Structured Low-Rank Approximation and (Some of) Its Applications

Ivan Markovsky

School of Electronics and Computer Science
University of Southampton

Examples

Let's start with the following well known examples:

- System realisation
- Discrete deconvolution
- Greatest common divisor of two polynomials

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

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

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
>
> $$\min_{\widehat{h}} \|h - \widehat{h}\| \quad \text{subject to} \quad \operatorname{rank}\big(\mathscr{H}(\widehat{h})\big) \leq \mathtt{n}$$

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 $\rightsquigarrow$ 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 $\Longrightarrow$ exact FIR fitting model
- approx. deconvolution $\Longrightarrow$ approx. FIR fitting model

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

Approximate deconvolution $\rightsquigarrow$ SLRA

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

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

Then the approximate deconvolution problem can be written as

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

$$\min_{\widehat{u},\,\widehat{y}} \|\mathrm{col}(u,y) - \mathrm{col}(\widehat{u},\widehat{y})\| \quad \text{subject to} \quad \mathrm{rank}\left(\begin{bmatrix} \mathscr{T}_{\mathrm{n}+1}(\widehat{u}) \\ \mathrm{row}(\widehat{y}) \end{bmatrix}\right) \leq (\mathrm{n}+1)\mathtt{m}$$

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

Greatest common divisor (GCD)

Consider the polynomials

$$a(z) := a_0 + a_1 z + \cdots + a_{n_a} z^{n_a}, \quad b(z) := b_0 + b_1 z + \cdots + b_{n_b} z^{n_b}$$

and define the Sylvester matrix

$$S(a,b) := \begin{bmatrix} a_0 & & & b_0 & & \\ \vdots & \ddots & & \vdots & \ddots & \\ a_{n_a} & & a_0 & b_{n_b} & & b_0 \\ & \ddots & \vdots & & \ddots & \vdots \\ & & a_{n_a} & & & b_{n_b} \end{bmatrix} \in \mathbb{R}^{(n_a+n_b)\times(n_a+n_b)}$$

The GCD of $a(z)$ and $b(z)$, has degree n, if and only if

$$\mathrm{rank}\left(S(a,b)\right) = n_a + n_b - n.$$

Approximate GCD $\iff$ Sylvester SLRA

Given $a(z)$, $b(z)$, and $n \in \mathbb{N}$, find

$$\min_{\widehat{a},\,\widehat{b}} \|\mathrm{col}(a,b) - \mathrm{col}(\widehat{a},\widehat{b})\| \quad \text{subject to} \quad \mathrm{rank}\left(S(a,b)\right) \leq n_a + n_b - n$$

Data matrix being low-rank

an exact property holds on the data $\iff$ a matrix constructed from data is low-rank

- h is realisable by an LTI system of order n $\iff$ $\mathrm{rank}\left(\mathscr{H}(h)\right) \leq \mathrm{n}$
- (u,y) is fitted by an n taps FIR system $\iff$ $\mathrm{rank}\left(\begin{bmatrix} \mathscr{T}_{\mathrm{n}+1}(u) \\ \mathrm{row}(y) \end{bmatrix}\right) \leq (\mathrm{n}+1)\mathtt{m}$
- $a(z), b(z)$ have GCD of deg. $\geq n$ $\iff$ $\mathrm{rank}\left(S(a,b)\right) \leq n_a + n_b - n$

Rank of the data matrix

complexity of an exact model fitting the data	$\leftrightarrow$	rank of the data matrix

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

Main issue: Low-rank approximation

With a bounding on the model complexity,

generically in the data space, exact property does not hold

$\Longrightarrow$ 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

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

Interpretation:

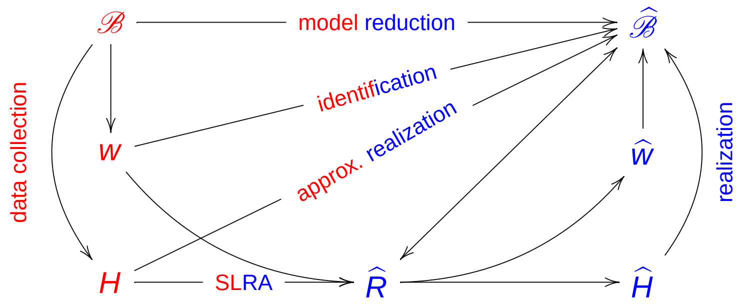
$\widehat{D}^* := \mathscr{S}(\widehat{p}^*)$ is optimal rank-r (or less) approx. of $D := \mathscr{S}(p)$,
within the class of matrices with the same structure as D.

Applications

- System theory
  1. Approximate realization
  2. Model reduction
  3. Errors-in-variables system identification
  4. Output error system identification
- Signal processing
  5. Output only (autonomous) system identification
  6. Finite impulse response (FIR) system identification
  7. Harmonic retrieval
  8. Image deblurring
- Computer algebra
  9. 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 $\widehat{\mathscr{B}}$



Errors-in-variables (EIV) identification

$\mathscr{L}_{\mathtt{m},\mathtt{l}}$ LTI model class of bounded complexity (#inputs$\leq \mathtt{m}$, lag$\leq \mathtt{l}$)

> Given $w_{\mathrm{d}} \in (\mathbb{R}^{\mathtt{w}})^T$ and complexity specification $(\mathtt{m}, \mathtt{l})$, find
>
> $$\widehat{\mathscr{B}}^* := \arg\min_{\widehat{\mathscr{B}}, \widehat{w}} \|w_{\mathrm{d}} - \widehat{w}\|_{\ell_2} \quad \text{subject to} \quad \widehat{w} \in \widehat{\mathscr{B}} \in \mathscr{L}_{\mathtt{m},\mathtt{l}}.$$

SLRA $(*)$ with $\mathscr{S}(p) = \mathscr{H}_{\mathtt{l}+1}(w_{\mathrm{d}})$, and $r = \mathtt{p}$.

EIV model: $w_{\mathrm{d}} = \overline{w} + \widetilde{w}, \quad \overline{w} \in \overline{\mathscr{B}} \in \mathscr{L}_{\mathtt{m},\mathtt{l}}, \quad \widetilde{w} \sim \mathrm{Normal}(0, \sigma^2 I)$

$\overline{w}$ — true data, $\overline{\mathscr{B}}$ — true model, $\widetilde{w}$ — measurement noise

$\widehat{\mathscr{B}}^*$ is a maximum likelihood estimate of $\overline{\mathscr{B}}$

consistent and assympt. normal $\implies$ confidence regions

Statistical vs. deterministic formulation

The EIV model gives a quality certificate to the method.

> The method works "well" (consistency) and is optimal (efficiency) under certain specified conditions.

However, the assumption that the data is generated by a true model with additive noise is sometimes not realistic.

Model-data mismatch is often due to a restrictive (LTI) model class being used and not (only) due to measurement noise.

$\implies$ The approximation aspect is often more important than the stochastic estimation one.

Outline

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

$$U =: \begin{array}{cc} r & n-r \\ \left[U_1\right. & \left.U_2\right] \end{array} \; m, \quad \Sigma =: \begin{array}{cc} r & n-r \\ \left[\begin{array}{cc} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{array}\right] \end{array} \begin{array}{c} r \\ n-r \end{array} \quad \text{and} \quad V =: \begin{array}{cc} r & n-r \\ \left[V_1\right. & \left.V_2\right] \end{array} \; m \; .$$

An optimal low-rank approximation solution is

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

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

Structured low-rank approximation

No closed form solution is known for the general SLRA problem

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

NP-hard, consider solution methods based on local optimization

Representing the constraint in a kernel form, the problem is

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

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.

Software implementation

The structure of $\mathscr{S}$ can be exploited for efficient $O(\dim(p))$ cost function and first derivative evaluations.

