Structured low-rank approximation as optimization on a Grassmann manifold

Konstantin Usevich and Ivan Markovsky
Department ELEC, Vrije Universiteit Brussel

Leuven, 10 July 2013

International Workshop on Advances in Regularization, Optimization, Kernel Methods and Support Vector Machines: theory and applications

Structured low-rank approximation problem

$$\begin{bmatrix} a & b & c & d \\ b & c & d & e \\ c & d & e & f \end{bmatrix} \approx \begin{bmatrix} \widehat{a} & \widehat{b} & \widehat{c} & \widehat{d} \\ \widehat{b} & \widehat{c} & \widehat{d} & \widehat{e} \\ \widehat{c} & \widehat{d} & \widehat{e} & \widehat{f} \end{bmatrix} = \begin{bmatrix} * \\ * \\ * \end{bmatrix} \underbrace{\begin{bmatrix} * & * & * & * \end{bmatrix}}_{}\Big\} r$$

$$(a, b, c, d, e, f) \approx (\widehat{a}, \widehat{b}, \widehat{c}, \widehat{d}, \widehat{e}, \widehat{f}) \in \mathfrak{M}_r$$

Data — model with complexity r

Structured low-rank approximation problem

$$\begin{bmatrix} a & b & c & d \\ b & c & d & e \\ c & d & e & f \end{bmatrix} \approx \begin{bmatrix} \widehat{a} & \widehat{b} & \widehat{c} & \widehat{d} \\ \widehat{b} & \widehat{c} & \widehat{d} & \widehat{e} \\ \widehat{c} & \widehat{d} & \widehat{e} & \widehat{f} \end{bmatrix} = \begin{bmatrix} * \\ * \\ * \end{bmatrix} \begin{bmatrix} * & * & * & * \end{bmatrix} \} \, r$$

$$\underset{\text{Data}}{(a,b,c,d,e,f)} \approx \underset{\text{model with complexity r}}{(\widehat{a},\widehat{b},\widehat{c},\widehat{d},\widehat{e},\widehat{f}) \in \mathfrak{M}_r}$$

Structure of this talk:

$$\overset{1}{\text{Structured low-rank approximation}} \overset{2}{\longleftrightarrow} \overset{3}{\text{Optimization on } \mathrm{Gr}(d,m)}$$

Hankel matrices

$$\mathscr{H}_m(p) := \begin{bmatrix} p(1) & p(2) & \cdots & p(T-m+1) \\ p(2) & \iddots & & p(T-m+2) \\ \vdots & \iddots & \iddots & \vdots \\ p(m) & p(m+1) & \cdots & p(T) \end{bmatrix},$$

$p = \begin{bmatrix} p(1) & p(2) & \cdots p(T) \end{bmatrix}$
univariate time series

Theorem. (Heinig, 1984) (evident if $r = m - 1$)

$$\operatorname{rank} \mathscr{H}_m(p) \leq r < m \iff \theta_0 p(t) + \theta_1 p(t+1) + \cdots + \theta_r p(t+r) = 0,$$

linear recurrence $t = 1 : T - r$

Hankel matrices

$$\mathscr{H}_m(p) := \begin{bmatrix} p(1) & p(2) & \cdots & p(T-m+1) \\ p(2) & \iddots & & p(T-m+2) \\ \vdots & \iddots & \iddots & \vdots \\ p(m) & p(m+1) & \cdots & p(T) \end{bmatrix}, \qquad p = \begin{bmatrix} p(1) & p(2) & \cdots p(T) \end{bmatrix}$$

univariate time series

Theorem. (Heinig, 1984) (evident if $r = m - 1$)

rank $\mathscr{H}_m(p) \leq r < m \iff \theta_0 p(t) + \theta_1 p(t+1) + \cdots + \theta_r p(t+r) = 0,$

$$t = 1 : T - r$$

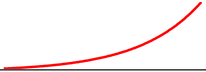
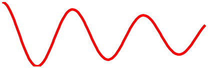
$$\longleftrightarrow p(t) = \sum_{k=1}^{d} \underbrace{P_k(t) \cdot \lambda_k^t}_{\text{polynomial} \cdot \text{exponential}}$$

where

- $\lambda_1, \ldots, \lambda_d$ — distinct roots of $\theta(z) = \sum_{j=0}^{r} \theta_j z^j$
- $\deg P_k(t) =$ (multiplicity of λ_k) - 1 ($P_k = const$ if λ_k is simple)

