text{row}(\widehat{y}) = \text{row}(\widehat{h})\mathscr{T}_{n+1}(\widehat{u}) \iff \text{rank}\left(\begin{bmatrix}\mathscr{T}_{n+1}(\widehat{u}) \\ \text{row}(\widehat{y})\end{bmatrix}\right) = (n+1)m$$

Then the approximate deconvolution problem can be written as

Given $u$, $y$, and $n \in \mathbb{N}$, find

$$\min_{\widehat{u},\,\widehat{y}} \|\text{col}(u,y) - \text{col}(\widehat{u},\widehat{y})\| \quad \text{subject to}$$

$$\text{rank}\left(\begin{bmatrix}\mathscr{T}_{n+1}(\widehat{u}) \\ \text{row}(\widehat{y})\end{bmatrix}\right) \leq (n+1)m$$

a SLRA problem with structure composed of two blocks:
Toeplitz block above an unstructured block.

## Rank of the data matrix

complexity of an exact model fitting the data ↔ rank of the data matrix

- order of the realization $=$ $\text{rank}\,(\mathscr{H}(h))$
- number of taps of an FIR system $=$ $\text{rank}\left(\begin{bmatrix}\mathscr{T}_{n+1}(u) \\ \text{row}(y)\end{bmatrix}\right)/m - 1$

## Main issue: Low-rank approximation

With a bounding on the model complexity,

generically in the data space, exact property does not hold

$\implies$ an approximation is needed.

Approximation paradigm:

modify the data as little as possible, so that the exact property holds for the modified data.

This paradigm leads to structured low-rank approximation.

## Structured low-rank approximation

Given

- a vector $p \in \mathbb{R}^{n_p}$,
- a mapping $\mathscr{S} : \mathbb{R}^{n_p} \to \mathbb{R}^{m \times n}$ (structure specification)
- a vector norm $\|\cdot\|$, and
- an integer $r$, $0 < r < \min(m,n)$,

find

$$\widehat{p}^* := \arg\min_{\widehat{p}} \|p - \widehat{p}\| \quad \text{subject to} \quad \text{rank}\,(\mathscr{S}(\widehat{p})) \leq r. \qquad (*)$$

Interpretation:

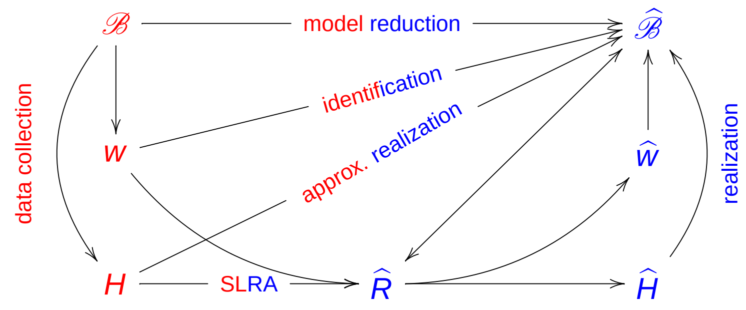
$\widehat{D}^* := \mathscr{S}(\widehat{p}^*)$ is optimal rank-$r$ (or less) approx. of $D := \mathscr{S}(p)$,
within the class of matrices with the same structure as $D$.

# Applications

- System theory
  1. Approximate realization
  2. Model reduction
  3. Errors-in-variables system identification
  4. Output error system identification
  5. Frequency domain system identification
- Signal processing
  6. Output only (autonomous) system identification
  7. Finite impulse response (FIR) system identification
  8. Harmonic retrieval
  9. Image deblurring
- Computer algebra
  10. Approximate greatest common divisor (GCD)

# System theory applications

| | | | |
|---|---|---|---|
| $\mathscr{B}$ | "true" (high order) model | $w$ | observed response |
| | | $H$ | observed impulse resp. |
| $\widehat{\mathscr{B}}$ | approximate (low order) model | $\widehat{w}$ | response of $\widehat{\mathscr{B}}$ |
| | | $\widehat{H}$ | impulse resp. of $\widehat{\mathscr{B}}$ |




# Unstructured low-rank approximation

$$\widehat{D}^* := \arg\min_{\widehat{D}} \|D - \widehat{D}\|_{\mathrm{F}} \quad \text{subject to} \quad \operatorname{rank}(\widehat{D}) \le r$$

