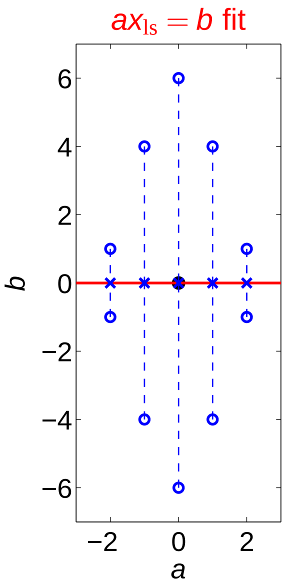
Structured Low-Rank Approximation and (Some of) Its Applications

Ivan Markovsky

K.U.Leuven, ESAT-SISTA

A line fitting example



Classic problem: Fit the points

$$d_1 = \begin{bmatrix}0\\6\end{bmatrix},\ d_2 = \begin{bmatrix}1\\4\end{bmatrix},\ \dots,\ d_{10} = \begin{bmatrix}-1\\4\end{bmatrix}$$

by a line passing through the origin.

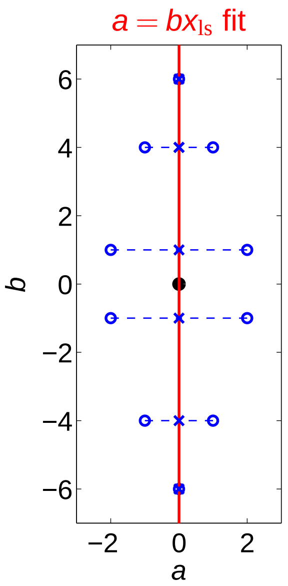
Classic solution: Define $d_i =: \operatorname{col}(a_i, b_i)$ and solve the least squares problem

$$\operatorname{col}(a_1,\dots,a_{10})x = \operatorname{col}(b_1,\dots,b_{10}).$$

The LS fitting line is given by $ax_{\text{ls}} = b$.

It minimizes the vertical distances from the data points to the fitting line.

A line fitting example (cont.)



Minimizing vertical distances does not seem appropriate in this example.

Revised LS problem:

$$\operatorname{col}(a_1,\dots,a_{10}) = \operatorname{col}(b_1,\dots,b_{10})x$$

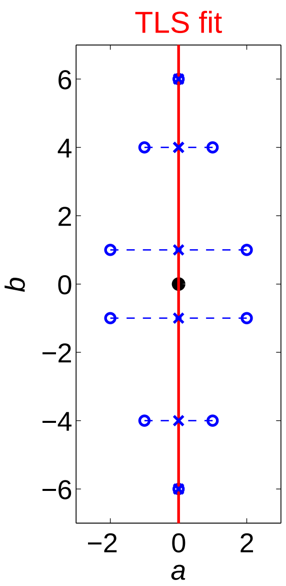
minimize the horizontal distances

The fitting line is now given by $a = bx_{\text{ls}}$.

Total least squares fitting:

minimize the orthogonal distances

A line fitting example (cont.)



Total least squares problem:

$$\min_{x,\widehat{a}_i,\widehat{b}_i} \ \sum_{i=1}^{10} \left((a_i - \widehat{a}_i)^2 + (b_i - \widehat{b}_i)^2 \right)$$

$$\text{subject to} \quad \widehat{a}_i x = \widehat{b}_i, \quad i = 1,\dots,10$$

However, x_{tls} does not exist! ($x_{\text{tls}} = \infty$)

If we represent the fitting line as an

image $d = Pl$ or kernel $Rd = 0$

TLS solutions do exist, *e.g.*,

$$P_{\text{tls}} = \operatorname{col}(0,1) \quad \text{and} \quad R_{\text{tls}} = \begin{bmatrix}1 & 0\end{bmatrix}.$$

What are the issues?

- LS is representation dependent
- TLS is representation invariant
- TLS using I/O representation might have no solution

The representation is a matter of convenience and should not affect the solution.

$\Longrightarrow$ Orthogonal distance minimization combined with image or kernel representation is a better concept.

In this talk . . .

