Riemannian geometry and geometric means of fixed-rank positive semi-definite matrices

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

Mines ParisTech
silvere.bonnabel@mines-paristech.fr

Workshop on Matrix Geometries and Applications
The Abdus Salam International Centre for Theoretical Physics
Trieste, July 10th, 2013

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 (complexity and storage):

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

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,

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

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]:

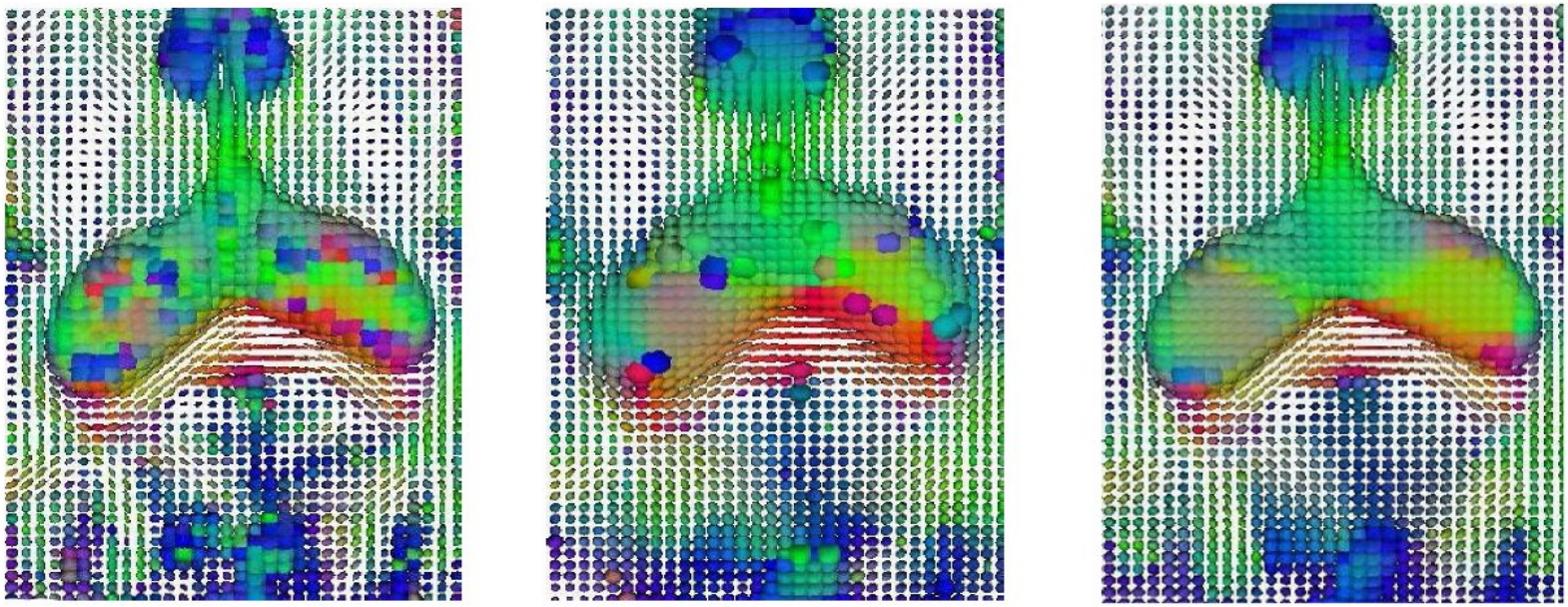


Figure: Original image “Vectorial" filtering “Riemannian" filtering

- Does geometry matter??

