Data fitting on positive-semidefinite matrices

Estelle Massart, Julien M. Hendrickx, P.-A. Absil

ICTEAM, UCLouvain, Belgium

13 June 2019
Positive-semidefinite matrices

\[ A \succeq 0 \iff A = A^\top \text{ and } \lambda_i \geq 0 \ \forall i = 1, \ldots, n \]

For \( n = 2 \):

For \( n = 2 \):

\[
A = \begin{pmatrix} a & b \\ b & c \end{pmatrix} \in \mathcal{P}_2
\]

\[
\begin{aligned}
\tau(A) &= a + c > 0 \\
\det(A) &= ac - b^2 > 0
\end{aligned}
\]
In this talk, two problems...

Data fitting on PSD matrices

Averaging on $\mathcal{P}_n$

Fitting a curve on $S_+(p, n)$
Part I: Averaging positive-definite matrices
Applications

How to define a mean on $\mathcal{P}_n$?
Swelling effect with the arithmetic mean

\[ A = \begin{pmatrix} 5 & 0 \\ 0 & 1 \end{pmatrix} \]

\[ B = \begin{pmatrix} 1 & 0 \\ 0 & 5 \end{pmatrix} \]

\[ \gamma(t) = (1-t)A + tB \]

Undesirable in several applications!
Solution: see \( \mathcal{P}_n \) as a Riemannian manifold, and change the metric!
What is a manifold?

- Euclidean space
- Sphere, torus, cylinder, ...

Some constrained matrix sets:
- Orthogonal matrices
- $\mathcal{P}_n$
- $S_+(p, n)$
Riemannian manifold

\[ g_X(\eta_X, \xi_X) = \text{tr}(\eta_X X^{-1} \xi_X X^{-1}). \]

\[ \delta(A, B) = \| \text{Log}(A^{-\frac{1}{2}} BA^{-\frac{1}{2}}) \|. \]

\[ \gamma(t) = A^{\frac{1}{2}} \left(A^{-\frac{1}{2}} BA^{-\frac{1}{2}} \right)^t A^{\frac{1}{2}}. \]
Riemannian barycenter

\[ G(A_1, \ldots, A_N) = \arg \min_{X \in \mathcal{P}_n} \sum_{i=1}^{N} \delta^2(X, A_i) \]

with

\[ \delta(A, B) = \| \text{Log}(A^{-\frac{1}{2}} BA^{-\frac{1}{2}}) \|_F \]
Part I: Averaging

\[ G(A, B) = A^{\frac{1}{2}} \left( A^{-\frac{1}{2}} BA^{-\frac{1}{2}} \right)^{\frac{1}{2}} A^{\frac{1}{2}} \]

\[ A = \begin{pmatrix} a & b \\ b & c \end{pmatrix} \in \mathbb{P}_2 \]

\[ \begin{cases} \tau(A) = a + c > 0 \\ \det(A) = ac - b^2 > 0 \end{cases} \]
Computing the Riemannian barycenter?

\[ G(A_1, \ldots, A_N) = \arg \min_{X \in \mathcal{P}_n} \sum_{i=1}^{N} \delta^2(X, A_i) \]

Many approaches rely on optimization on manifolds, e.g., Steepest Descent (Pennec, Manton, Bini & Iannazzo, Jeuris et al)
Exponential and logarithm maps

\[ \mathcal{T}_{X_1} \mathcal{M} \]

\[ X_1 \cdot \eta_{X_1} \]

\[ \cdot X_2 = \text{Exp}_{X_1} \eta_{X_1} \]
Steepest descent on manifolds

\[ \mathcal{T}_X \mathcal{M} \]

\[ X \cdot -\nabla f(X) \]

\[ \cdot X_2 = \text{Exp}_X(-\nabla f(X)) \]

Part I: Averaging
Our approach

Incremental gradient descent (IGD) on manifolds.

\[ f(X) = \sum_{i=1}^{N} f_i(X) \]

At each iteration, pick some \( i \) and perform a gradient step to decrease \( f_i \). In our case:

\[ f(X) = \sum_{i=1}^{N} \delta^2(X, A_i). \]
IGD for computing the Riemannian barycenter.

\[ X_1 = A_1 \]
\[ X_2 \]
\[ X_3 \]
\[ X_4 \]
\[ X_5 = X^* \]

Figure: In \( \mathbb{R}^2 \).

