Geometric means of fixed-rank positive semi-definite matrices

Silvère Bonnabel
Co-workers: Anne Collard, Rodolphe Sepulchre (Ulg)

Mines ParisTech
Centre de Robotique
Mathématiques et Systèmes
silvere.bonnabel@mines-paristech.fr

Matrix Information Geometry workshop
Ecole Polytechnique
Feb 24th, 2011

Motivations

Positive definite matrices appear in many contexts:

- statistics and information geometry: covariance matrices;
- optimization: unknowns in convex and semidefinite programming;
- system theory: unknowns in Lyapunov equations and LMIs;
- machine learning: kernels for distance learning and classification;
- biomedical imaging: tensors in diffusion MRI;
- radar signal processing, ...

BUT

Low-rank approximations are necessary to handle large-scale problems:

$$O(n^3) \rightarrow O(np^2)$$

A basic and common issue

Define:

$$P_+(n) = \{X \in \mathbb{R}^{n \times n} \mid X = X^T \succ 0\}$$

and, for any $1 \leq p < n$ the sets,

$$S^+(p, n) = \{X \in \mathbb{R}^{n \times n} \mid X = X^T \succeq 0, \ \mathrm{rank X = p}\}$$

Take two points in those sets: distance? connecting geodesic (shortest path)? Geometric mean? i.e: what is the right Riemannian geometry?

Essential to optimization, filtering, interpolation, fusion, completion, learning, ...

Illustrative example 1

Vector-valued image and tensor computing

Results of several filtering methods on a 3D DTI of the brain[1]:

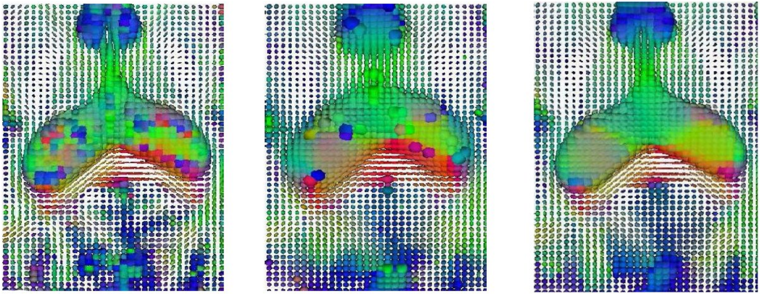


Figure: Original image "Vectorial" filtering "Riemannian" filtering

- Large scales problems involving basic operations on positive definite matrices
- Does geometry matter??

[1] Courtesy from Xavier Pennec (INRIA Sophia Antipolis)

Illustrative example 2

Vector-valued image and Doppler radar

Results of filtering methods for detection of small targets in sea clutter with HF coastal Doppler radar to detect a small target[2]:

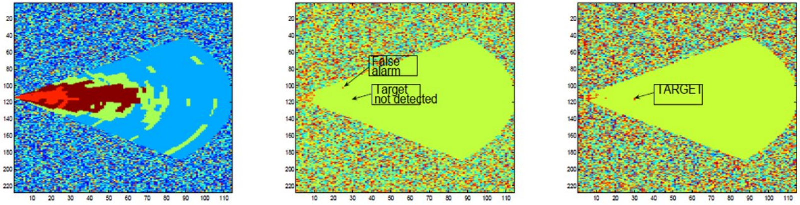


Figure: Doppler Radar Fourier filtering “Riemannian" filtering

[2] Courtesy of Frederic Barbaresco (Thales Air Systems, Strategy Technology & Innovation Department)

Illustrative example 3

Computing with kernels

- Data fusion and kernel combination:
  a kernel is constructed for each data source i. Problem: infer an average kernel for classification purposes.

- Kernel completion: construct a kernel from incomplete data.[3]

- Kernel online learning: integrate new data in an existing kernel[4].