In fact, line fitting is a low-rank approximation (LRA) problem:

$$\boxed{\text{approximate } D := \begin{bmatrix} d_1 & \cdots & d_{10} \end{bmatrix} \text{ by a rank-one matrix,}}$$

. . . a representation free concept applying to general multivariable static and dynamic linear fitting problems.

LRA is closely related to:

- principle component analysis PCA
- latent semantic analysis LSA
- various other methods for dimensionality reduction

Outline

Low-rank approximation as data modeling

Applications

Algorithms

Related problems

Low-rank approximation

Given

- a matrix $D \in \mathbb{R}^{\mathrm{d} \times N}$, $\mathrm{d} \leq N$
- a matrix norm $\|\cdot\|$, and
- an integer $\mathtt{m}$, $0 < \mathtt{m} < \mathrm{d}$,

find

$$\widehat{D}^* := \arg\min_{\widehat{D}} \|D - \widehat{D}\| \quad \text{subject to} \quad \operatorname{rank}(\widehat{D}) \leq \mathtt{m}.$$

Interpretation:

$\widehat{D}^*$ is optimal rank-$\mathtt{m}$ (or less) approximation of D (w.r.t. $\|\cdot\|$).

Why low-rank approximation?

D is low-rank	$\iff$	D is generated by a linear model
so that LRA	$\iff$	data modeling

Suppose

$$\mathtt{m} := \operatorname{rank}(D) < \mathtt{d} := \operatorname{row\,dim}(D).$$

Then there is a full rank $R \in \mathbb{R}^{\mathtt{p} \times \mathtt{d}}$, $\mathtt{p} := \mathtt{d} - \mathtt{m}$, such that $RD = 0$.

The columns $d_1, \dots, d_N$ of D obey $\mathtt{p}$ independent linear relations $r_i d_j = 0$, given by the rows $r_1, \dots, r_{\mathtt{p}}$ of R.

$Rd = 0$ is a kernel representation of the model $\mathscr{B} := \{\, d \mid Rd = 0 \,\}$.

LRA as data modeling

Given

- N, $\mathtt{d}$-variable observations $\begin{bmatrix} d_1 & \cdots & d_N \end{bmatrix} := D \in \mathbb{R}^{\mathtt{d} \times N}$
- a matrix norm $\|\cdot\|$, and
- model complexity $\mathtt{m}$, $0 < \mathtt{m} < \mathtt{d}$,

find

$$\widehat{\mathscr{B}}^* := \arg\min_{\widehat{\mathscr{B}}, \widehat{D}} \|D - \widehat{D}\| \quad \text{subject to} \quad \begin{array}{c} \operatorname{col\,span}(\widehat{D}) \subseteq \widehat{\mathscr{B}} \\ \dim(\widehat{\mathscr{B}}) \leq \mathtt{m} \end{array}$$

Interpretation:

$\widehat{\mathscr{B}}^*$ is optimal (w.r.t. $\|\cdot\|$) approximate model for D with bounded complexity: $\dim(\widehat{\mathscr{B}}) \leq \mathtt{m} \iff$ # inputs $\leq \mathtt{m}$.

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

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

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.

Why structured low-rank approximation?

$D = S(p)$ is low-rank and (Hankel) structured	$\iff$	p is generated by a LTI dynamic model

Example: $D = \mathscr{H}_{\mathtt{l}+1}(w_{\mathrm{d}})$ block Hankel and rank deficient

$\exists R$, such that $R\mathscr{H}_{\mathtt{l}+1}(w_{\mathrm{d}}) = 0$. Taking into account the structure

$$\begin{bmatrix} R_0 & R_1 & \cdots & R_{\mathtt{l}} \end{bmatrix} \begin{bmatrix} w_{\mathrm{d}}(1) & w_{\mathrm{d}}(2) & \cdots & w_{\mathrm{d}}(T - \mathtt{l}) \\ w_{\mathrm{d}}(2) & w_{\mathrm{d}}(3) & \cdots & w_{\mathrm{d}}(T - \mathtt{l} + 1) \\ \vdots & \vdots & & \vdots \\ w_{\mathrm{d}}(\mathtt{l}+1) & w_{\mathrm{d}}(\mathtt{l}+2) & \cdots & w_{\mathrm{d}}(T) \end{bmatrix} = 0$$

we have a vector difference equation for w_{d} with $\mathtt{l}$ lags

$$R_0 w_{\mathrm{d}}(t) + R_1 w_{\mathrm{d}}(t+1) + \cdots + R_{\mathtt{l}} w_{\mathrm{d}}(t+\mathtt{l}) = 0 \quad \text{for } t = 1, \dots, T - \mathtt{l}.$$