[1] Courtesy of 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]:

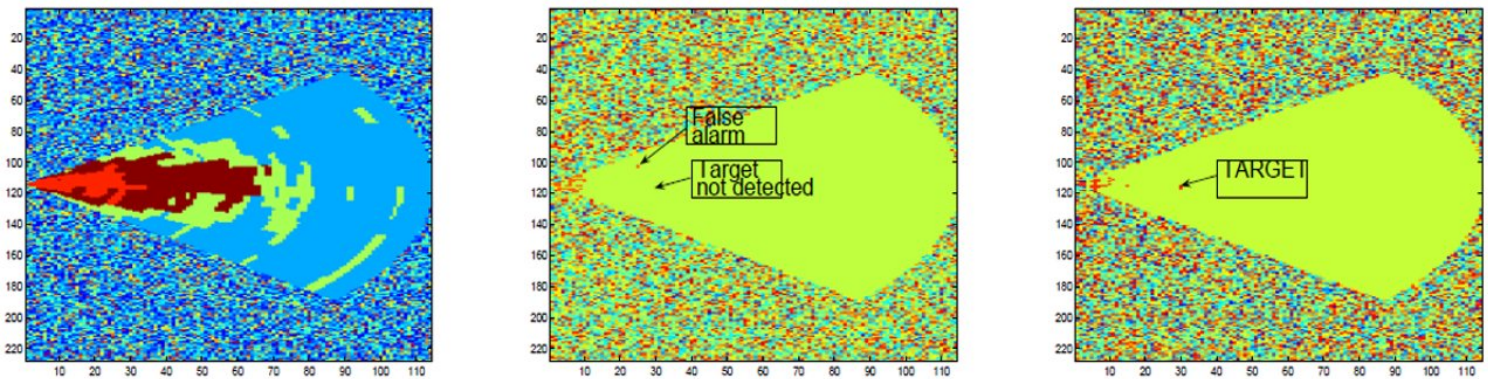


Figure: Doppler Radar Fourier filtering “Riemannian" filtering

[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

1. Recalling the positive definite case
2. A Riemannian metric for the semidefinite fixed rank case
3. A rank preserving geometric mean for the semidefinite fixed rank case
4. Some applications
   - Image classification via Mahalanobis distance learning
   - Low rank Kalman filters for data assimilation in oceanography

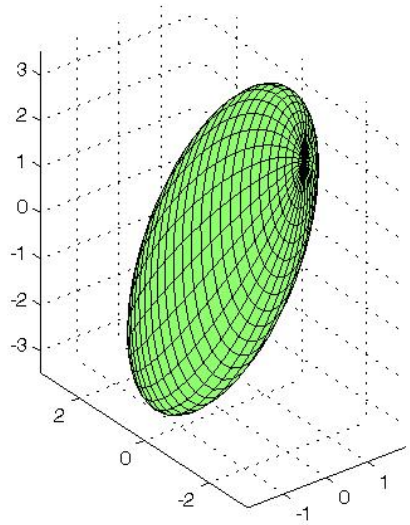
A basic and common issue in the cone

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

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

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, Li and Mathias, Moakher, Petz and Temesi, Barbaresco, Bhatia, Lim, Holbrook, Bini, Palfia, Lawson...

Properties of the natural metric

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

  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?

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 Cramer Rao Bounds, IEEE Transaction on Signal Processing, 2005

Outline

1. Recalling the positive definite case
2. A Riemannian metric for the semidefinite fixed rank case
3. A rank preserving geometric mean for the semidefinite fixed rank case
4. Some applications

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

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 $X_1 \circ X_2 = u r^2 u^T$ with

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

midpoint for the metric $ds^2 = d\theta^2 + (dr/r)^2$

[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:

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

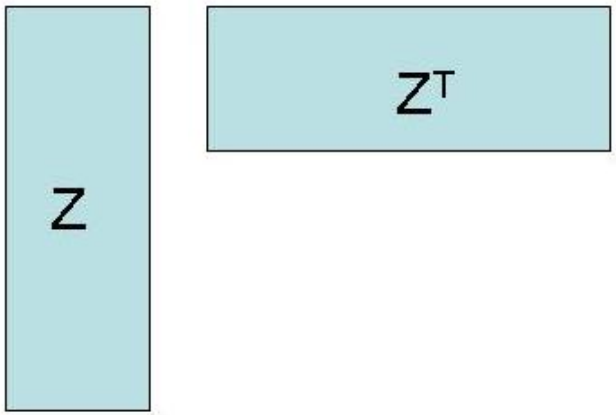
square-root decomposition

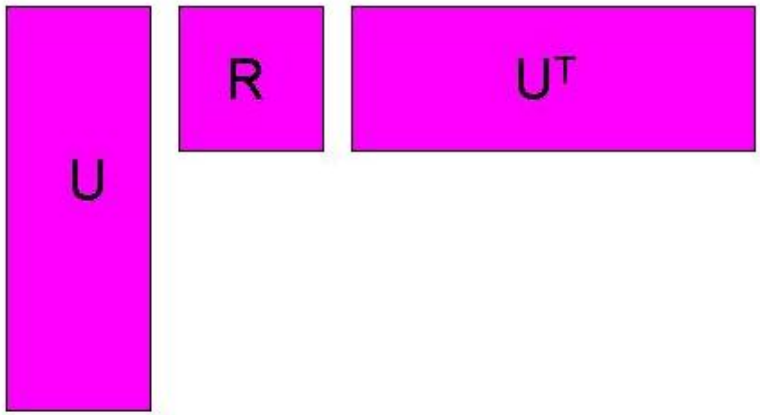
$$Z \in \mathbb{R}_*^{n \times p}$$

polar decomposition

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

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





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 $S_+(p, n)$:

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

$$Z \approx ZO$$

$$S_+(p, n) \approx (St(p, n) \times P_+(p))/O(p)$$

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

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

Let us focus on the **polar decomposition**

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

where U is $n \times p$ with orthonormal columns and $R^2 \in P^+(p)$.

1. UU^T is a projector onto a subspace of dimension p
2. R^2 is a (small) positive definite matrix

Grassman distance$\rightarrow$ natural distance between subspaces (principal angles)

Natural metric of $P_+(p)\rightarrow$ natural distance between positive definite matrices

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

Let us focus on the polar decomposition

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

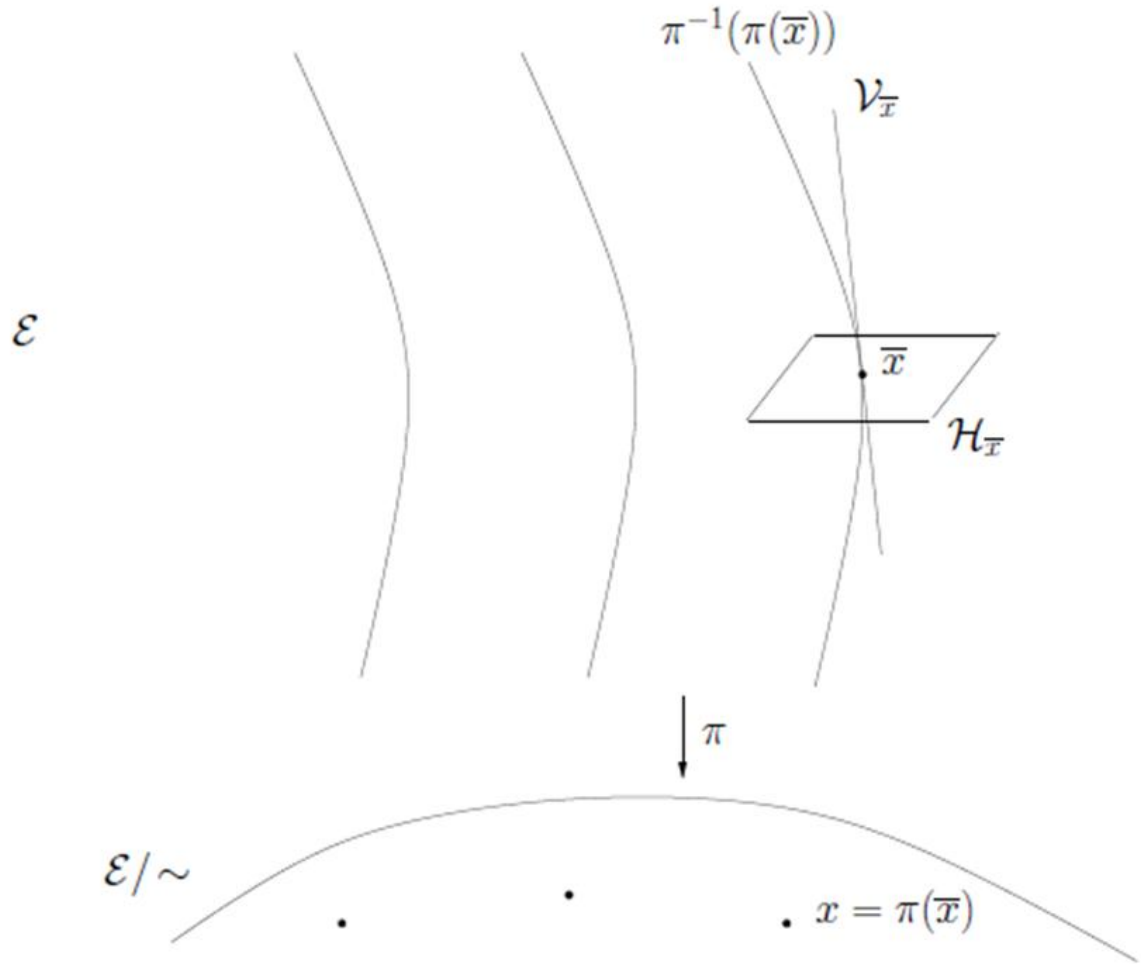
A sensible metric: for close enough $X_1 = U_1R_1^2U_1$ and $X_2 = U_2R_2^2U_2$ we propose the infinitesimal distance:

$$\begin{aligned}\delta(X_1, X_2)^2 &= \delta(U_1R_1^2U_1, U_2R_2^2U_2)^2 \\ &\simeq d^2_{Grassman}(U_1U_1^T, U_2U_2^T)^2 + d^2_{P_+(p)}(R_1^2, R_2^2)\end{aligned}$$

Wait ! $S_+(p,n)$ is a quotient manifold. The metric not well-defined on the *quotient*. Indeed:

$$X_1 = U_1 R_1^2 U_1^T = U_1 O (O^T R_1^2 O)(U_1 O)^T \quad \text{for} \quad O \in O(p)$$

$U_1 U_1^T$ unchanged **BUT** $d_{P_+(p)}(R_1, R_2) \neq d_{P_+(p)}(O^T R_1^2 O, R_2^2)$



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

Let us focus on the polar decomposition

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

Proposed metric: for close enough $X_1 = U_1R_1^2U_1$ and $X_2 = U_2R_2^2U_2$ we propose the infinitesimal distance:

$$\begin{aligned}\delta(X_1,X_2)^2 &= \delta(U_1R_1^2U_1, U_2R_2^2U_2)^2 \\ &\simeq d^2_{Grassman}(U_1U_1^T, U_2U_2^T) + d^2_{P_+(p)}(R_1^2, R_2^2)\end{aligned}$$

with well-chosen representatives (U_1, R_1^2) and (U_2, R_2^2) !

Proper Riemannian metric formulation with well-chosen tangent vectors at U and R^2

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

Theorem 1 The space $S_+(p,n) \cong (St(p,n) \times P_+(p))/O(p)$ endowed with the proposed metric is a well-defined (quotient) Riemannian manifold. Moreover it is geodesically complete.

Theorem 2 Furthermore, the metric is invariant with respect to orthogonal transformations, scalings, and pseudo-inversion.

Conclusion: the metric retains some desirable invariance properties of the natural metric of $P_+(n)$.[8]

[8] S. Bonnabel and R. Sepulchre. Riemannian metric and geometric mean for positive semidefinite matrices of fixed rank. SIAM J. Matrix Anal. Appl. 2009.

