Differential geometrical approach to covariance estimation

Antoni Musolas

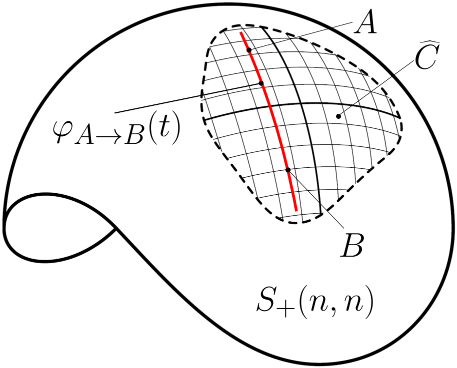
Center for Computational Engineering
Aerospace Computational Design Lab

Massachusetts Institute of Technology (MIT)
musolas@mit.edu

22 April 2016

Covariance estimation problem

- **Broad problem:** Given a parameterized family of covariances and some samples, what is the most representative member of the family?
- **Goals of the presentation:**
  1. Can we use a geodesic line between two symmetric positive definite matrices to define a covariance matrix family?
  2. Can we look at the problem of covariance estimation geometrically?



Geometry of the manifold of positive definite matrices

Let A_1 and A_2 belong to $S_+(n,n)$.

- There exists a distance that satisfies:

$$d(A_1, A_2) = d(A_1^{-1}, A_2^{-1}),$$

$$d(A_1, A_2) = d(ZA_1Z^T, ZA_2Z^T).$$

- Closed form expression for the distance:

$$d(A_1, A_2) = \sqrt{\sum_{k=1}^{n} \log^2(\lambda_k)},$$

where λ_k are the generalized eigenvalues of (A_1, A_2).

- A parametrization of the geodesic between A_1 and A_2 is given by:

$$\varphi_{A_1 \to A_2}(t) = A_1^{\frac{1}{2}} \exp_m(t \log_m(A_1^{-\frac{1}{2}} A_2 A_1^{-\frac{1}{2}})) A_1^{\frac{1}{2}} = A_1^{\frac{1}{2}} U \Lambda^t U^T A_1^{\frac{1}{2}},$$

where $\varphi_{A_1 \to A_2}(t) \in S_+(n,n)$ for all $t \in \mathbb{R}$, and $\Lambda = diag(\lambda_k)$.

Geodesic as covariance function

Definition (Covariance function)

A one-parameter covariance function is a one-parameter group $\varphi\colon \mathbb{R} \to S_+(n,n)$.

Lemma (Geodesic as covariance function)

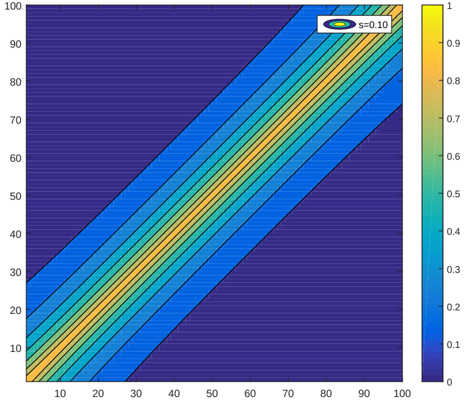
Let A_1 and A_2 be two elements in $S_+(n,n)$. Then $\varphi_{A_1 \to A_2}(t)$ is a one-parameter covariance function.

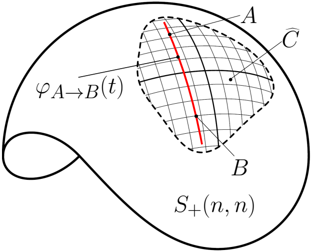
Two possible generalizations:

1. Let A_1 and A_2 be two elements in $S_+(n,r)$ for $r < n$.
2. Let φ to be $\varphi_{A_1 \to A_2}\colon \mathbb{R}^p \to S_+(n,r)$, for p-variate covariance function.

- **Idea:** Interpolation of covariance matrices through a geodesic.
- **Example:** A log-permeability field $Y(x, \omega)$ is defined as a Gaussian process with mean $\mu_Y = 1$ and covariance kernel.

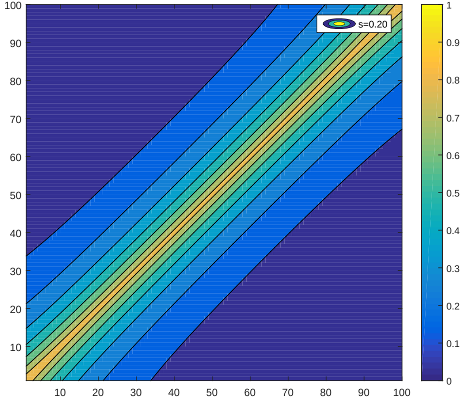
$$C(x, \bar{x}) = \sigma_Y^2 \exp\left(-\frac{1}{p}\left(\frac{|x-\bar{x}|}{L}\right)^p\right), \qquad L = 0.3,\ \sigma_Y^2 = 1,\ p = 1$$