SLRA as time-series modeling

Given

- T samples, $\mathtt{w}$ variables, vector time series $w_{\mathrm{d}} \in (\mathbb{R}^{\mathtt{w}})^T$,
- a signal norm $\|\cdot\|$, and
- model complexity $(\mathtt{m}, \mathtt{l})$, $0 \le \mathtt{m} < \mathtt{w}$,

find

$$\widehat{\mathscr{B}}^* := \arg\min_{\widehat{\mathscr{B}}, \widehat{w}} \|w_{\mathrm{d}} - \widehat{w}\| \quad \text{s.t.} \quad \begin{array}{c} \widehat{w} \in \widehat{\mathscr{B}}, \\ \dim(\widehat{\mathscr{B}}) \le T\mathtt{m} + \mathtt{l}(\mathtt{w} - \mathtt{m}) \end{array} \qquad (*)$$

Interpretation:

$\widehat{\mathscr{B}}^*$ is optimal (w.r.t. $\|\cdot\|$) model for the time series w_{d}
with a bounded complexity: # inputs $\le \mathtt{m}$ and lag $\le \mathtt{l}$.

(Go back to page 26.)

Kernel, image, and input/output representations

A static model $\mathscr{B}$ with $\mathtt{d}$ variables is a subset of $\mathbb{R}^{\mathtt{d}}$.

How to represent a linear model $\mathscr{B}$ (a subspace) by equations?

Representations:

- kernel: $\mathscr{B} = \ker(R)$, $R \in \mathbb{R}^{\mathtt{p} \times \mathtt{d}}$
- image: $\mathscr{B} = \operatorname{col\,span}(P)$, $P \in \mathbb{R}^{\mathtt{d} \times \mathtt{m}}$
- input/output: $\mathscr{B}_{\mathrm{i/o}} = \mathscr{B}(X)$, $X \in \mathbb{R}^{\mathtt{m} \times \mathtt{p}}$

$$\mathscr{B}_{\mathrm{i/o}}(X) := \{\, d := \operatorname{col}(d_{\mathrm{i}}, d_{\mathrm{o}}) \in \mathbb{R}^{\mathtt{d}} \mid d_{\mathrm{i}} \in \mathbb{R}^{\mathtt{m}},\ d_{\mathrm{o}} = X^\top d_{\mathrm{i}} \,\}$$

In terms of D, the I/O repr. is $AX \approx B$, where $\begin{bmatrix} A & B \end{bmatrix} := D^\top$.

$\Longrightarrow$ Solving $AX \approx B$ approximately by LS, TLS, ...
is LRA using I/O representation

Links among the parameters R, P, and X

Define the partitionings

$$R =: \begin{bmatrix} R_{\mathrm{i}} & R_{\mathrm{o}} \end{bmatrix}, \quad R_{\mathrm{o}} \in \mathbb{R}^{\mathtt{p} \times \mathtt{p}} \quad \text{and} \quad P =: \begin{bmatrix} P_{\mathrm{i}} \\ P_{\mathrm{o}} \end{bmatrix}, \quad P_{\mathrm{i}} \in \mathbb{R}^{\mathtt{m} \times \mathtt{m}}.$$

We have the following links among R, P, and X:

$$\begin{tikzcd}
\mathscr{B} = \ker(R) \arrow[rr, leftrightarrow, "RP=0"] \arrow[dr, shift left, "X^\top = -R_{\mathrm{o}}^{-1} R_{\mathrm{i}}"] & & \mathscr{B} = \operatorname{col\,span}(P) \arrow[dl, shift left, "X^\top = P_{\mathrm{o}} P_{\mathrm{i}}^{-1}"'] \\
& \mathscr{B} = \mathscr{B}_{\mathrm{i/o}}(X) \arrow[ul, shift left, "R = [X^\top \ -I]"] \arrow[ur, shift left, "P^\top = [I \ \ X]"'] &
\end{tikzcd}$$

Nonuniqueness of an input/output partition

In general, many I/O partitions of the variables w are possible.