Outline

1. Recalling the positive definite case
2. A Riemannian metric for the semidefinite fixed rank case
3. A rank preserving geometric mean for the semidefinite fixed rank case
4. Some applications

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 $\{X_1(k)\}$ and $\{X_2(k)\}$ are monotonic decreasing sequence (in the Lowner matrix ordering) converging to X_1, X_2 then $\lim M(X_1(k), X_2(k)) = M(X_1, X_2)$.

(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 !

Result: 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

$$M(QX_1Q^T, QX_2Q^T) \geq QM(X_1,X_2)Q^T$$

  If Q is the orthoprojector on $Ker(X_1)$ we have $M(QX_1Q^T, QX_2Q^T) = 0$.

- as soon as p is small enough, the intersection of the subspaces supporting X_1 and X_2 is almost surely null

□

A preliminary result: relax P6

Does it matter ??

Proposition: Let R_ϵ be a rotation arbitrarily close to identity. When the rank is low it is always possible to choose it s.t.

$$M(X, R_\epsilon X R_\epsilon^T) = 0 \quad \text{(far from } X \text{ !)}$$

Filtering: averaging is usually a simple way to filter out noise, because of noise cancellations in the sum.

The noise can not be filtered out by averaging without destroying all the relevant information in the signal !

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[9] . 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.

[9] Remark : invariance to scalings is mandatory for a geometric mean (P2). Invariance to rotations is physically meaningful, e.g. north vs west.

Construction of the mean

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

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

where the columns of $U \in \mathbb{R}^{n\times p}$ 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 supporting the ellipsoid lives
- 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

Consider N matrices $X_1 = U_1 R_1^2 U_1^T, \cdots X_N = U_N R_N^2 U_N^T$ of $S^+(p, n)$.

Proposition[10]: if $U_1 U_1^T, \cdots, U_N U_N^T$ belong to a geodesic ball of radius $\pi/4$ in the Grassman manifold, their Karcher mean exists and is unique. Let U_{mean} be a basis of the mean subspace.

Definition 1: if $X_1 = U_1 U_1^T$, $X_2 = U_2 U_2^T$ ... that is,

$$R_1 = R_2 = \cdots = R_N = I_p$$

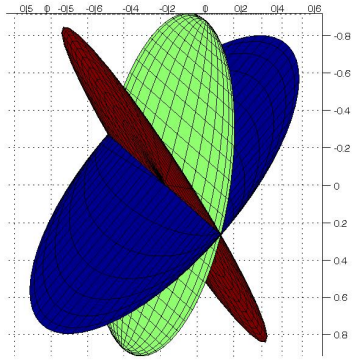
we propose as a rank preserving geometric mean

$$X_{mean} = U_{mean} U_{mean}^T$$

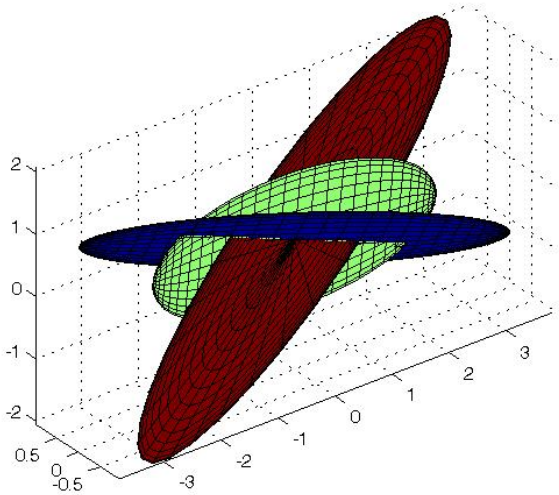
[10] Applying a result of Afsari.