---

[3] Bach, Lankriet, Jordan (ICML, 2004) De Bie, Tranchevent, Van Oeffelen, Moreau (Bioinformatics, 2007)

[4] Tsuda, Ratsch, Warmut (JMLR 2005) Kulis, Sustik, Dhillon (ICML 2006) Meyer, Bonnabel, Sepulchre (JMLR 2011)

Outline

Positive definite case : natural metric and information geometry

Semidefinite positive case: two quotient geometries to compute distances and means

Polar geometry: geometric mean of an arbitray number of low rank positive semidefinite matrices

Some filtering illustrations

More details about the construction of the mean

The positive definite case

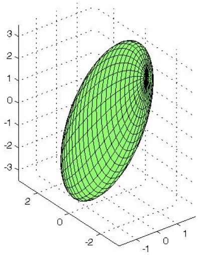
A basic and common issue in the cone

Define:

$$P_+(n) = \{X \in \mathbb{R}^{n\times n} \mid X = X^T \succ 0\}$$

Take two points in those sets: distance? connecting geodesic (shortest path)? Geometric mean? i.e: what is the right Riemannian geometry?

Essential to optimization, filtering, interpolation, fusion, completion, learning, ...



A much studied issue in the cone[5]

Let $X_1 = X_1^T \succ 0$ and $X_2 = X_2^T \succ 0$ in $P_+(n)$

The natural geometry of the cone leads to the following distance

$$d(X_1, X_2) = \| \log(X_1^{-\frac{1}{2}} X_2 X_1^{-\frac{1}{2}}) \| = \Big(\sum_k (\log \lambda_k)^2\Big)^{\frac{1}{2}}$$

with $\det(X_1 X_2^{-1} - \lambda_k I) = 0$. The geodesics have the form

$$X_1^{1/2} \exp(t \log(X_1^{-1/2} X_2 X_1^{-1/2})) X_1^{1/2}$$

The mean point is called “geometric" mean and is obtained for $t = 1/2$:

$$X_1 \circ X_2 = X_1^{1/2}(X_1^{-1/2} X_2 X_1^{-1/2})^{1/2} X_1^{1/2} \quad (= \sqrt{X_1 X_2} \text{ for } X_1, X_2 \in \mathbb{R}_*^+)$$

[5] Ando et al. (2004), Moakher (2005), Petz and Temesi (2005), Smith (2006), Arsigny et al. (2007), Barbaresco (2008), Bhatia (2006) . . .

Properties of the natural metric

- Invariance to invertible transformations and inversion. For any $P \in GL(n)$:
  - $d(X_1, X_2) = \left(\sum_k (\log \lambda_k)^2\right)^{\frac{1}{2}} = d(PX_1P^T, PX_2P^T)$
  - $d(X_1, X_2) = d(X_1^{-1}, X_2^{-1})$ (ex: precision matrix)

Indeed $\det(X_1 - \lambda_k X_2) = 0 \Leftrightarrow \det(P(X_1 - \lambda_k X_2)P^T) = 0$.

- Geodesically complete: $P_+(n)$ is NOT a vector space :
  - $X_1 + t(X_2 - X_1)$ does not remain in $P_+(n)$.
  - $X_1^{1/2} \exp(t \log(X_1^{-1/2} X_2 X_1^{-1/2})) X_1^{1/2}$ does for all $t > 0$

Natural geometry of the cone.

Does it generalize to the facets?

Fisher Information Metric and Gaussian distributions

For multivariate Gaussian distributions with zero mean $X \sim \mathcal{N}(0, \Sigma)$ (we suppose $\Sigma \in P_+(n)$):

$$f(x \mid \Sigma) = \frac{1}{(2\pi)^{n/2} \mid \Sigma \mid^{1/2}} \exp(-x^T \Sigma^{-1} x)$$

The Fisher Information metric yields the following distance

$$d_F(\Sigma_1, \Sigma_2) = \| \log(\Sigma_1^{-\frac{1}{2}} \Sigma_2 \Sigma_1^{-\frac{1}{2}}) \| = \Big(\sum_k (\log \lambda_k)^2\Big)^{\frac{1}{2}}$$

independent of a arbitrary choice of coordinate $X \mapsto PX$, and $d(\Sigma_1, \Sigma_2) = d(\Sigma_1^{-1}, \Sigma_2^{-1})$.