Theorem (closed form solution)

Let $D = U\Sigma V^\top$ be the SVD of $D$ and define

$$U =: \begin{bmatrix} \overset{r}{U_1} & \overset{n-r}{U_2} \end{bmatrix} \; m, \quad \Sigma =: \begin{bmatrix} \overset{r}{\Sigma_1} & \overset{n-r}{0} \\ 0 & \Sigma_2 \end{bmatrix} \begin{matrix} r \\ n-r \end{matrix} \quad \text{and} \quad V =: \begin{bmatrix} \overset{r}{V_1} & \overset{n-r}{V_2} \end{bmatrix} \; m .$$

An optimal low-rank approximation solution is

$$\widehat{D}^* = U_1 \Sigma_1 V_1^\top, \qquad (\widehat{\mathscr{B}}^* = \ker(U_2^\top) = \operatorname{colspan}(U_1)).$$

It is unique if and only if $\sigma_r \ne \sigma_{r+1}$.

# Structured low-rank approximation

No closed form solution is known for the general SLRA problem

$$\widehat{p}^* := \arg\min_{\widehat{p}} \|p - \widehat{p}\| \quad \text{subject to} \quad \operatorname{rank}\big(\mathscr{S}(\widehat{p})\big) \le r.$$

NP-hard, consider solution methods based on local optimization

Representing the constraint in a kernel form, the problem is

$$\min_{R,\; RR^\top = I_{m-r}} \Big( \min_{\widehat{p}} \|p - \widehat{p}\| \quad \text{subject to} \quad R\mathscr{S}(\widehat{p}) = 0 \Big)$$

Note: Double minimization with bilinear equality constraint.

There is a matrix $G(R)$, such that $R\mathscr{S}(\widehat{p}) = 0 \iff G(R)\widehat{p} = 0$.

## Variable projection vs. alternating projections

Two ways to approach the double minimization:

- Variable projections (VARPRO):
  solve the inner minimization analytically

$$\min_{R,\ RR^\top = I_{m-r}} \operatorname{vec}^\top\left(R\mathscr{S}(\widehat{p})\right)\left(G(R)G^\top(R)\right)^{-1}\operatorname{vec}\left(R\mathscr{S}(\widehat{p})\right)$$

  $\rightsquigarrow$ a nonlinear least squares problem for $R$ only.

- Alternating projections (AP):
  alternate between solving two least squares problems

VARPRO is globally convergent with a super linear conv. rate.

AP is globally convergent with a linear convergence rate.

## Variations on low-rank approximation

- Cost functions
  - weighted norms $(\operatorname{vec}^\top(D)W\operatorname{vec}(D))$
  - information criteria $(\log\det(D))$
- Constraints and structures
  - nonnegative
  - structured, sparse
  - missing data and exact data
- Data structures
  - nD data $\leftrightarrow$ tensors
  - nonlinear models $\leftrightarrow$ kernel methods
- Optimization algorithms
  - convex relaxations

## Summary

- SLRA is a generic problem for data modeling.

  has many applications in machine learning as well

- In general, SLRA is an NP-complete problem.

  search for special cases that have “nice” solutions
  *e.g.*, circulant SLRA can be computed by DFT.

- The SLRA framework leads to conceptual unification.

## Summary

- Efficient local solution methods

- Different rank representations (kernel, image, $AX = B$)
  lead to equivalent parameter optimization problems.

  Computationally, however, these problems are different.

  For example, the kernel representation leads to
  optimization on a Grassman manifold.

  Currently, it is unexplored which parameterization is
  computational most beneficial.

## Summary

- Effective heuristics, based on convex relaxations
- Practical advantage: one algorithm (and a piece of software) can solve a variety of problems
- Extensions of SLRA for tensors and nonlinear models

A framework with a potential for much to be done.

## Thank you

## Weighted low-rank approximation

In the EIV model, LRA is ML assuming $\operatorname{cov}(\operatorname{vec}(\widetilde{D})) = I$.

Motivation: incorporate prior knowledge $W$ about $\operatorname{cov}(\operatorname{vec}(\widetilde{D}))$

$$\min_{\widehat{D}} \operatorname{vec}^\top(D - \widehat{D}) W^{-1} \operatorname{vec}(D - \widehat{D}) \quad \text{subject to} \quad \operatorname{rank}(\widehat{D}) \leq r$$

Known in chemometrics as maximum likelihood PCA.