> Choosing an I/O partition amounts to choosing a full rank $\mathtt{p} \times \mathtt{p}$ submatrix of R or a full rank $\mathtt{m} \times \mathtt{m}$ submatrix of P.

Often there is no a priori reason to prefer one partition over another.

$\Longrightarrow$ $AX \approx B$ is often not a natural starting point for data modeling

In addition, it might lead to ill-posed computational problems.

LTI models of bounded complexity

A dynamic model $\mathscr{B}$ with $\mathtt{w}$ variables is a subset of $(\mathbb{R}^{\mathtt{w}})^{\mathbb{Z}}$.

$\mathscr{B}$ is LTI $:\iff$ $\mathscr{B}$ is a shift-invariant subspace of $(\mathbb{R}^{\mathtt{w}})^{\mathbb{Z}}$.

Let $\mathscr{B}$ be LTI with $\mathtt{m}$ inputs, $\mathtt{p}$ outputs, of order $\mathtt{n}$ and lag $\mathtt{l}$,

$$\dim\left(\mathscr{B}|_{[0,T]}\right) = \mathtt{m}T + \mathtt{n} \leq \mathtt{m}T + \mathtt{p}\mathtt{l}, \quad \text{for } T \geq \mathtt{l}.$$

$\dim(\mathscr{B})$ is an indication of the model complexity.

$\implies$ The complexity of $\mathscr{B}$ is specified by $(\mathtt{m},\mathtt{n})$ or $(\mathtt{m},\mathtt{l})$.

Notation: $\mathscr{L}^{\mathtt{w}}_{\mathtt{m},\mathtt{l}}$ — LTI model class with bounded complexity
\# inputs $\leq \mathtt{m}$ and lag $\leq \mathtt{l}$.

SISTA

LTI model representations

- Kernel representation (parameter $R(z) := \sum_{i=0}^{\mathtt{l}} R_i z^i$)

$$R_0 w(t) + R_1 w(t+1) + \cdots + R_{\mathtt{l}} w(t+\mathtt{l}) = 0$$

- Impulse response represent (parameter $H : \mathbb{Z} \to \mathbb{R}^{p\times m}$)

$$w = \operatorname{col}(u,y), \qquad y(t) = \sum_{\tau=-\infty}^{t} H(\tau)u(t-\tau)$$

- Input/state/output representation (parameter (A,B,C,D))

$$w = \operatorname{col}(u,y), \qquad \begin{aligned} x(t+1) &= Ax(t) + Bu(t) \\ y(t) &= Cx(t) + Du(t) \end{aligned}$$

Transitions among R, H, (A,B,C,D) are classic problems, *e.g.*,

R or $H \mapsto (A,B,C,D)$ are realization problems.

SISTA

Basic uses of a model $\mathscr{B}$

- Simulation: given an input u and initial conditions find the corresponding output y, $\operatorname{col}(u,y) \in \mathscr{B}$
- Smoothing: given w_{d}, find its best approximation in $\mathscr{B}$

$$\hat{w}^* := \arg\min_{\hat{w}} \|w_{\mathrm{d}} - \hat{w}\| \text{ subject to } \hat{w} \in \mathscr{B}$$

- Filtering: smoothing in real-time
- Prediction: given past data $w_{\mathrm{p}} = \big(w_{\mathrm{d}}(1),\dots,w_{\mathrm{d}}(t)\big)$
find prediction $\hat{w}_{\mathrm{f}} = \big(\hat{w}(t+1),\hat{w}(t+2),\dots\big)$

$\mathscr{B}$ subspace $\implies$ these are linear problems (projections on $\mathscr{B}$)

There are efficient algorithms for carrying out the computations.

They are ingredients of the approximate modeling algorithms.