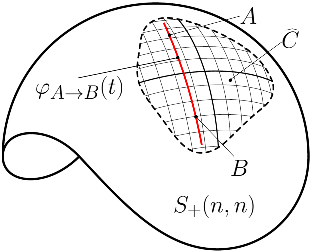




- **Idea:** Interpolation of covariance matrices through a geodesic.
- **Example:** A log-permeability field $Y(x,\omega)$ is defined as a Gaussian process with mean $\mu_Y = 1$ and covariance kernel.

$$C(x,\bar{x}) = \sigma_Y^2 \exp\left(-\frac{1}{p}\left(\frac{|x-\bar{x}|}{L}\right)^p\right), \qquad L = 0.3,\ \sigma_Y^2 = 1,\ p = 1$$

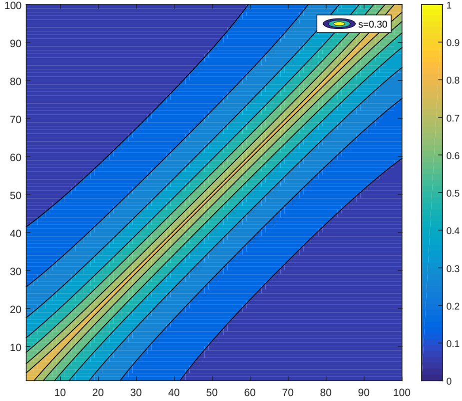


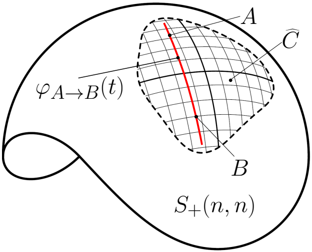


Geodesic as covariance function

- **Idea:** Interpolation of covariance matrices through a geodesic.
- **Example:** A log-permeability field $Y(x, \omega)$ is defined as a Gaussian process with mean $\mu_Y = 1$ and covariance kernel.

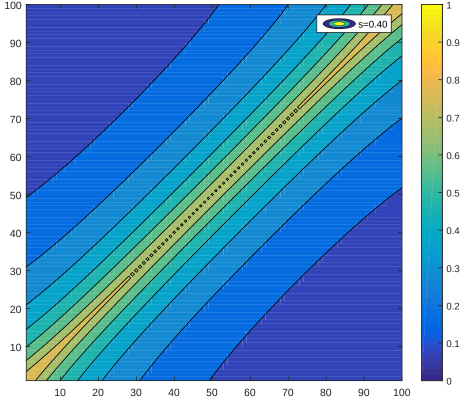
$$C(x, \bar{x}) = \sigma_Y^2 \exp\left(-\frac{1}{p}\left(\frac{|x-\bar{x}|}{L}\right)^p\right), \qquad L = 0.3,\ \sigma_Y^2 = 1,\ p = 1$$

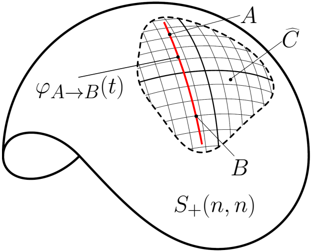




- **Idea:** Interpolation of covariance matrices through a geodesic.
- **Example:** A log-permeability field $Y(x,\omega)$ is defined as a Gaussian process with mean $\mu_Y = 1$ and covariance kernel.

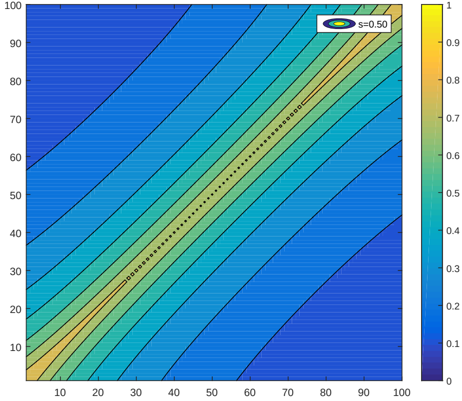
$$C(x,\bar{x}) = \sigma_Y^2 \exp\left(-\frac{1}{p}\left(\frac{|x-\bar{x}|}{L}\right)^p\right), \qquad L = 0.3,\ \sigma_Y^2 = 1,\ p = 1$$

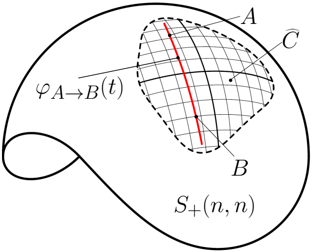




- **Idea:** Interpolation of covariance matrices through a geodesic.
- **Example:** A log-permeability field $Y(x,\omega)$ is defined as a Gaussian process with mean $\mu_Y = 1$ and covariance kernel.