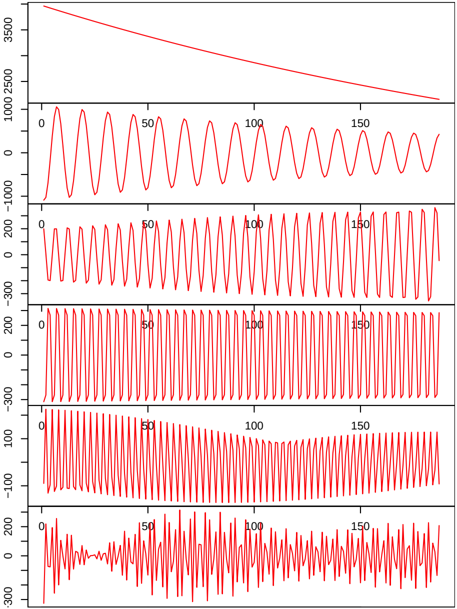
Low-rank Hankel matrices: examples

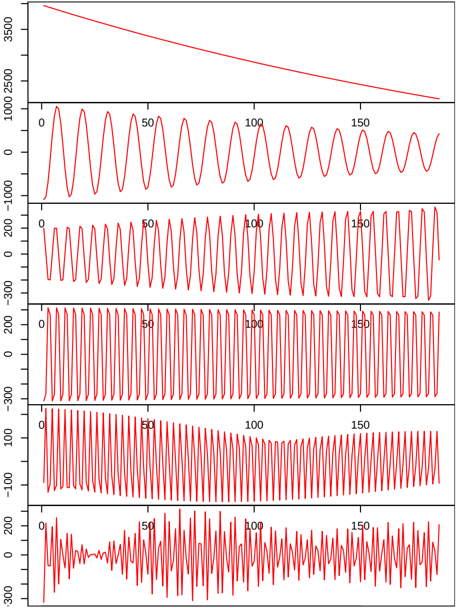
$$p(t) = \sum_{k=1}^{d} P_k(t)\lambda_k^t \iff \theta_0 p(t) + \cdots + \theta_r p(t+r) = 0$$

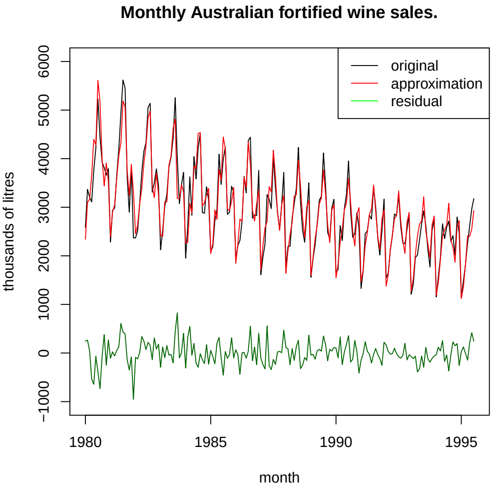
Real p:	picture	formula	r	$\sum_{j=0}^{r} \theta_j z^j$
exponential		$c\rho^t$	1	$(z-\rho)$
damped sine		$c\rho^t \cos(\omega t + \phi)$	2	$(z-\rho e^{i\omega})\cdot (z-\rho e^{-i\omega})$
polynomial		$\sum_{k=0}^{3} c_k t^k$	4	$(z-1)^4$