Natural metric on $P_+(n)$: advantages

- Links with information geometry (natural for covariance matrices)
- Invariance to geometric transformation (units etc.)
- Invariance to matrix inversion
- Induced geometric mean : robustness to outliers
- Interior points methods and geodesic completeness

The invariance properties have many interests :

- Change of coordinates
- In sample covariance matrix estimation, the intrinsic Cramer-Rao bound does not depend on the underlying covariance matrix (Smith 05[6] - homogeneous space):

$$\mathbb{E}_\Sigma[d^2(\hat{\Sigma},\Sigma)] = \mathbb{E}_\Sigma[d^2(\Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2}, I)] \geq \frac{n(n+1)}{k}$$

[6] S. T. Smith, Covariance, Subspace, and Intrinsic CramérRao Bounds, IEEE Transaction on Signal Processing, 2005

Two quotient geometries to compute means on $S^+(p, n)$

How to define a mean in $S_+(1, 2)$[7]

[7] Recall $S_+(p, n) = \{X \in \mathbb{R}^{n \times n} \mid X = X^T \succeq 0, \text{ rank X} = \text{p}\}$

How to define a mean in $S_+(1,2)$[7]

1. Density argument: for $\epsilon > 0$, $X_i + \epsilon I \succ 0 \Rightarrow$ use the mean in $P_+(2)$ and take the limit as $\epsilon \to 0$.
   Such means are NOT rank-preserving.

[7] Recall $S_+(p,n) = \{X \in \mathbb{R}^{n\times n} \mid X = X^T \succeq 0,\ \mathrm{rank\ X = p}\}$

How to define a mean in $S_+(1, 2)$[7]

2. Use of cartesian coordinates: $z \in \mathbb{R}^2$
   $X_i = z_i z_i^T$. Define $X_1 \circ X_2 = zz^T$ with $z = (z_1 + z_2)/2$.

[7] Recall $S_+(p, n) = \{X \in \mathbb{R}^{n\times n} \mid X = X^T \succeq 0, \text{ rank X} = \text{p}\}$

How to define a mean in $S_+(1,2)$[7]

3. Use of polar coordinates: $z = ru$
   $X_i = u_i r_i^2 u_i^T$ with

   $$u_i = (\cos\theta_i, \sin\theta_i) \quad \text{and} \quad r_i = \| z_i \|$$

   Define $X_1 \circ X_2 = ur^2u^T$ with

   $$\theta = \theta_1 + 0.5(\theta_2 - \theta_1) \quad \text{and} \quad r^2 = r_1 r_2$$

---

[7] Recall $S_+(p,n) = \{X \in \mathbb{R}^{n\times n} \mid X = X^T \succeq 0, \ \text{rank X} = \text{p}\}$

Extending the concept to $S_+(p, n)$

Generalize the previous idea using the matrix factorization:

$$X = ZZ^T = UR^2U^T$$

square-root decomposition → $Z \in \mathbb{R}_*^{n \times p}$

polar decomposition → $(U, R^2) \in St(p, n) \times P_+(p)$

$U^TU = I_p \qquad R^2 = R^{2T} \succ 0$

Extending the concept to $S_+(p, n)$

Generalize the previous idea using the matrix factorization:

$$X = ZZ^T = UR^2U^T$$

square-root decomposition

$$Z \in \mathbb{R}_*^{n \times p}$$

polar decomposition

$$(U, R^2) \in St(p, n) \times P_+(p)$$

$$U^T U = I_p \qquad R^2 = R^{2T} \succ 0$$

But if $O \in O(p)$, Z and ZO represent the same matrix.

