Optimization on the Grassmann manifold: a case study

Konstantin Usevich and Ivan Markovsky
Department ELEC, Vrije Universiteit Brussel

28 March 2013

32nd Benelux Meeting on Systems and Control, Houffalize, Belgium
Introduction

Optimization on the Grassmann manifold

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

\[
\min_{\mathcal{L} \in \text{Gr}(d, m)} f(\mathcal{L})
\]

Example: fitting \( x_1, \ldots, x_N \in \mathbb{R}^m \) to a \( d \)-dimensional subspace \( \mathcal{L} \)

\[
f_1(\mathcal{L}) = \sum_{k=1}^{N} \rho(x_k, \mathcal{L})
\]

Examples: computer vision, machine learning, system identification, ...

..., matrix/tensor low-rank factorization/approximation, ...
What is this talk about

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

$$\text{minimize } f(L) \quad \text{subject to } L \in \text{Gr}(d, m)$$

(Absil, Mahony, Sepulchre 2008):
How to define derivatives, how to optimize; atlases, tangent bundles, Riemannian metric, ...
What is this talk about

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

$$\minimize_{\mathcal{L} \in \text{Gr}(d, m)} f(\mathcal{L})$$

(Absil, Mahony, Sepulchre 2008):

How to define derivatives, how to optimize; atlases, tangent bundles, Riemannian metric, ...

This talk: (re)statement of the problem, reformulations as $\min_{X \in \mathbb{R}^n} f(X)$
minimize $f(R)$ subject to $R \in \mathcal{R}_f$,

where $\mathcal{R}_f := \{ R \in \mathbb{R}^{d \times m} \mid \text{rank } R = d \}$,

and $f(R) = f(UR) \quad \forall$ nonsingular $U \in \mathbb{R}^{d \times d}$.
Problem (re)statement

\[
\text{minimize } f(R) \text{ subject to } R \in \mathcal{R}_f,
\]

where \( \mathcal{R}_f := \{ R \in \mathbb{R}^{d \times m} \mid \text{rank } R = d \} \),

and \( f(R) = f(U R) \quad \forall \text{nonsingular } U \in \mathbb{R}^{d \times d} \).

Problems:

- How to impose the constraint \( R \in \mathcal{R}_f \)？
- If \( R_\ast \) is a minimum, then \( UR_\ast \) is also a minimum...

A solution:

- Reduce the search space
Reduction of the search space

\begin{align*}
\text{minimize} & \quad f(R) \quad \text{subject to} \quad R \in \mathcal{R}_f, \\
\text{where} & \quad \mathcal{R}_f := \{R \in \mathbb{R}^{d \times m} \mid \text{rank } R = d\}, \\
\text{and} & \quad f(R) = f(U R) \quad \forall \text{ nonsingular } U \in \mathbb{R}^{d \times d}.
\end{align*}

How to reduce the search space:

1. \(RR^\top = I_d\) — orthonormal basis
   represents all subspaces, nonlinear constraint
Reduction of the search space

\[
\begin{align*}
\text{minimize} & \quad f(R) \quad \text{subject to} \quad R \in \mathcal{R}_f, \\
\text{where} & \quad \mathcal{R}_f := \{ R \in \mathbb{R}^{d \times m} \mid \text{rank } R = d \}, \\
\text{and} & \quad f(R) = f(U R) \quad \forall \text{ nonsingular } U \in \mathbb{R}^{d \times d}.
\end{align*}
\]

How to reduce the search space:

1. \( R R^\top = I_d \) — orthonormal basis
   represents all subspaces, nonlinear constraint

2. \( R = \begin{bmatrix} X & I_d \end{bmatrix} \), where \( X \in \mathbb{R}^{d \times (m-d)} \)
   represents almost all subspaces
   (all \( R \) with nonsingular \( R_{:, (m-d+1):m} \))
Outline

Introduction: problem (re)statement

Optimization over $[X \quad I_d]\Pi$

Optimization with $RR^\top = I_d$

A case study
Parametrizations with permutation matrices

\[ [X \quad I_d] \] do not represent all \( d \)-dimensional subspaces

Solution: consider all permutations \( \Pi \in \mathbb{R}^{m \times m} \)

and matrices of the form \( [X \quad I_d] \, \Pi \),

Corresponds to fixing \( R_{\cdot,\mathcal{J}} = I_d \)
Parametrizations with permutation matrices

\([X \quad I_d]\) do not represent all \(d\)-dimensional subspaces

Solution: consider all permutations \(\Pi \in \mathbb{R}^{m \times m}\)
and matrices of the form \([X \quad I_d] \Pi\),

Corresponds to fixing \(R_{:,J} = I_d\)

\(d=1, \quad \mathbb{R}^m\)
Parametrizations with permutation matrices

\[ [X, I_d] \] do not represent all \( d \)-dimensional subspaces

Solution: consider all permutations \( \Pi \in \mathbb{R}^{m \times m} \) and matrices of the form \( [X, I_d] \Pi \),