Low-rank approximation $\mathscr{H}_m(p) \approx$ l.r. $\mathscr{H}_m(\widehat{p})$ $\leftrightarrow$ Sparse approximation with ∞ dictionary

$$p \approx \sum_{k=1}^{d} c_k f_k, \quad f_k(t) \in \left\{ t^\alpha \lambda^t \right\}_{\substack{\alpha \in \mathbb{N}, \\ \lambda \in \mathbb{C}}}$$


3500
2500
1000
0
−1000
200
0
−300
200
0
−300
100
−100
200
0
-300
0
50
100
150




Monthly Australian fortified wine sales.
original
approximation
residual
thousands of litres
month
1980
1985
1990
1995

Block-Hankel matrices

$$\mathscr{H}_{\ell+1}(w) := \begin{bmatrix} w(1) & w(2) & \cdots & w(T-\ell) \\ w(2) & \iddots & & w(T-\ell+1) \\ \vdots & \iddots & \iddots & \vdots \\ w(\ell+1) & w(\ell+2) & \cdots & w(T) \end{bmatrix}, \quad w = \begin{bmatrix} w(1) & w(2) & \cdots w(T) \end{bmatrix} \in \mathbb{R}^{q \times T}$$

q-variate time series

$$\operatorname{rank} \mathscr{H}_{\ell+1}(w) \leq \underbrace{(\ell+1)q}_{\text{number of rows}} - p \iff \begin{bmatrix} R_0 & \cdots & R_\ell \end{bmatrix} \mathscr{H}_{\ell+1}(w) = 0, \ R_k \in \mathbb{R}^{p \times q}$$

$$\Updownarrow$$

$$R_0 w(t) + \cdots + R_\ell w(t+\ell) = 0, \quad t=1:T-\ell$$

Block-Hankel matrices

$$\mathscr{H}_{\ell+1}(w) := \begin{bmatrix} w(1) & w(2) & \cdots & w(T-\ell) \\ w(2) & \iddots & & w(T-\ell+1) \\ \vdots & \iddots & \iddots & \vdots \\ w(\ell+1) & w(\ell+2) & \cdots & w(T) \end{bmatrix}, \quad w = \begin{bmatrix} w(1) & w(2) & \cdots w(T) \end{bmatrix} \in \mathbb{R}^{q \times T}$$

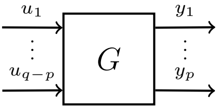
q-variate time series

$$\operatorname{rank} \mathscr{H}_{\ell+1}(w) \leq \underbrace{(\ell+1)q}_{\text{number of rows}} - p \iff \begin{bmatrix} R_0 & \cdots & R_\ell \end{bmatrix} \mathscr{H}_{\ell+1}(w) = 0, \; R_k \in \mathbb{R}^{p \times q}$$

$$\Updownarrow$$

$$R_0 w(t) + \cdots + R_\ell w(t+\ell) = 0, \quad t=1{:}T-\ell$$

$\updownarrow$ (Willems, 1986)



$w = \Pi \operatorname{col}(u, y)$ — trajectory of a dynamical system with lag $\leq \ell$ and $\leq q - p$ inputs

Low-rank approximation $\mathscr{H}_m(w) \approx$ l.r. $\mathscr{H}_m(\widehat{w})$	$\leftrightarrow$	System identification from w G — ?

Sylvester matrices

Two polynomials: $a(z) = \sum_{k=0}^{m} a_k z^k$ and $b(z) = \sum_{k=0}^{m} b_k z^k$

Sylvester matrix: $S(a,b) := \begin{bmatrix} a_0 & & & b_0 & & \\ \vdots & \ddots & & \vdots & \ddots & \\ \vdots & & a_0 & \vdots & & b_0 \\ a_m & & \vdots & b_m & & \vdots \\ & \ddots & \vdots & & \ddots & \vdots \\ & & a_m & & & b_m \end{bmatrix} \in \mathbb{R}^{2m \times 2m}$

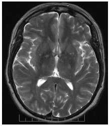
Theorem. $\deg \gcd(a,b) \geq d \iff \operatorname{rank} S(a,b) \leq 2m - d$,