Two quotient geometries for $S_+(p, n)$:

$$S_+(p, n) \approx \mathbb{R}_*^{n \times p} / O(p) \qquad S_+(p, n) \approx (St(p, n) \times P_+(p)) / O(p)$$

$$Z \approx ZO \qquad (U, R^2) \approx (UO, O^T R^2 O)$$

Polar geometry : rank-preserving geometric means on $S_+(p, n)$

Prof. Bhatia reminded us yesterday of Pascal's quotation stating that man is a geometric mean: “Qu'est-ce que l'homme dans la nature ? Un néant à l'égard de l'infini, un tout à l'égard du néant, un milieu entre rien et tout"

According to the fundamental and axiomatic approach of Ando, a geometric mean enjoys the following properties

(P1) Consistency with scalars: if X_1, X_2 commute $M(X_1, X_2) = (X_1 X_2)^{1/2}$.

(P2) Joint homogeneity

$$M(\alpha X_1, \beta X_2) = (\alpha\beta)^{1/2} M(X_1, X_2).$$

(P3) Permutation invariance $M(X_1, X_2) = M(X_2, X_1)$.

(P4) Monotonicity. If $X_1 \leq X_1'$ (i.e. $(X_1' - X_1)$ is a positive matrix) and $X_2 \leq X_2'$, the means are comparable and verify $M(X_1, X_2) \leq M(X_1', X_2')$.

(P5) Continuity from above. If $\{X_1(n)\}$ and $\{X_2(n)\}$ are monotonic decreasing sequence (in the Lowner matrix ordering) converging to X_1, X_2 then $\lim M(X_1(n), X_2(n)) = M(X_1, X_2)$.

(P6) Congruence invariance. For any $G \in \mathrm{Gl(n)}$ we have $M(GX_1G^T, GX_2G^T) = GM(X_1, X_2)G^T$.

(P7) Self-duality $M(X_1, X_2)^{-1} = M(X_1^{-1}, X_2^{-1})$.

Is there a rank preserving mean satisfying all the properties ?

We try to find a mean on $S^+(p,n)$, the set of positive semidefinite matrices of rank p which verifies properties P1-P7.

A few remarks:

- (P4) Monotonicity. If $X_1, X_2 \in S^+(p,n)$, X_1 and X_2 are comparable only if they have the same range.
- (P7) Self-duality. Matrices of $S^+(p,n)$ are not invertible. Inversion must be replaced by pseudo-inversion.
- (P6) We are going to prove one can **not** find a mean

$$M : S^+(p,n) \times S^+(p,n) \to S^+(p,n)$$

which verifies P6 if the rank p is small.

A preliminary result: P6 must be relaxed !

Proposition: Congruence invariance (transformer equation) implies the geometric mean of two matrices of $S^+(p, n)$ is almost surely null if $p < n/2$.

Proof.

- If Q is any matrix, continuity and monotonicity of the mean imply
$$M(QAQ^T, QBQ^T) \geq QM(X_1, X_2)Q^T$$
If Q is the orthoprojector on $Ker(X_1)$ we have $M(QAQ^T, QBQ^T) = 0$.
- as soon as p is small enough, the intersection of the ranges is almost surely null

□

Conclusion: the extension by continuity is moot for applications involving low rank.

A preliminary result: conclusion

Congruence invariance (P6) must be relaxed. However, invariance properties make sense in applications. We propose to replace P6-P7 with

(P6') Invariance to scalings and rotations[8] . For $(\mu, P) \in \mathbb{R}_+^* \times \mathrm{O(n)}$ we have $(\mu P^T X_1 \mu P) \circ (\mu P^T X_2 \mu P) = \mu P^T (X_1 \circ X_2) \mu P$.

(P7') Self-duality $M(X_1, X_2)^\dagger = M(X_1^\dagger, X_2^\dagger)$, where $\dagger$ is the pseudo-inversion.

[8] Remark : invariance to scalings is mandatory for a geometric mean (P2). Invariance to rotations is strongly desirable. Without it, for instance in Diffusion Tensor Imaging, the filtering process would depend on the orientation of the laboratory axes (e.g. the first vector is pointing North or East) !! (see last section for links with filtering).

Construction of the mean 1

In the polar geometry, a matrix X of $S^+(p, n)$ writes

$$X = UR^2U^T$$

where the columns of $U \in \mathbb{R}^{p\times n}$ form a p dimensional orthonormal basis of the span of X, and $R^2 \in P_+(p)$.