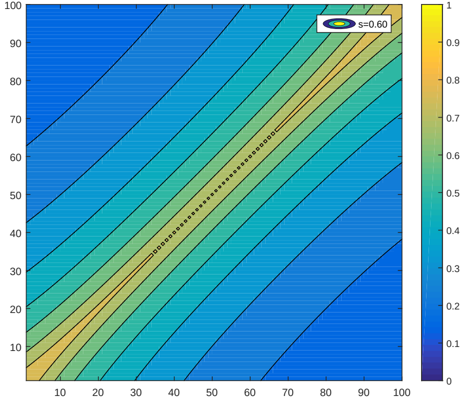
$$C(x,\bar{x}) = \sigma_Y^2 \exp\left(-\frac{1}{p}\left(\frac{|x-\bar{x}|}{L}\right)^p\right), \qquad L = 0.3,\ \sigma_Y^2 = 1,\ p = 1$$

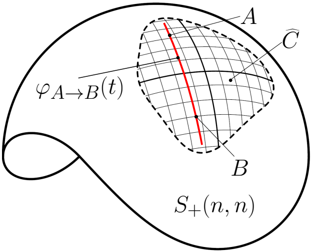




- **Idea:** Interpolation of covariance matrices through a geodesic.
- **Example:** A log-permeability field $Y(x,\omega)$ is defined as a Gaussian process with mean $\mu_Y = 1$ and covariance kernel.

$$C(x,\bar{x}) = \sigma_Y^2 \exp\left(-\frac{1}{p}\left(\frac{|x-\bar{x}|}{L}\right)^p\right), \qquad L = 0.3,\ \sigma_Y^2 = 1,\ p = 1$$

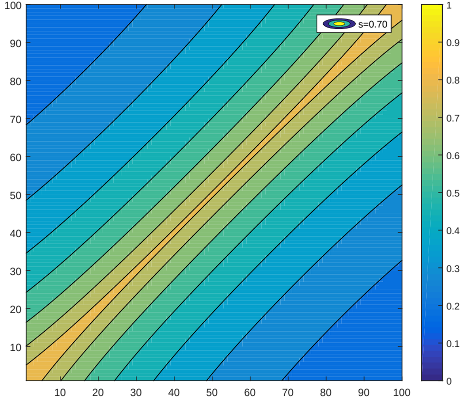


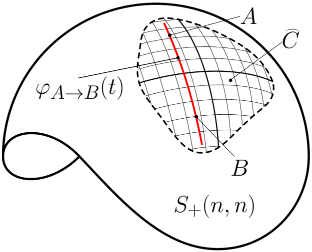


Geodesic as covariance function

- **Idea:** Interpolation of covariance matrices through a geodesic.
- **Example:** A log-permeability field $Y(x,\omega)$ is defined as a Gaussian process with mean $\mu_Y = 1$ and covariance kernel.

$$C(x,\bar{x}) = \sigma_Y^2 \exp\left(-\frac{1}{p}\left(\frac{|x-\bar{x}|}{L}\right)^p\right), \qquad L = 0.3,\ \sigma_Y^2 = 1,\ p = 1$$

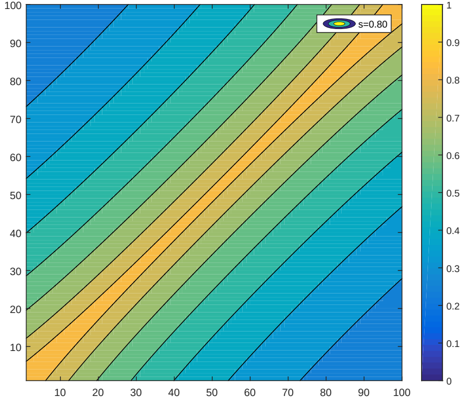


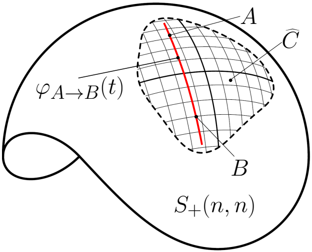


Geodesic as covariance function

- **Idea:** Interpolation of covariance matrices through a geodesic.
- **Example:** A log-permeability field $Y(x,\omega)$ is defined as a Gaussian process with mean $\mu_Y = 1$ and covariance kernel.

$$C(x,\bar{x}) = \sigma_Y^2 \exp\left(-\frac{1}{p}\left(\frac{|x-\bar{x}|}{L}\right)^p\right), \qquad L = 0.3,\ \sigma_Y^2 = 1,\ p = 1$$