Figure: In \( \mathcal{P}_2 \).
Bias on $\mathcal{P}_2$

$\delta(X_N, A_i) / \delta(G, A_i)$

Part I: Averaging
Extending the sequence

\[ X_1 = A_1 \]

\[ X_2, X_3, X_4, X_5, X_6, X_7 \]

\[ A_2, A_3, A_4, A_5 \]

Part I: Averaging
Principle: move towards the data in a cyclic way

Order:

\[
\begin{align*}
A_1 & \quad A_2 & \quad \ldots & \quad A_N \\
& \quad \overbrace{\text{iteration 1}} & & \\
A_1 & \quad A_2 & \quad \ldots & \quad A_N \\
& \quad \overbrace{\text{iteration 2}} & & \\
A_1 & \quad A_2 & \quad \ldots & \quad A_N \\
& \quad \overbrace{\text{iteration 3}} & & \\
A_1 & \quad A_2 & \quad \ldots & \quad A_N & \quad \ldots \\
& \quad \overbrace{\text{iteration 4}} & & \\
\end{align*}
\]
Remember...

Part I: Averaging

\[ \frac{\delta(X_N,A_i)}{\delta(G,A_i)} \]
What if we regularly shuffle the data points?
Part one: reverse

\[ A_1 A_2 \ldots A_N \]

iter 1

\[ A_1 A_2 \ldots A_N \]

iter 2

\[ A_1 A_2 \ldots A_N \]

iter 3

\[ A_1 A_2 \ldots A_N \]

iter 4

\[ \downarrow \]

\[ A_1 A_2 \ldots A_N \]

iter 1

\[ A_N A_{N-1} \ldots A_1 \]

iter 2 = reverse(iter 1)

\[ A_1 A_2 \ldots A_N \]

iter 3

\[ A_N A_{N-1} \ldots A_1 \]

iter 4 = reverse(iter 3)

Part I: Averaging
Part two: *in-shuffle*

\[
\begin{align*}
A_1 & \ A_2 & \cdots & A_N \\
& \text{iter 1} & \\
A_N & A_{N-1} & \cdots & A_1 \\
& \text{iter 2=} \text{reverse(iter 1)} & \\
A_{p(1)} & \cdots & A_{p(N)} \\
& \text{iter 3} & \\
A_{p(N)} & \cdots & A_{p(1)} \\
& \text{iter 4=} \text{reverse(iter 3)}
\end{align*}
\]

iter 3 = *in-shuffle* (iter 1)
Choice of the permutations: in-shuffling

Part I: Averaging
Choice of the permutations: in-shuffling

Part I: Averaging
Choice of the permutations: in-shuffling
Choice of the permutations: in-shuffling
Improvement of the convergence rate

Part I: Averaging

![Graph showing the convergence rate for different methods: Circular, Random without replacement, Random with replacement, and Shuffled. The x-axis represents iterations from 1 to 10, and the y-axis represents the relative error $E_{rel}$ on a logarithmic scale from $10^{-3}$ to $10^0$. The graph illustrates how the convergence rate improves with different methods as the number of iterations increases.](image-url)
We proposed an **accelerated IGD algorithm**, trying to remedy to the **bias** arising with the cyclic ordering.

A **faster convergence** is observed numerically, even faster than SGD while being deterministic.

The IGD / SGD algorithms have a sublinear convergence rate, with a fast initial decrease. Possible to build **hybrid methods**, starting with IGD and moving to GD.