Corresponds to fixing \( R_{:, \mathcal{J}} = I_d \)
Parametrizations with permutation matrices

\[
\begin{bmatrix} X & I_d \end{bmatrix} \text{ do not represent all } d\text{-dimensional subspaces}
\]

Solution: consider all permutations \( \Pi \in \mathbb{R}^{m \times m} \) and matrices of the form \( \begin{bmatrix} X & I_d \end{bmatrix} \Pi \),

Corresponds to fixing \( R_{:,\mathcal{J}} = I_d \)
Parametrizations with permutation matrices

\[ [X \quad I_d] \] do not represent all \( d \)-dimensional subspaces

Solution: consider all permutations \( \Pi \in \mathbb{R}^{m \times m} \) and matrices of the form \( [X \quad I_d] \Pi \),

Corresponds to fixing \( R_{:,\mathcal{I}} = I_d \)

Moreover, for \( d = 1 \) we can select \( \Pi: |X_{1,j}| \leq 1 \)

Question: for \( d > 1 \), can we consider only \( X \) from a bounded subset of \( \mathbb{R}^{d \times (m-d)} \)?
Bounded parametrizations with permutation matrices

**Theorem** (Knuth, 1980)

For any $R \in \mathbb{R}^{d \times m}$, rank $R = d$, there exist $U$ and $\Pi$ such that

$$UR = [X \quad I_d] \Pi, \quad |X_{ij}| \leq 1 \quad \text{for all } (i, j)$$
Bounded parametrizations with permutation matrices

**Theorem** (Knuth, 1980)
For any $R \in \mathbb{R}^{d \times m}$, rank $R = d$, there exist $U$ and $\Pi$ such that

$$UR = \begin{bmatrix} X & I_d \end{bmatrix} \Pi, \quad |X_{i,j}| \leq 1 \quad \text{for all } (i, j)$$

**Proof:** (sketch)

1. Find $d \times d$ submatrix with maximal determinant (maximal volume)

\[ J_0 = \arg \max_{\mathcal{J}} |\det R_{:,\mathcal{J}}| \quad \Rightarrow \quad R = \begin{bmatrix} \text{maximal volume} \end{bmatrix} \]

2. Take $\Pi_0$ that permutes $R_{:,J_0}$ and $R_{:, (m-d+1):m}$

\[ R = \begin{bmatrix} \text{maximal volume} \end{bmatrix} \quad \rightarrow \quad R\Pi_0 = \begin{bmatrix} \text{permuted} \end{bmatrix} \]

3. Take $\begin{bmatrix} X & I_d \end{bmatrix} = (R_{:,J_0})^{-1} R\Pi_0$

By Kramer’s rule we have that $|X_{i,j}| \leq 1$
Optimization over $[X \quad I_d] \Pi$

Optimization over permutations

$$\minimize_{R \in \mathcal{R}_f} f(R)$$

is equivalent to

$$\minimize_{\Pi \text{ – perm.}} \min_{X \in [-1; 1]^{d \times (m-d)}} f([X \quad I_d] \Pi)$$

There are \(\frac{m!}{(m-d)!}\) permutations \(\Pi\)...
Optimization over permutations

\[
\minimize_{\Pi \in \text{perm.}} \min_{X \in [-1;1]^{d \times (m-d)}} f \left( [X \ I_d] \ \Pi \right)
\]

There are \( \frac{m!}{(m-d)!} \) permutations \( \Pi \)...

.. we can switch the permutation in the course of optimization, if \( |X_{i,j}| > \delta \geq 1 \) for some \( i, j \).
Outline

Introduction: problem (re)statement

Optimization over $[X \quad I_d] \Pi$

Optimization with $RR^\top = I_d$

A case study
Optimization with $RR^\top = I_d$

Orthonormal basis

\[
\text{minimize } f(R) \quad \text{subject to } RR^\top = I_d
\]

Disadvantage: nonlinear constraint

Possible solutions:
1. Lagrange multipliers,
2. penalty: $f(R) + \gamma\|RR^\top - I_d\|_F^2$

$\gamma \to \infty$? not at all.
Penalty method for orthonormal bases

\begin{align*}
\text{minimize } & f(R) \quad \text{subject to } \quad RR^\top = I_d \\
\end{align*}

where \( f(R) = f(UR) \forall \text{nonsingular } U \)

**Theorem.** For any \( \gamma > 0 \), the local minima of

\begin{align*}
\text{minimize } & f(R) + \gamma \| RR^\top - I_d \|_F^2 \\
\end{align*}

coincide with the local minima of \((*)\).
Penalty method for orthonormal bases

\[
\begin{align*}
\text{minimize } f(R) & \quad \text{subject to } RR^\top = I_d \\
& \quad R \in \mathcal{R}_f
\end{align*}
\]  

(\*)

where \( f(R) = f(U R) \) \( \forall \) nonsingular \( U \)

**Theorem.** For any \( \gamma > 0 \), the local minima of

\[
\begin{align*}
\text{minimize } f(R) + \gamma \| RR^\top - I_d \|_F^2 \\
& \quad R \in \mathcal{R}_f
\end{align*}
\]

coincide with the local minima of (\*).