Geometric intuition for the mean

Simple case: $R_1 = R_2 = I_p$
the mean ellipsoid spans the mean subspace and has p-isotropic form



General case: $R_1, R_2 \in P_+(p)$
both ellipsoids are rotated into the mean subspace via a minimal rotation and then averaged using a full rank geometric mean on $P_+(p)$



Intuitive definition of the mean

Consider N matrices

$$X_1 = U_1 R_1^2 U_1^T, \cdots, X_N = U_N R_N^2 U_N^T \in S^+(p, n)$$

Proposition: Let U_{mean} be a basis of the mean subspace. The compact SVD of $U_i^T U_{mean}$ yields two bases V_i and Y_i such that

$$Y_i V_i^T X_i V_i Y_i^T$$

is the ellipsoid X_i rotated into the subspace spanned by U_{mean} by the rotation being the closest to identity I_n.

Definition [rank preserving geometric mean]: Rotate all ellipsoids to span(U_{mean}) with minimal rotations. Apply any p-full rank matrix geometric mean (Ando, Riemannian (ls) etc.) having properties $(P1) - (P7)$.

Results

Main result[11]: The proposed mean is

• rank preserving on $S_+(p, n)$

• satisfies properties $(P1) - (P5)$ and $(P6)' - (P7)'$ and thus deserves to be called "geometric"

Remark: Unfortunately not the Karcher (ls) mean in the sense of the proposed Riemannian metric.

[11] S. Bonnabel, A. Collard and R. Sepulchre. Rank preserving geometric means of positive semidefinite matrices. Linear Algebra and its Appl. 2013.

Outline

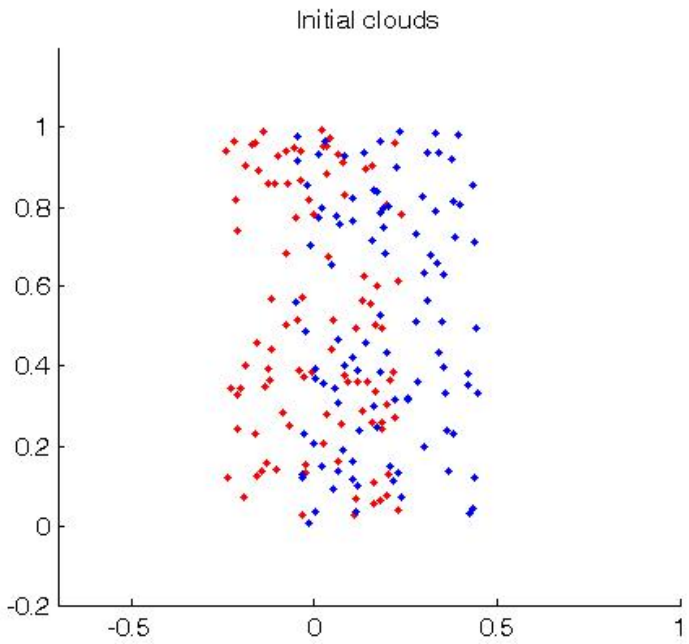
1. Recalling the positive definite case
2. A Riemannian metric for the semidefinite fixed rank case
3. A rank preserving geometric mean for the semidefinite fixed rank case
4. Some applications
   - Image classification via Mahalanobis distance learning
   - Low rank Kalman filters for data assimilation in oceanography