SLICOT library includes high quality FORTRAN implementation of algorithms for block Toeplitz matrices.

SLRA C software using I/O repr. and VARPRO approach `http://www.esat.kuleuven.be/~imarkovs`

Based on the Levenberg–Marquardt alg. implemented in MINPACK.

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
  - sparse
- Data structures
  - nonlinear models
  - tensors
- Optimization algorithms
  - convex relaxations

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: $P^{(k+1)} = \arg\min_P \|D - PL^{(k)}\|$ 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:

$$\begin{aligned} d \in \mathscr{B}(A,b,c) &\iff d^\top A d + b^\top d + c = 0 \\ &\iff \big\langle \underbrace{\operatorname{col}(d \otimes_s d, d, 1)}_{d_{\text{ext}}}, \underbrace{\operatorname{col}\big(\operatorname{vec}_s(A), b, c\big)}_{\theta} \big\rangle = 0 \end{aligned}$$

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

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

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

Notes:

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

Consistency in the errors-in-variables setting

Assume that the data is collected according to the EIV model

$$d_i = \bar{d}_i + \tilde{d}_i, \quad \text{where} \quad \bar{d}_i \in \mathscr{B}(\bar{\theta}), \quad \tilde{d}_i \sim \mathrm{N}(0, \sigma^2 I).$$

LRA of D_{ext} (kernel PCA) $\rightsquigarrow$ inconsistent estimator

$$\tilde{d}_{\mathrm{ext},i} := \mathrm{col}(\tilde{d}_i \otimes_{\mathrm{s}} \tilde{d}_i, \tilde{d}_i, 0) \text{ is not Gaussian}$$

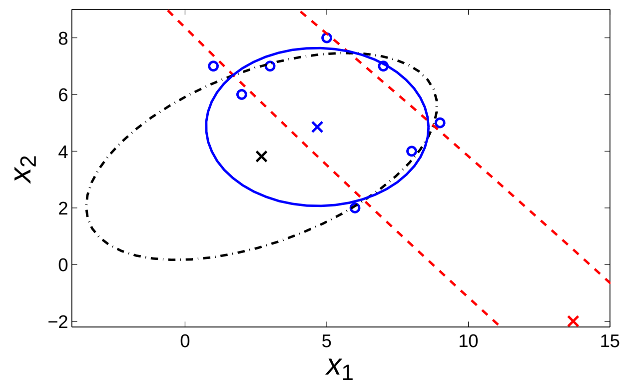
proposed method — incorporate bias correction in the LRA

Notes:

- works on the sample covariance matrix $D_{\mathrm{ext}} D_{\mathrm{ext}}^{\top}$
- the correction depends on the noise variance σ^2
- the core of the proposed method is the σ^2 estimator (possible link with methods for choosing regularization par.)

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

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

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

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

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

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

Structured pseudospectra

$\Lambda(A)$ — the set of eigenvalues of $A \in \mathbb{C}^{n\times n}$

$\mathbb{M}$ — a set of matrices ($\mathbb{M} = \{\mathscr{S}(p) \mid p \in \mathbb{R}^{n_p}\}$)

Using the structured pseudospectra

$$\Lambda_\varepsilon(A) := \{ z \in \mathbb{C} \mid z \in \Lambda(B),\ B \in \mathbb{M},\ \|A - B\|_2 \leq \varepsilon \}$$

one can determine the distance to singularity

$$d(A) := \min_{\Delta A \in \mathbb{M}} \|\Delta A\|_2 \quad \text{subject to} \quad A + \Delta A \text{ is singular}$$

which is a special SLRA problem with

1. square data matrix
2. perturbation measured by spectral norm, and
3. focus on minimum (vs minimizer) and singularity (vs rank).

Summary

- SLRA is a generic problem for data modeling.

  search for more applications (pole placement, μ-analysis, ...)

- 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