SISTA

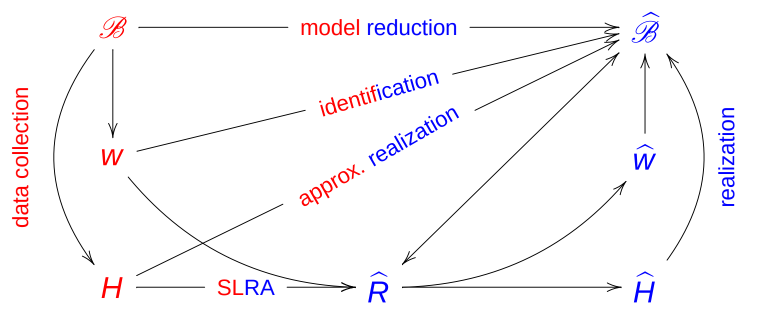
Applications

- 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)

SISTA

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 $\widehat{\mathscr{B}}$



Generic problem: structured LRA

The applications are special cases of the SLRA problem:

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

for specific choices of p, $\mathscr{S}$, and r.

$\Longrightarrow$ Algorithms and software for SLRA can be readily used.

Notes:

- In many applications, $\mathscr{S}(\cdot)$ is composed of blocks that are:

  (H) block Hankel, (U) Unstructured, or (F) Fixed.
- Of interest is the model $\widehat{\mathscr{B}}^*$, given, *e.g.*, by $\operatorname{left\,ker}\big(\mathscr{S}(\widehat{p}^*)\big)$.
- The algorithms compute $\widehat{R}$, such that $\widehat{R}\mathscr{S}(\widehat{p}^*) = 0$.

Errors-in-variables identification

Statistical name for the fitting problem $(*)$ considered before.

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

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

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

$\widehat{\mathscr{B}}^*$ is a maximum likelihood estimate of $\bar{\mathscr{B}}$, in the EIV model

consistent and assympt. normal $\Longrightarrow$ 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.

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

System theory $\leftrightarrow$ Signal proc. $\leftrightarrow$ Computer algebra

The Toeplitz matrix–vector product $y = \mathscr{T}(H)u = \mathscr{T}(u)H$ is equivalent to (may describe):

$(u,y) \in \mathscr{B}(H)$	$\Longleftrightarrow$	$y = H \star u$	$\Longleftrightarrow$	$y(z) = H(z)u(z)$
FIR sys. traj.		convolution		polyn. multipl.

Multivariable case: block Toeplitz structure

multivariable systems	$\Longleftrightarrow$	matrix valued time series	$\Longleftrightarrow$	matrix valued polynomials

2D case: block Toeplitz–Toeplitz block structure

multidim. system	$\Longleftrightarrow$	function of several indep. variables	$\Longleftrightarrow$	polyn. of several var.

(F) Forward problem define $y := \mathscr{T}(u)H$

(I) Inverse problem solve $y = \mathscr{T}(u)H$ for H

	System theory	Signal proc.	Computer algebra
F	FIR sys. simulation	convolution	polyn. multipl.
I	FIR sys. identification	deconv.	polyn. division

Typically $y = \mathscr{T}(u)H$ is an overdetermined system of eqns

$\Longrightarrow$ With "rough data $w_{\mathrm{d}} = (u_{\mathrm{d}}, y_{\mathrm{d}})$", there is no exact solution.

$\rightsquigarrow$ approximate identification, deconvolution, polyn. division.

SLRA: find the smallest modification of the data w_{d} that allows the modified data $\widehat{w}$ to have an exact solution.

Outline

Low-rank approximation as data modeling

Applications

Algorithms

Related problems

Unstructured low-rank approximation

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

Theorem (closed form solution)

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

$$U =: \begin{array}{c} \begin{array}{cc} \mathtt{m} & \mathtt{p} \end{array} \\ \begin{bmatrix} U_1 & U_2 \end{bmatrix} \end{array} \mathtt{d}\,, \quad \Sigma =: \begin{array}{cc} \begin{array}{cc} \mathtt{m} & \mathtt{p} \end{array} & \\ \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix} & \begin{array}{c} \mathtt{m} \\ \mathtt{p} \end{array} \end{array} \quad \text{and} \quad V =: \begin{array}{c} \begin{array}{cc} \mathtt{m} & \mathtt{p} \end{array} \\ \begin{bmatrix} V_1 & V_2 \end{bmatrix} \end{array} N\,.$$

An optimal LRA 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 $\sigma_{\mathtt{m}} \neq \sigma_{\mathtt{m}+1}$.