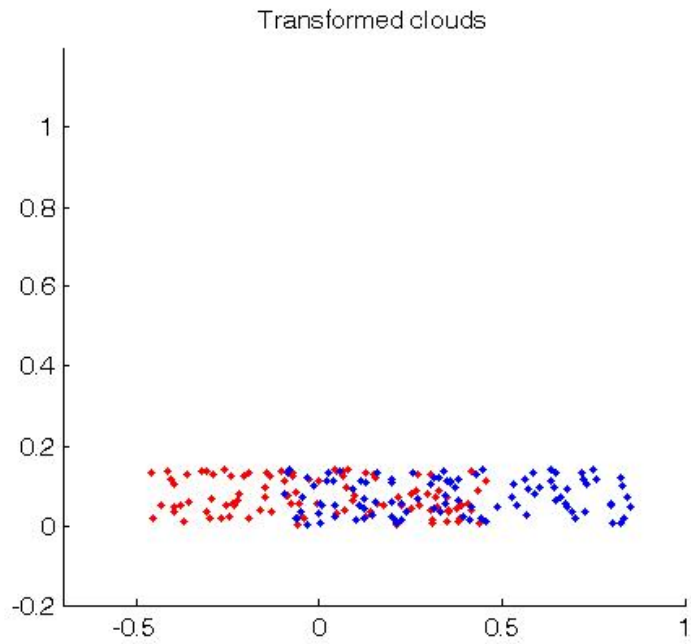
Data mining and classification

Mahalanobis distance: parameterized by a positive semidefinite matrix W

$$d_W^2(x_i, x_j) = (x_i - x_j)^T W (x_i - x_j)$$

Machine learning: Let $W = GG^T$. Then d_W^2 simple Euclidian squared distance for transformed data $\tilde{x}_i = Gx_i$.