The intuition lies in the fact that matrices of $S^+(p, n)$ can be viewed as flat ellipsoids where

- U defines the space in which the ellipsoid lives (the range of X).
- R^2 defines the form of the ellipsoid (contains the information on the eigenvalues, i.e. the principal axes of the ellipsoid).

Construction of the mean 2

A good starting point is to consider the subset of rank p projectors:

$$\{P \in \mathbb{R}^{n\times n} /\ P^T = P,\ P^2 = P,\ \text{trace } P = p\}, \qquad (1)$$

which is in bijection with the Grassman manifold of p-dimensional subspaces. Such matrices belong to $S^+(p, n)$.[9]

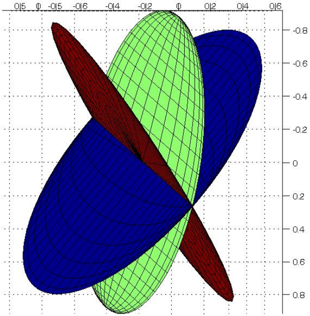
[9] A few definitions are needed

- Grassman manifold $Gr(p, n)$ is the set of p-dimensional subspaces in total space of dimension n. The Grassman squared distance between two subspaces is the sum of the squared principal angles between the subspaces.
- $St(p, n)$ is the Stiefel manifold of p-dimensional orthonormal bases of p-dimensional subspaces in space of dimension n.
- if $U \in \mathbb{R}^{n\times p}$ is such a p-dimensional basis $U^T U = I_p$. If $O \in O(p)$ then UO is a basis of the same subspace. We have thus the identification

$$Gr(p, n) \approx St(p, n)/O(p)$$

Geometric mean of rank p projectors[10]

- The Riemannian mean in the Grassman manifold is rotation-invariant and rank-preserving.
- Consider N projectors. The dominant eigenspace of dimension p of $\sum P_i$ is an approximation of the Grassman mean (see Sarlette and Sepulchre 2007).



[10] Any rank p projector writes $P = UU^T$ with $U \in St(p, n)$. It is a matrix of $S^+(p, n)$ with all eigenvalues $= 1$. It can be viewed as a flat disk.

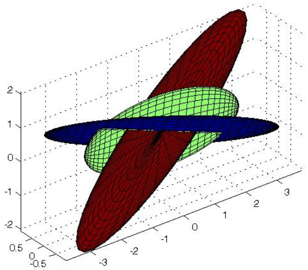
Geometric mean of two matrices

Suppose we have a geometric mean $M_{P_+(n)}$ on the set of full rank positive definite matrices and we want to extend it to $S^+(p,n)$. Let $X_1 = U_1 R_1^2 U_1^T$ and $X_2 = U_2 R_2^2 U_2^T$. A tempting generalization is

$$M_{S^+(p,n)}(X_1, X_2) = UR^2 U^T$$

where

$$UU^T = M_{Gr(p,n)}(U_1 U_1^T, U_2 U_2^T), \qquad R^2 = M_{P_+(p)}(R_1^2, R_2^2)$$



It is
RANK PRESERVING

Geometric mean of two matrices

With the proposed definition the calculation of the mean decouples into two separate problems

- calculation of the mean of the ranges $U_1 U_1^T$ and $U_2 U_2^T$ in $Gr(p, n)$
- calculation of the full rank geometric mean $M_{P_+(p)}$ of the factors R_1^2 and R_2^2.

BUT

- $U_1 R_1^2 U_1^T = (U_1 O)(O^T R_1^2 O)(U_1 O)^T$
- the Grassman mean is unchanged by the transformation $U_1 \mapsto U_1 O$
- whereas $M_{P_+(p)}(O^T R_1^2 O, R_2^2) \neq M_{P_+(p)}(R_1^2, R_2^2)$!!

Conclusion: The representatives U_1, R_1^2, U_2, R_2^2 must be carefully chosen.

Construction of the mean 3

Proposed algorithm

Consider N matrices $X_1, \cdots, X_N$ of $S^+(p,n)$ and their factorizations $X_i = U_i R_i^2 U_i^T$.