Structured low-rank approximation

No closed form solution is known for the general SLRA problem

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

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 $R\mathscr{S}(\widehat{p}) = 0 \iff G(R)p = 0$.

Low-rank approximation as data modeling Applications Algorithms Related problems

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.

Low-rank approximation as data modeling Applications Algorithms Related problems

Software implementation

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

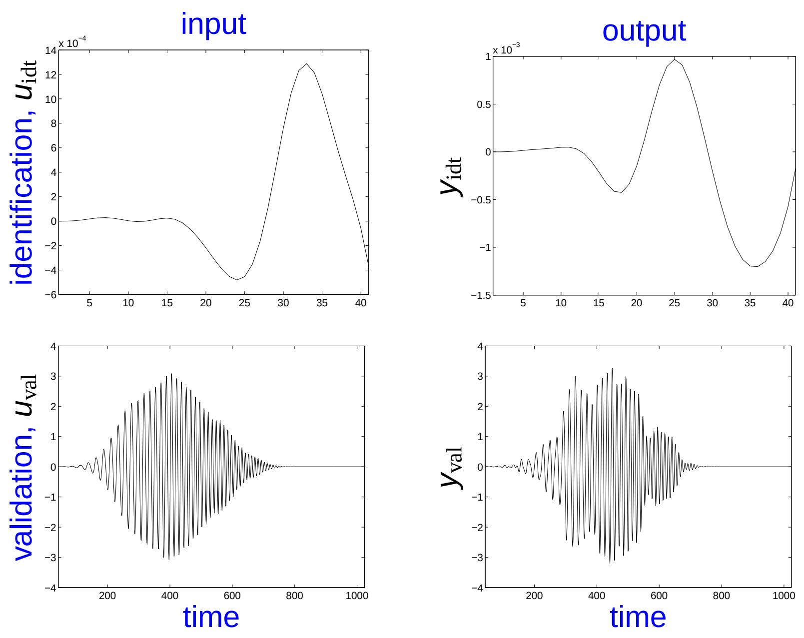
SLICOT library includes high quality FORTRAN implementation of algorithms for block Toeplitz matrices.

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.

Low-rank approximation as data modeling Applications Algorithms Related problems

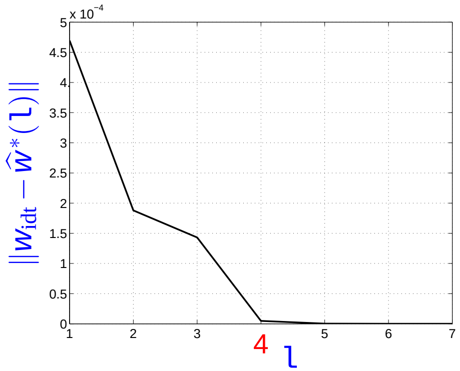
Example: data of a wing flutter



Example: choice of the model class

Consider the LTI model class $\mathscr{L}_{\mathtt{m},\mathtt{l}}$, where $\mathtt{m} = 1$ is given and $\mathtt{l}$ is an unknown hyper parameter.

We choose $\mathtt{l}$ from the misfit vs complexity tradeoff curve:



Example: optimal model

The optimal model $\hat{\mathscr{B}}^*$ in the model class $\mathscr{L}_{1,4}$ is

$$\widehat{\mathscr{B}}^* = \ker\big(\widehat{R}^*(\sigma)\big), \quad \text{where}$$