Low-rank approximation $S(a,b) \approx$ l.r. $S(\widehat{a}, \widehat{b})$	$\leftrightarrow$	Approximate common divisor problem $\big(a(z), b(z)\big) \approx \big(p(z)h(z), q(z)h(z)\big)$

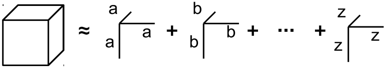
Other structures

- Multipolynomial Sylvester matrices
  — approximate GCD of multiple polynomials
- Multivariate Sylvester-like matrices (Macaulay matrices)
  — multivariate polynomials: approximate Gröbner bases, GCD
- Multilevel (nested) Hankel matrices
  — processing of n-way arrays

image processing
(MRI, textures)



symmetric tensor decomposition
(ICA, nonlinear SYSID)



Structured low-rank approximation: formulation

Linear structure: linear map $\mathscr{S} : \mathbb{R}^{n_\mathsf{p}} \to \mathbb{R}^{m \times n}$

Structured low-rank approximation: Given $\mathscr{S}$, $\|\cdot\|$, $p \in \mathbb{R}^{n_\mathsf{p}}$, $r < m$

$$\underset{\widehat{p} \in \mathbb{R}^{n_\mathsf{p}}}{\text{minimize}} \; \|p - \widehat{p}\| \quad \text{subject to} \quad \operatorname{rank} \mathscr{S}(\widehat{p}) \leq r,$$

Structured low-rank approximation: formulation

Linear structure: linear map $\mathscr{S} : \mathbb{R}^{n_\mathsf{p}} \to \mathbb{R}^{m\times n}$

Structured low-rank approximation: Given $\mathscr{S}$, $\|\cdot\|$, $p \in \mathbb{R}^{n_\mathsf{p}}$, $r < m$

$$\operatorname*{minimize}_{\widehat{p}\in\mathbb{R}^{n_\mathsf{p}}} \; \|p-\widehat{p}\| \quad \text{subject to} \quad \operatorname{rank} \mathscr{S}(\widehat{p}) \le r,$$

Weighted Euclidean semi-norm:

$$\|p\|_w^2 = \sum_{k=1}^{n_\mathsf{p}} w_k p_k^2, \qquad w_k \in [0;+\infty]$$

- $w_k = +\infty \iff$ constraint $p_k = \widehat{p}_k$ — fixed values
- $w_k = 0 \iff p_k$ and $\widehat{p}_k$ do not matter — missing values

Weighted semi-norm: examples

1. fixed values

$$\begin{bmatrix} & & a_0 & & & b_0 \\ & \iddots & \vdots & & \iddots & \vdots \\ a_0 & & \vdots & b_0 & & \vdots \\ \vdots & & a_m & \vdots & & b_m \\ \vdots & \iddots & & \vdots & \iddots & \\ a_m & & & b_m & & \end{bmatrix}$$

flipped Sylvester matrix

$\updownarrow$

Hankel matrices with fixed zeros

2. missing values

   ◦ $\begin{bmatrix} p_{1,1} & p_{1,2} & ? & p_{1,4} \\ p_{2,1} & p_{2,2} & p_{2,3} & p_{2,4} \\ p_{3,1} & p_{3,2} & p_{3,3} & ? \\ p_{4,1} & ? & p_{4,3} & p_{4,4} \end{bmatrix}$ — approximate matrix completion

   ◦ System identification with missing data, in particular:

   $\begin{bmatrix} \mathscr{H}_\ell(u_1) & \mathscr{H}_\ell(?) \\ \mathscr{H}_\ell(y_1) & \mathscr{H}_\ell(y_{\mathrm{ref}}) \end{bmatrix}$ — data-driven control