1- Compute the Riemannian mean (geodesic midpoint) in the Grassman manifold of the projectors $U_1 U_1^T, \cdots, U_N U_N^T$. Let the mean projector (or subspace) be WW^T.

2- Map all the flat ellipsoids A_i to the mean subspace WW^T via a rotation “of minimal energy"[11].

3- Compute any full-rank geometric mean $M_{P_+(p)}$ of the rotated ellipsoids.

[11] i.e., the rotation that maps the range of X_1 to the range of X_2 and that is the closest to identity. Its calculation can be avoided by picking good representatives U_1 and U_2. See at the end of the talk for more details

Properties of the mean

Let $M_{P_+(p)}$ be any geometric mean on $P_+(p)$.

Proposition 1:

- On the subset of rank p projectors, the proposed mean coincides with the Grassman Riemannian mean.
- On the other hand, when the rank p matrices all have the same range, the proposed mean coincides with the geometric mean induced by $M_{P_+(p)}$ on the common range subspace of dimension p.

Proposition 2: Consider N matrices. Suppose[12] the principal angles between the ranges are all $< \pi/2$. If $M_{P_+(p)}$ satisfies properties P1-P7 the proposed mean is well-defined, rank-preserving, and it deserves the appellation “geometric" as it satisfies the properties P1-P5, P6’-P7’.

[12] the assumption is satisfied in the generic case

Can it be viewed as a Riemannian mean ?

In the rank 1 case the mean is

$$ur^2u^T, \quad \text{where} \quad \theta = \theta_1 + \frac{\theta_2 - \theta_1}{2}, \; r = \sqrt{r_1 r_2}$$

It is the Riemannian mean for metric

$$ds^2 = k d\theta^2 + (r^{-1}\, dr)^2$$

In $S^+(p, n)$ we generalize it with metric (being vague)[13]

$$\text{“}ds^2 = k \sum_1^p d\theta_i^2 + \mathrm{Tr}\big((R^{-1}\, dR)^2\big)\text{”}$$

Proposition 3: The proposed mean (with $M_{P_+(p)}$ the Ando mean) tends to the Riemannian mean for the metric above when $k \to \infty$.

[13] Bonnabel, Sepulchre, SIMAX, 2009.

Some filtering illustrations

Filtering a constant rank p positive semi-definite matrix

In continuous-time, a first-order filter meant to filter a constant noisy signal $y(t)$ writes

$$\tau \frac{d}{dt} x = -x + y$$

In discrete-time, using a semi-implicit numerical scheme

$$x_{i+1} = \frac{y_i dt + \tau x_i}{dt + \tau}$$

which is a weighted mean. On $S^+(1,2)$ it becomes

$$u_{i+1} = (\cos(\frac{\tau\theta_i + \theta_{Y_i} dt}{dt + \tau}), \sin(\frac{\tau\theta_i + \theta_{Y_i} dt}{dt + \tau}))$$

$$r_{i+1}^2 = \exp(\frac{\tau \log r_i^2 + dt \log s_i^2}{dt + \tau})$$

filtering a constant rank p positive semi-definite matrix

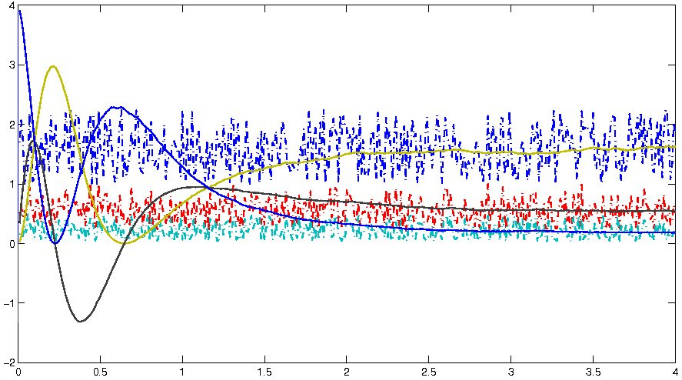


Figure: Filtering on $S^+(1,2)$: plot of the 3 coefficients of the measured matrix (dashed line) and the filtered matrix (plain line) with a 50% measurement noise, and $\tau = 50dt$. The filter allows to denoise the measured rank-1 symmetric matrix.