**Convergence** of the algorithm is proven in the general framework of NPC spaces, under weak assumptions on the shuffling strategy.
Conclusion Part I

- Applied to **EEG classification** (incremental algorithms allow online learning), see *Estelle Massart, Sylvain Chevallier, Inductive means and sequences applied to online classification of EEG, Geometric Science of Information: Third International Conference (GSI 2017), Paris, France, 2017*

- For more information, see: *Estelle M. Massart, Julien M. Hendrickx, P-A. Absil, Matrix geometric means based on shuffled inductive sequences, Linear Algebra and its Applications, 252, pp. 334-359 (2018)*
Part II: Interpolation on $S_+(p, n)$
Applications:

- Generation of transitions between protein conformations
- Action recognition in videos
- Wind field modeling
- Parametric model order reduction
- Graph theory
The manifold of fixed-rank PSD matrices

- The set of PSD matrices is difficult to deal with → restrict the analysis to **fixed-rank PSD matrices**.

**Definition:**

\[ S_+ (p, n) = \{ C \in \mathbb{R}^{n \times n} | C = C^\top, x^\top C x \geq 0 \forall x \in \mathbb{R}^n, \text{rank}(C) = p \} \].

For using curve fitting algorithms on \( S_+ (p, n) \): need expressions for

- Riemannian exponential (already known)
- Riemannian logarithm → computed here :-)

Part II: Interpolation
The manifold of fixed-rank PSD matrices

Proposition:

\[ A \in S_+(p, n) \iff A = YY^T, \ Y \in \mathbb{R}_*^{n \times p}. \]

\[
\begin{bmatrix}
A \\
n \times n
\end{bmatrix}
= 
\begin{bmatrix}
Y \\
n \times p
\end{bmatrix}
\begin{bmatrix}
Y^T \\
p \times n
\end{bmatrix}
\]
The manifold of fixed-rank PSD matrices

Proposition:

\[ A \in S_+(p, n) \iff A = YY^\top, \ Y \in \mathbb{R}^{n \times p}. \]

\[
\begin{bmatrix}
A_{n \times n} \\
\end{bmatrix} = 
\begin{bmatrix}
Y_{n \times p} \\
\end{bmatrix}
QQ^\top
\begin{bmatrix}
Y^\top_{p \times n} \\
\end{bmatrix}
\]

Non-unique decomposition!
Quotient geometry

Each point of $S_+(p, n)$ is identified to a set of points in $\mathbb{R}^{n \times p}_*$

\[
\{ Y | YY^\top = C_2 \}
\]

\[
\{ Y | YY^\top = C_3 \}
\]

\[
\{ Y | YY^\top = C_1 \}
\]
Example: case $S_+(1, 2)$

\[
\begin{bmatrix}
1 & 0 \\
0 & 0
\end{bmatrix}
= \begin{bmatrix}
1 \\
0
\end{bmatrix}
\begin{bmatrix}
1 & 0
\end{bmatrix}

= \begin{bmatrix}
-1 \\
0
\end{bmatrix}
\begin{bmatrix}
-1 & 0
\end{bmatrix}
\]
Example: case $\mathcal{S}_+(1, 2)$

\[
\begin{bmatrix}
1 & 0 \\
0 & 0
\end{bmatrix} = \begin{bmatrix}
1 \\
0
\end{bmatrix} \begin{bmatrix}
1 & 0
\end{bmatrix} = \begin{bmatrix}
-1 \\
0
\end{bmatrix} \begin{bmatrix}
-1 & 0
\end{bmatrix}
\]

\[
\begin{bmatrix}
0 & 0 \\
0 & 1
\end{bmatrix} = \begin{bmatrix}
0 \\
1
\end{bmatrix} \begin{bmatrix}
0 & 1
\end{bmatrix} = \begin{bmatrix}
0 \\
-1
\end{bmatrix} \begin{bmatrix}
0 & -1
\end{bmatrix}
\]
Example: case $S_+(2, 3)$

$$
\begin{bmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 0
\end{bmatrix}
= \begin{bmatrix}
1 & 0 \\
0 & 1 \\
0 & 0
\end{bmatrix}
\begin{bmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 0
\end{bmatrix}
= \begin{bmatrix}
1 & 0 \\
0 & 1 \\
0 & 0
\end{bmatrix}QQ^\top\begin{bmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 0
\end{bmatrix}
$$