Reparameterization of the problem

$$\underset{\widehat{p}\in\mathbb{R}^{n_\mathsf{p}}}{\text{minimize}} \ \ \|p - \widehat{p}\|_w^2 \quad \text{subject to} \quad \operatorname{rank} \mathscr{S}(\widehat{p}) \leq r \qquad \text{(SLRA)}$$

rank constraint

kernel form

$$\operatorname{rank} \mathscr{S}(\widehat{p}) \leq r \quad \iff \quad d\ \overset{m}{\boxed{R}} \cdot \overset{n}{\boxed{\mathscr{S}(p)}} = 0,$$

corank (at least) $d := m - r$

a full row rank matrix

Reparameterization of the problem

$$\underset{\widehat{p}\in\mathbb{R}^{n_p}}{\text{minimize}} \ \|p-\widehat{p}\|_w^2 \quad \text{subject to} \quad \operatorname{rank} \mathscr{S}(\widehat{p}) \le r \qquad \text{(SLRA)}$$

rank constraint

kernel form

$$\operatorname{rank} \mathscr{S}(\widehat{p}) \le r \quad \iff \quad \begin{array}{c} \\ d \end{array} \overset{m}{\boxed{R}} \cdot \overset{n}{\boxed{\mathscr{S}(p)}} = 0,$$

a full row rank matrix

corank (at least) $d := m - r$

$$\text{(SLRA)} \iff \underset{R\in\mathbb{R}^{d\times m},\operatorname{rank} R=d}{\text{minimize}} \ f(R),$$

outer minimization

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

inner minimization

Inner minimization problem

$$f(R) := \left(\min_{\widehat{p} \in \mathbb{R}^{n_\mathrm{p}}} \|p - \widehat{p}\|_w^2 \ \text{ subject to } \ R\mathscr{S}(\widehat{p}) = 0 \right), \quad w_k \in [0; +\infty]$$

$\updownarrow$ see (Markovsky, Usevich, 2013)

$$\min_{x,u} \|x\|_2^2 \ \text{ subject to } \ A(R)x + \underbrace{B(R)u}_{\text{missing data}} = s(R)$$

generalized least-norm problem

$$f(R) = s^\top B_\perp^\top \left(B_\perp A (B_\perp A)^\top\right)^{-1} B_\perp s, \quad \text{where } B_\perp : \begin{array}{c}\mathrm{rowspan}(B_\perp) = \\ (\mathrm{colspan}(B))_\perp\end{array}$$

(Usevich, Markovsky, 2013): for $\mathscr{S}(p) \in \mathbb{R}^{m \times n}$ mosaic Hankel, $w_k > 0$, f, ∇f and Hessian can be evaluated in $O(m^2 n)$ flops.

Also, $f(R) = \frac{P(R)}{Q(R)}$

Outer minimization problem

$$f(R) := \Big(\min_{\widehat{p}\in\mathbb{R}^{n_\mathrm{p}}} \|p-\widehat{p}\|_w^2 \quad \text{subject to} \quad \begin{array}{cc} & m \\ d & \boxed{R} \end{array} \boxed{\mathscr{S}(p)} = 0, \Big),$$

full row rank

Note. f depends only on the row space of R:

$$\mathrm{rowspan}(R_1) = \mathrm{rowspan}(R_2) \Rightarrow f(R_1) = f(R_2)$$

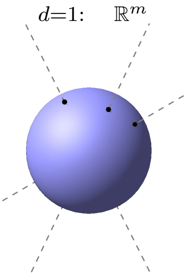
$$\Rightarrow \underset{R\in\mathbb{R}^{d\times m},\, \text{rank } R=d}{\text{minimize}} \; f(R) \quad \longleftrightarrow \quad \underset{\mathscr{L}\in\mathrm{Gr}(d,m)}{\text{minimize}} \; f(\mathscr{L})$$

Grassmann manifold: $\mathrm{Gr}(d,m) := \{d\text{-dim. subspaces of } \mathbb{R}^m\}$

Constrained minimization

$$\underset{\mathscr{L} \in \mathrm{Gr}(d,m)}{\text{minimize}} \; f(\mathscr{L}) \text{— ?},$$