**Proof.** Let \( R = U \Sigma V \) be an SVD of \( R \). Then

\[
f(R) + \gamma \| RR^\top - I_d \|_F^2 \geq f(V) + \gamma \| V V^\top - I_d \|_F^2 = f(V) = f(R)
\]
Outline

Introduction: problem (re)statement

Optimization over \([X \quad I_d]\)\(\Pi\)

Optimization with \(RR^\top = I_d\)

A case study
Structured low-rank approximation

System identification, model reduction, etc. → structured low-rank approximation

Structured low-rank approximation: given \( p \in \mathbb{R}^{np} \), \( \mathcal{S} \) and \( r < m \)

\[
\minimize_{\hat{p}} \|p - \hat{p}\|_2^2 \quad \text{subject to} \quad \text{rank} \mathcal{S}(\hat{p}) \leq r. \quad (\text{SLRA})
\]
Structured low-rank approximation

System identification, model reduction, etc. → structured low-rank approximation

Structured low-rank approximation: given $p \in \mathbb{R}^n$, $\mathcal{S}$ and $r < m$

$$
\minimize_{\hat{p}} \| p - \hat{p} \|_2^2 \quad \text{subject to} \quad \text{rank } \mathcal{S}(\hat{p}) \leq r.
$$

( SLRA )

Kernel representation + variable projection:

$$
\minimize_{R \in \mathbb{R}^{d \times m}} \ f(R) := \left( \min_{\hat{p}} \| p - \hat{p} \|_2^2 \quad \text{subject to} \quad R \mathcal{S}(\hat{p}) = 0 \right),
$$

where $d = m - r$. 
Structured low-rank approximation: an example

Hankel SLRA: given \( w = (w_1, \ldots, w_n) \),

\[
g([r_1 \ r_2 \ r_3]) := \min \| w - \hat{w} \|_2^2 \quad \text{subject to} \]

\[
\begin{bmatrix} r_1 & r_2 & r_3 \\ r_2 & r_3 & \hat{w}_1 \\ r_3 & \hat{w}_2 & \hat{w}_3 \\ \hat{w}_1 & \hat{w}_2 & \hat{w}_3 & \cdots & \hat{w}_{n-2} \\ \hat{w}_2 & \hat{w}_3 & \hat{w}_{n-2} & \hat{w}_{n-1} \\ \hat{w}_3 & \cdots & \hat{w}_{n-2} & \hat{w}_{n-1} & \hat{w}_n \end{bmatrix} = 0
\]

\( \min f(R) \) — solves the SLRA problem

\( f(\alpha R) = f(R) \quad \forall \alpha \neq 0 \)
A case study: $f([x_1 \, x_2 \, 1])$ (3 × 10 Hankel matrix)
Mosaic Hankel low-rank approximation (LTI system identification of DAISY datasets), approximation errors are measured as

\[ 100\% \cdot (1 - \frac{\|\hat{p}^* - p\|_2^2}{\|p\|_2^2}) \]

<table>
<thead>
<tr>
<th>t. #</th>
<th>([X \  I_d])</th>
<th>([X \  I_d]\Pi)</th>
<th>genrtr</th>
<th>(RR^\top = I_d)</th>
<th>(|RR^\top - I_d|_F^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>99.96</td>
<td>99.96</td>
<td>99.96</td>
<td>99.96</td>
<td>99.96</td>
</tr>
<tr>
<td>2</td>
<td>99.93</td>
<td>99.93</td>
<td>99.92</td>
<td>99.92</td>
<td>99.91</td>
</tr>
<tr>
<td>3</td>
<td>98.66</td>
<td>98.66</td>
<td>98.66</td>
<td>97.8</td>
<td>98.58</td>
</tr>
<tr>
<td>4</td>
<td>96.45</td>
<td>96.45</td>
<td>95.69</td>
<td>95.71</td>
<td>95.69</td>
</tr>
<tr>
<td>5</td>
<td>90.26</td>
<td>90.27</td>
<td>86.46</td>
<td>85.69</td>
<td>82.71</td>
</tr>
</tbody>
</table>

**Table:** Percentage fits of the methods
Conclusions

• 2 reformulations as optimization on Euclidean space $\rightarrow$ freedom in choosing the optimization method

• Issues: bound on the number of switches, choosing regularization parameter...

References:


• [http://github.com/slra/slra/](http://github.com/slra/slra/) — software and reproducible experiments