Filtering a constant rank p positive semi-definite matrix

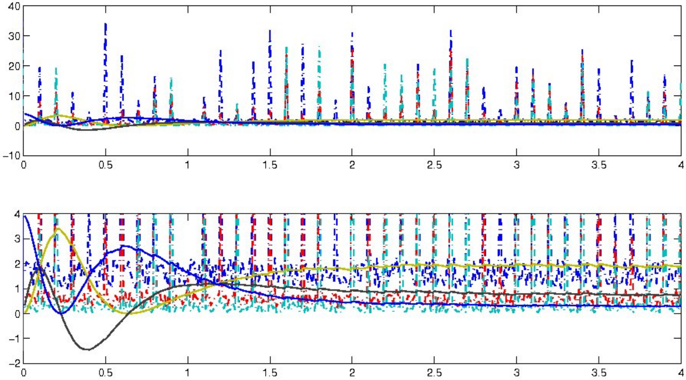


Figure: same setting but an outlier is inserted at each 10 steps. Below : zoom on the filtered signals (to be compared to the previous figure). The filter shows good robustness properties to outliers.

Anisotropic diffusion on Grassman [14]

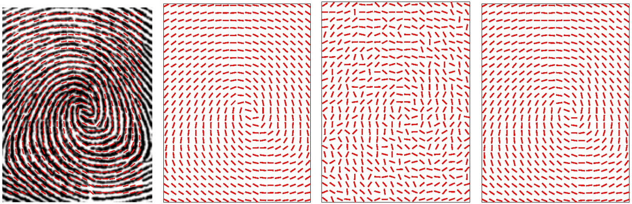


Figure: Implementation of a Perona-Malik filter on $S^+(1,2)$ field. (a): A fingerprint and its ridge directions. (b): The ridge directions of (a). (c): A Gaussian noise was added to the original image (b). (d): Regularized image.

The discrete perona-malik equation is

$$I_{i,j}^{t+1} = I_{i,j}^{t} + \lambda(c_N \bigtriangledown_N I + c_S \bigtriangledown_S I + c_E \bigtriangledown_E I + c_W \bigtriangledown_W I)|_{i,j}^{t}$$

with c a function of the gradient which is high for low gradient.

[14] See Y. Gur and N. Sochen. Regularizing flows over Lie groups. 2009.

Anisotropic diffusion: another illustration (just an)

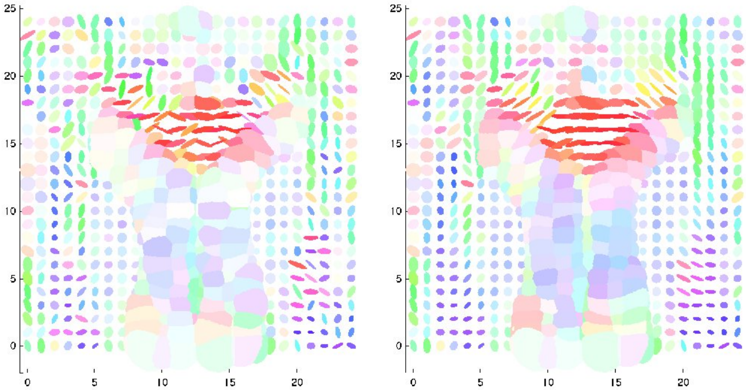


Figure: Perona-Malik filter on a Diffusion Tensor Image using means on $S^+(2,3)$. Image courtesy of the Cyclotron Research Centre of the University of Liège. (left noisy, right filtered)

Anisotropic diffusion: another illustration (just an)