NP-hard problem, alternating projections is effective heuristic

## Nonnegative low-rank approximation

Constrained LRA arise in Markov chains and image mining

$$\min_{\widehat{D}} \|D - \widehat{D}\| \quad \text{subject to} \quad \operatorname{rank}(\widehat{D}) \leq r \text{ and } \widehat{D}_{ij} \geq 0 \text{ for all } i, j.$$

Using an image representation, an equivalent problem is

$$\min_{P \in \mathbb{R}^{m \times r},\ L \in \mathbb{R}^{r \times n}} \|D - PL\| \quad \text{subject to} \quad P_{ik}, L_{kj} \geq 0 \text{ for all } i, k, j.$$

Alternating projections algorithm:

- Choose an initial approximation $P^{(0)}$ and set $k := 0$.
- Solve: $L^{(k)} = \arg\min_L \|D - P^{(k)} L\|$ subject to $L \geq 0$.
- Solve: $P^{(k+1)} = \arg\min_P \|D - P L^{(k)}\|$ subject to $P \geq 0$.
- Repeat until convergence.

## Data fitting by a second order model

$$\mathscr{B}(A,b,c) := \{ d \in \mathbb{R}^{\mathrm{d}} \mid d^\top A d + b^\top d + c = 0 \}, \quad \text{with } A = A^\top$$

Consider first exact data:

$$d \in \mathscr{B}(A,b,c) \iff d^\top A d + b^\top d + c = 0$$

$$\iff \langle \underbrace{\operatorname{col}(d \otimes_s d, d, 1)}_{d_{\text{ext}}}, \underbrace{\operatorname{col}\big(\operatorname{vec}_s(A), b, c\big)}_{\theta} \rangle = 0$$

$$\{d_1, \dots, d_N\} \in \mathscr{B}(\theta) \iff \theta \in \text{left ker} \underbrace{\begin{bmatrix} d_{\text{ext},1} & \cdots & d_{\text{ext},N} \end{bmatrix}}_{D_{\text{ext}}}, \quad \theta \neq 0$$

$$\iff \operatorname{rank}(D_{\text{ext}}) \leq \mathrm{d} - 1$$

Therefore, for measured data $\rightsquigarrow$ LRA of $D_{\text{ext}}$.

Notes:

- Special case $\mathscr{B}$ an ellipsoid (for $A > 0$ and $4c < b^\top A^{-1} b$).
- Related to kernel PCA

## Rank minimization

Approximate modeling is a trade-off between:

- fitting accuracy and
- model complexity

Two possible scalarizations of the bi-objective optimization are:

LRA: minimize misfit under a constraint on complexity

RM: minimize complexity under a constraint ($\mathscr{C}$) on misfit

$$\min_X \operatorname{rank}(X) \quad \text{subject to} \quad X \in \mathscr{C}$$

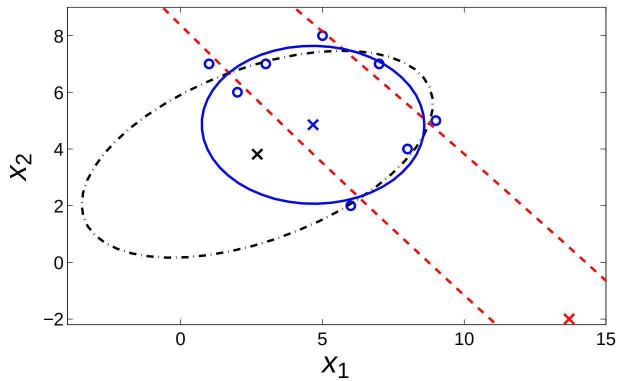
RM is also NP-hard, however, there are effective heuristics, *e.g.*,

with $X = \operatorname{diag}(x)$, $\operatorname{rank}(X) = \operatorname{card}(x)$,

$$\ell_1 \text{ heuristic:} \quad \min_x \|x\|_1 \quad \text{subject to} \quad \operatorname{diag}(x) \in \mathscr{C}$$

## Example: ellipsoid fitting

benchmark example of (Gander *et al.* 94), called "special data"




dashed — LRA    solid — proposed method

dashed-dotted — orthogonal regression (geometric fitting)

$\circ$ — data points    $\times$ — centers