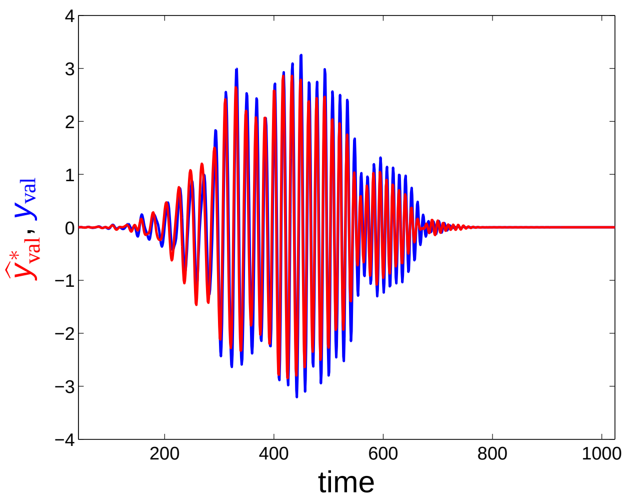
$$\begin{aligned}\widehat{R}^*(z) = & \begin{bmatrix} 9.55 & -0.09 \end{bmatrix} z^0 + \begin{bmatrix} -32.18 & 1.50 \end{bmatrix} z^1 + \\ & \begin{bmatrix} 43.77 & -3.56 \end{bmatrix} z^2 + \begin{bmatrix} -28.62 & 3.05 \end{bmatrix} z^3 + \\ & \begin{bmatrix} 7.57 & -1.00 \end{bmatrix} z^4 .\end{aligned}$$

Notes:

- Relatively simple model for the flutter phenomenon.
- Computed in 0.4 sec on a desktop computer by the SLRA software.

Example: validation

$\widehat{\mathscr{B}}^*$ fits well the output $y_{\rm val}$ on the validation data $w_{\rm val}$



Variations on low-rank approximation

- Cost functions
  - weighted norms $\quad (\text{vec}^\top(D) W \text{vec}(D))$
  - information criteria $\quad (\log\det(D))$
- 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 $\text{cov}(\text{vec}(\widetilde{D})) = I$.

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

$$\min_{\widehat{D}} \text{vec}^\top(D - \widehat{D}) W \,\text{vec}(D - \widehat{D}) \quad \text{subject to} \quad \text{rank}(\widehat{D}) \leq \mathtt{m}$$

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

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

Using an image representation, an equivalent problem is

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

Alternating projections algorithm:

- Choose an initial approximation $P^{(0)} \in \mathbb{R}^{\mathtt{d} \times \mathtt{m}}$ 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

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

Consider first exact data:

$$d \in \mathscr{B}(A,b,c) \iff d^\top A d + b^\top d + c = 0$$
$$\iff \langle \underbrace{\text{col}(d \otimes_{\text{s}} d, d, 1)}_{d_{\text{ext}}}, \underbrace{\text{col}\left(\text{vec}_{\text{s}}(A), b, d\right)}_{\theta} \rangle = 0$$

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

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

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

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

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

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

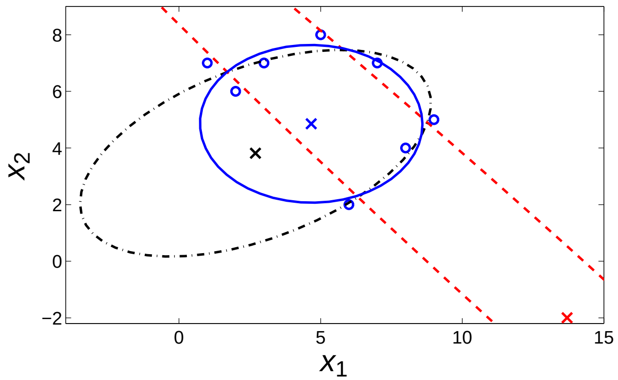
proposed method — incorporate bias correction in the LRA

Notes:

- works on the sample covariance matrix $D_{\text{ext}} D_{\text{ext}}^\top$
- 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 tradeoff 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 $X = \operatorname{diag}(x)$, $\operatorname{rank}(X) = \operatorname{card}(x)$,

ℓ_1 heuristic: $\quad \min_x \|x\|_1 \quad \text{subject to} \quad \operatorname{diag}(x) \in \mathscr{C}$

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

- LRA $\iff$ linear data modeling (in the behavioral setting)
- rank and behavior $\rightsquigarrow$ representation-free problems
- however, different repr. are convenient for different goals
- $AX \approx B$ is LRA with fixed I/O repr. $\rightsquigarrow$ lack of solution
- applications in system theory, signal processing, and computer algebra
- links with rank minimization, structured pseudospectra, and positive rank