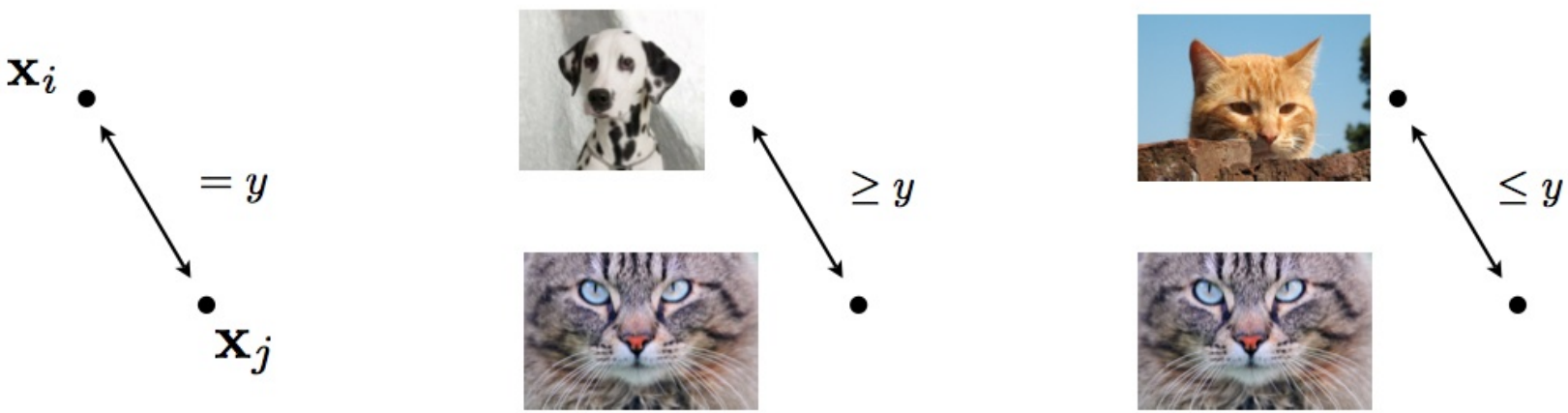




Mahalanobis distance learning

Goal: integrate new constraints to an existing W

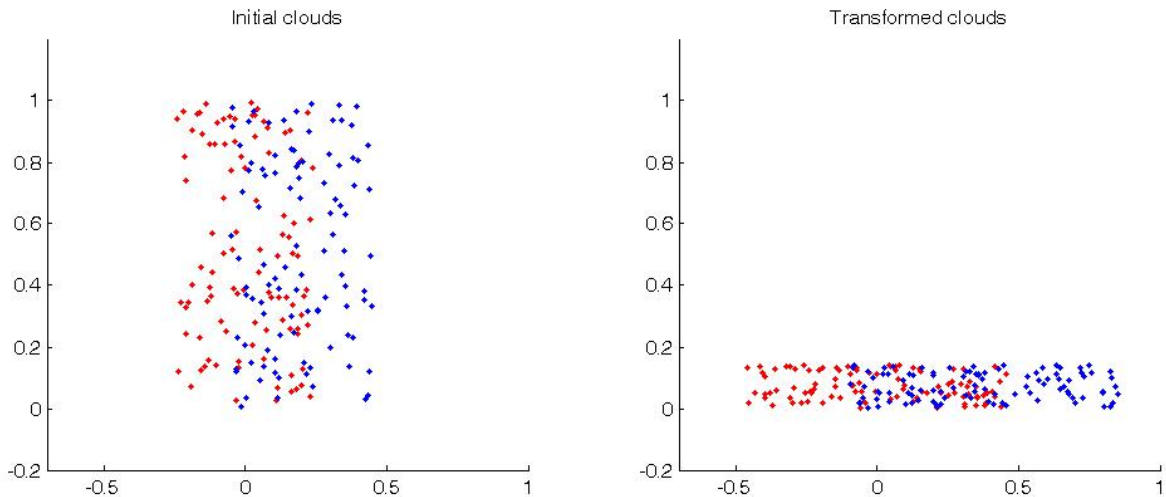
- equality constraints: $d_W(x_i, x_j) = y$
- similarity constraints: $d_W(x_i, x_j) \leq y$
- dissimilarity constraints: $d_W(x_i, x_j) \geq y$



Computational cost significantly reduced when W is low rank !

Low rank Mahalanobis distance

One could have projected everything on a horizontal axis ! For large datasets low rank allows to derive algorithm with linear complexity in the data space dimension d.



Method

Semi-definite positive matrices of fixed rank

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

Mathematical formulation: $y_t = d(x_i, x_j)$, error: $(\hat{y} - y)^2$ where $\hat{y}$ is the predicted distance with current W.

Simple scalable method: perform gradient descent on the prediction error $(\hat{y} - y)^2$

Problem: $W - \gamma_t \nabla_W((\hat{y} - y)^2)$ has NOT same rank as W.

Remedy: do Riemannian gradient descent on the manifold !

Outline

1. Recalling the positive definite case
2. A Riemannian metric for the semidefinite fixed rank case
3. A rank preserving geometric mean for the semidefinite fixed rank case
4. Some applications
   - Image classification via Mahalanobis distance learning
   - Low rank Kalman filters for data assimilation in oceanography

Oceanography problem: satellites measure height differences in the ocean $\rightarrow$ estimate currents (e.g. Gulf Stream) ?

Kalman filter: in the presence of measurement uncertainty compute the most probable state (currents) along with covariance matrix $P \in P_+(n)$ of the error through Riccati eq.

Numerical complexity: updating P requires $O(n^2)$ operations with n up to 10 millions (model based on PDE's) $\rightarrow$ work with low rank approximations leading to $O(pn)$ complexity.

Riccati equation is a contraction for the natural metric of the cone[12] leading to desirable stability properties of the filter $\rightarrow$ some contraction properties are inherited by the low-rank Riccati equation in the sense of the polar metric proposed above.[13]

---

[12] P. Bougerol. Kalman filtering with random coefficients and contractions. SIAM J. Contr. Opt. 1993.

[13] S. Bonnabel and R. Sepulchre. The geometry of Low rank Kalman filters. Matrix Information Geometry. Springer. 2012.

Conclusion

In this talk: we have introduced a Riemannian structure on $S_+(p, n)$ and a rank preserving geometric mean.

Some open questions remain: e.g. closed form for geodesics, discrepancy between proposed geometric mean and Karcher (ls) mean in the sense of the proposed metric $\rightarrow$ both challenges could be tackled theoretically or numerically.

Interpolation the mean allows for interpolation/fusion between low-rank covariance matrices $\rightarrow$ there definitely should be some more applications to tackle.

Optimization: a convergence result for (stochastic) Riemannian gradient descent on $S_+(p, n)$ for Mahalanobis distance learning was recently obtained[14].

[14] S. Bonnabel. Stochastic gradient descent on Riemannian manifolds. IEEE Tran. on Autom. Contr. In press. 2013.