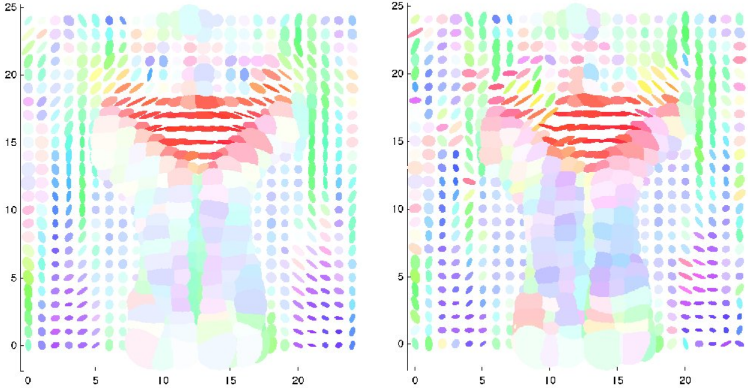


Figure: Perona-Malik filter on a Diffusion Tensor Image using means on $S^+(2,3)$. Image courtesy of the Cyclotron Research Centre of the University of Liège. (left true image, right filtered)

Conclusion

Conclusion

- Positive semidefinite matrices enjoy important quotient geometries rooted in Cartesian and Polar decomposition of matrices.
- The proposed mean rooted in the polar quotient geometry inherits from desirable properties of the geometric mean on the cone and it is rank-preserving.
- Low computation cost: the mean is calculated via a compact SVD (np^2 operations) and a mean on the cone.
- Maybe a way to extend other functions (means (harmonic), metrics, divergences, entropies $\cdots$) to the boundary of the cone ?

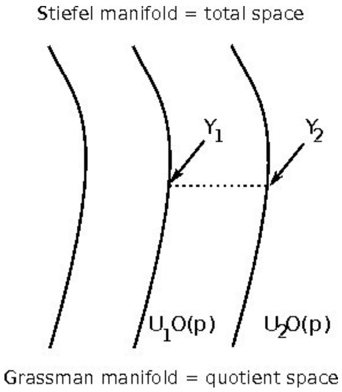
- For more results on the use of the cartesian coordinates see G. Meyer, S. Bonnabel, R. Sepulchre, JMLR, 2011.
- A comprehensive and insightful book on positive definite matrices and geometric means is R. Bhatia's book.

How to compute the minimal energy rotation ?

How to compute the “minimal energy" rotation ?

Recall $Gr(p,n) = St(p,n)/O(p)$. What are the bases of the subspaces $U_1O(p)$ and $U_2O(p)$ which are linked by a minimal energy rotation ? They are (Y_1, Y_2) s.t.

$$d_{\mathrm{St(n,p)}}(Y_1, Y_2) = \min_{(O_1,O_2)\in O(p)} d_{\mathrm{St(n,p)}}(U_1O_1, U_2O_2) \qquad (2)$$



Mean mathematical definition of two matrices

- Let Y_1, Y_2 be bases of the ranges of X_1 and X_2 linked by a minimal energy rotation.
- Let $Y_1^T X_1 Y_1 = Q_1^2$ and $Y_2^T X_2 Y_2 = Q_2^2$ be the corresponding representatives on $P_+(p)$.
- Let Y be the Riemannian mean of the bases in $St(p, n)$. With Y_1, Y_2 defined this way Y is a basis of the Grassman mean subspace.

The rotated ellipsoids[15] merely write

$$YQ_1^2 Y^T, \quad YQ_2^2 Y^T$$

The proposed mean is

$$Y \, M_{P_+(p)}(Q_1, Q_2) \, Y^T$$

[15] by a rotation of minimal energy

Illustration for several matrices

Let $X_1 = U_1 R_1^2 U_1^T, \cdots, X_N = U_N R_N^2 U_N^T$.

- the pairs (W_i, Y_i) are linked by a minimal energy rotation
- the rotated ellispoids are $W_i Q_i W_i^T$ where $A_i = Y_i Q_i Y_i^T$
- express them in a common basis W_0, i.e. $W_i Q_i W_i^T = W_0 T_i W_0^T$

The mean is

$$W_0 [M_{P_+(p)}(T_1^2, \cdots, T_N^2)] W_0^T$$