Y(:,1) Y(:,2)
Geodesics in $S_+ (p, n)$

**Proposition** [Journée et al, 2010]
Geodesics in $S_+ (n, p)$ are obtained by projection of straight lines in $\mathbb{R}_*^{n \times p}$, under the map $Y \rightarrow YY^T$, between well-chosen representatives of the equivalence classes.
Geodesics in $S_+(p, n)$

$$C(t) = Y(t)Y(t)^\top$$

\[ \{ Y | YY^\top = C_1 \} \]

\[ \{ Y | YY^\top = C_2 \} \]
Example of geodesic on $S_+(1, 2)$.
Example of geodesic on $S_+(1, 2)$. 

Part II: Interpolation
Example of geodesic on $\mathcal{S}_+(1, 2)$.

\[
C(t) = \bar{Y}(t)\bar{Y}(t)^T
\]
**Proposition:** [M., Absil, 2018]

Let $C_1, C_2 \in S_+(p, n)$, and let $Y_1, Y_2 \in \mathbb{R}^{n \times p}$ be such that $C_1 = Y_1 Y_1^\top$ and $C_2 = Y_2 Y_2^\top$. Then, there exist several shortest paths between $C_1$ and $C_2$ if and only if

$$\det(Y_1^\top Y_2) = 0.$$ 

**Remark:**

This condition does not depend on the choice of the $Y$-factors.
Illustration in \( S_+(1, 2) \):
Injectivity radius on $S_+(p, n)$

Proposition: [M., Absil, 2018]  
The injectivity radius of $S_+(p, n)$ at $C \in S_+(p, n)$ is:

$$r_{\text{Inj}}(C) = \sqrt{\lambda_p(C)}.$$

Corollary: [M., Absil, 2018]  

$$r_{\text{Inj}}(C) \to 0 \text{ when } C \to \text{ border of the manifold.}$$
Proposition: [M., Absil, 2018]
Let $Y_0, Y_1 \in \mathbb{R}^{n \times p}$ be such that $\det(Y_0^\top Y_1) \neq 0$. Then, the horizontal lift at $Y_0$ of the Riemannian logarithm $\text{Log}_\pi(Y_0)\pi(Y_1)$ is

$$\text{Log}_\pi(Y_0)\pi(Y_1)|_{Y_0} = Y_1 Q^\top - Y_0,$$

where $Q$ is the orthogonal factor of the (unique) polar decomposition of $Y_0^\top Y_1$. 
The associated Riemannian distance corresponds to the Wasserstein distance between degenerate Gaussian distributions.

**Corollary: [M., Absil, 2018]**

The associated Riemannian distance is:

\[
d(S_1, S_2) = \left[ \text{tr} S_1 + \text{tr} S_2 - 2\text{tr} \left( S_1^{1/2} S_2 S_1^{1/2} \right)^{1/2} \right]^{1/2}.
\]
Conclusions

- Several metrics exist for $S_+(p, n)$, we consider here the identification $S_+(p, n) \simeq \mathbb{R}_{\ast}^{n \times p} / O_p$.

- We obtained expressions for the Riemannian logarithm, injectivity radius and Riemannian distance.

- The associated Riemannian distance coincides with the Wasserstein metric between degenerate Gaussian distributions.

- Now that expressions for the Riemannian exponential and logarithm are available, we can apply curve fitting algorithms on $S_+(p, n)$. 

Part II: Interpolation
Codes for the Riemannian exponential and logarithm on $S_+(p, n)$ have been recently added to Manopt, a widely used toolbox for optimization on manifolds.

For more information, see: Estelle Massart, P.-A. Absil, Quotient geometry with simple geodesics for the manifold of fixed-rank positive-semidefinite matrices, Technical report (2018)

Thanks for your attention!

Refs:
- Estelle Massart, Sylvain Chevallier, Inductive means and sequences applied to online classification of EEG, Geometric Science of Information: Third International Conference (GSI 2017), Paris, France.