$\mathrm{Gr}(d,m) := \{d\text{-dim. subspaces of } \mathbb{R}^m\}$



Constrained minimization:

$$\underset{R \in \mathbb{R}^{d\times m}}{\text{minimize}} \; f(R) \quad \text{subject to} \quad \underbrace{RR^\top = I}_{\text{orthonormal basis}}$$

For example, use penalty:

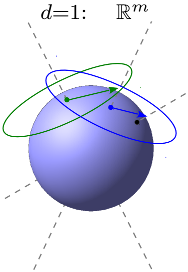
$$\underset{R \in \mathbb{R}^{d\times m}}{\text{minimize}} \; f(R) + \gamma \|RR^\top - I\|_F^2 \quad \text{— exact}$$

Retraction-based methods

$$\underset{\mathscr{L}\in\mathrm{Gr}(d,m)}{\text{minimize}}\; f(\mathscr{L}) \text{— ?},$$

(Absil, Mahoney, Sepulchre, 2008), and others ...

1. From $x_k \in \mathrm{Gr}(d,m)$
   choose direction ξ_k in the tangent space
2. Set $x_{k+1} = R_{x_k}(\xi_k)$ (retraction)
3. Go to step 1 (is stopping criteria not satisfied).



Optimization methods
on manidolds: gradient descent, trust-region, ...

Parametrizations with permutation matrices

For any full-rank $R \in \mathbb{R}^{d\times m}$ there exist d lin. indep. columns: 

$$\Downarrow$$

For any $\mathscr{L} \in \mathrm{Gr}(d,m)$ there exist $X \in \mathbb{R}^{d\times(m-d)}$ and permutation Π such that $\mathscr{L} = \mathrm{rowspan}\left([X\ I_d]\Pi\right)$

Take the permutation matrix:  $\Pi =$ 

Parametrizations with permutation matrices

For any full-rank $R \in \mathbb{R}^{d\times m}$ there exist d lin. indep. columns: 

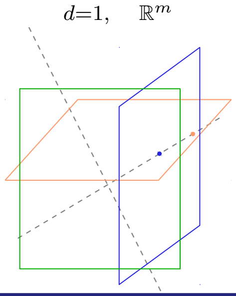
$$\Downarrow$$

For any $\mathscr{L} \in \mathrm{Gr}(d, m)$ there exist $X \in \mathbb{R}^{d\times(m-d)}$ and permutation Π such that $\mathscr{L} = \mathrm{rowspan}\left([X\ I_d]\Pi\right)$

Take the permutation matrix:  $\Pi =$ 

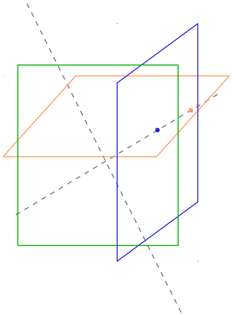
$$\underset{\mathscr{L}\in\mathrm{Gr}(d,m)}{\text{minimize}}\ f(\mathscr{L})$$

$$\iff \underset{\Pi\ \text{— perm.}}{\text{minimize}} \quad \min_{X\in\mathbb{R}^{d\times(m-d)}} f\left([X\ I_d]\Pi\right)$$



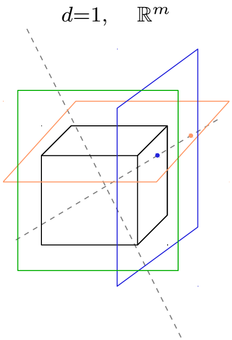
Optimization with permutations

$$\underbrace{\underset{\Pi \text{ — perm.}}{\text{minimize}}}_{\binom{m}{d} \text{ possibilities}} \quad \underbrace{\min_{X \in \mathbb{R}^{d \times (m-d)}}}_{\text{unbounded}} f\left([X \ I_d]\Pi\right)$$



Optimization with permutations

$$\underbrace{\underset{\Pi \text{ — perm.}}{\text{minimize}}}_{\binom{m}{d} \text{ possibilities}} \quad \underbrace{\min_{X \in [-1;1]^{d\times(m-d)}}}_{\cancel{\text{unbounded}}} f\left([X\ I_d]\Pi\right)$$



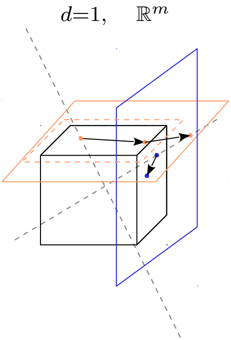
Theorem. (Knuth, 1985)
For any subspace $\mathscr{L} \in \mathrm{Gr}(d, m)$ there exists a representation $[X\ I_d]\Pi$ with $|X_{k,l}| \leq 1$.

Easy to prove for $d = 1$: if $R = \begin{bmatrix} r_1 & \cdots & r_m \end{bmatrix}$,

$$\text{then } \frac{R}{\max_k r_k} = \begin{array}{c} \begin{array}{ccc} & k & \end{array} \\ \begin{bmatrix} * & \pm 1 & * \end{bmatrix} \end{array}$$

Optimization with switching permutations

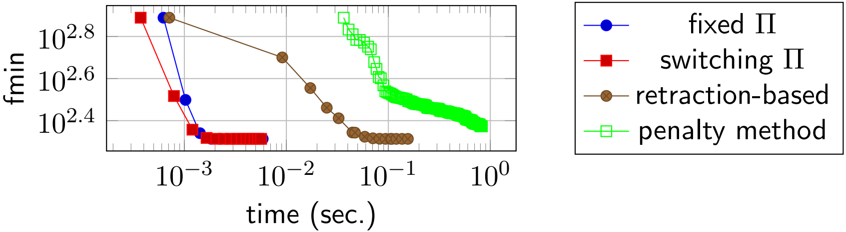
$$\underbrace{\underset{\Pi \text{ — perm.}}{\text{minimize}}}_{\binom{m}{d} \text{ possibilities}} \quad \underbrace{\min_{X\in[-1;1]^{d\times(m-d)}}}_{\cancel{\text{unbounded}}} f\left([X\ I_d]\Pi\right)$$



Switching permutations:

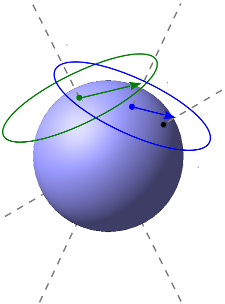
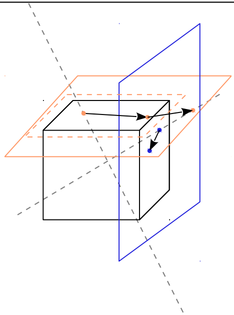
1. Perform local optimization of $f\left([X\ I_d]\Pi\right)$ until convergence, and unless $|X_{k,l}| \leq \Delta$ (where $\Delta > 1$)
2. If $|X_{k,l}| > \Delta$, switch the permutation, and go to step 1; otherwise stop.

Comparison of the methods



Structured low-rank approximation for system identification: an example from DAISY database, 6×801 block-Hankel matrix with 2×1 blocks (example #3 from the abstract).

Conclusions

	Retraction-based	permutation-based
		
$+++$	more adapted to the local geometry	simple, any optimization method can be used
$---$	complicated, every method needs to be adapted/tuned	no bounds on the number of switches

- SLRA as optimization on a Grassmann manifold

  K. Usevich and I. Markovsky. (2013).
  Optimization on a Grassmann manifold with application to system identification.
  Preprint. http://homepages.vub.ac.be/~kusevich/preprints.html

- SLRA with missing values

  I. Markovsky and K. Usevich (2013).
  Structured low-rank approximation with missing data.
  SIAM J. Matrix Anal. Appl. 34(2), 814-830.

- Fast cost function evaluation

  K. Usevich and I. Markovsky. (2013).
  Variable projection for affinely structured low-rank approximation in weighted 2-norms.
  J. Comput. Appl. Math. (doi:10.1016/j.cam.2013.04.034)

- Software (Matlab/R)

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

Thank you!