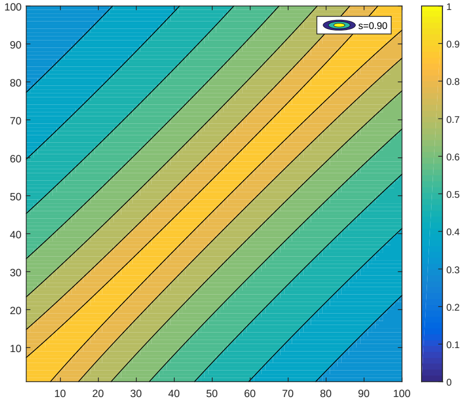


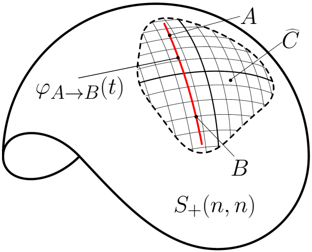


Geodesic as covariance function

- **Idea:** Interpolation of covariance matrices through a geodesic.
- **Example:** A log-permeability field $Y(x,\omega)$ is defined as a Gaussian process with mean $\mu_Y = 1$ and covariance kernel.

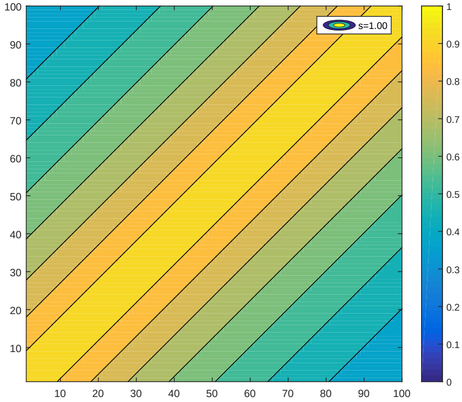
$$C(x,\bar{x}) = \sigma_Y^2 \exp\left(-\frac{1}{p}\left(\frac{|x-\bar{x}|}{L}\right)^p\right), \qquad L = 0.3,\ \sigma_Y^2 = 1,\ p = 1$$

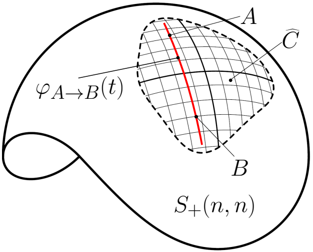




- **Idea:** Interpolation of covariance matrices through a geodesic.
- **Example:** A log-permeability field $Y(x,\omega)$ is defined as a Gaussian process with mean $\mu_Y = 1$ and covariance kernel.

$$C(x,\bar{x}) = \sigma_Y^2 \exp\left(-\frac{1}{p}\left(\frac{|x-\bar{x}|}{L}\right)^p\right), \qquad L = 0.3,\ \sigma_Y^2 = 1,\ p = 1$$





Duality of the covariance estimation problem

- Let $y^{(1)}, \ldots, y^{(q)}$ be observations from an n-variate normal dist.
- Let $\widehat{C}$ be a full rank sample covariance matrix of the $y^{(1)}, \ldots, y^{(q)}$.
- Consider two covariance matrices of interest, A and B, and $\varphi_{A\to B}(t)$.

Maximum likelihood approach to covariance estimation

$$\underset{t\in(-\infty,\infty)}{\text{maximize}} \;\; p_X(y^{(1)}, \ldots, y^{(q)} | t) \tag{1}$$

$$\text{s.t. } X \sim N(0, \varphi_{A\to B}(t))$$

Minimization of distance approach to covariance estimation

$$\underset{t\in(-\infty,\infty)}{\text{minimize}} \;\; d(\varphi_{A\to B}(t), \widehat{C}) \tag{2}$$

Divergence measures as spectral functions

Definition (Spectral function)

Let A_1 and A_2 be two elements in $S_+(n, n)$. A function $f(\lambda^{(A_1,A_2)})$ is a spectral function if it is a differentiable and symmetric map of the n generalized eigenvalues of (A_1, A_2) to the reals.

- **Examples of spectral functions**
  - Natural distance in $S_+(n, n)$:
$$d(A_1, A_2) = \sqrt{\sum_{k=1}^{n} \log^2(\lambda_k)}.$$
  - Kullback-Leibler divergence for multivariate normal:
$$D_{KL}(N(0, A_1)||N(0, A_2)) = \sum_{k=1}^{n} (\lambda_k + \log^2(\lambda_k) + 1)/2.$$
  - Hellinger distance for multivariate normal:
$$d_{Hell}(N(0, A_1), N(0, A_2)) = 1 - 2^{l/2} \prod_{k=1}^{n} \lambda_k^{1/4}(1 + \lambda_k)^{-1/2}.$$

Minimizing an spectral function

Lemma (Spectral function minimization)

Let f be a spectral function, then:

- *Minimizing $f(\lambda^{(\varphi_{A\to B}(t),\widehat{C})})$ over t is equivalent to finding t^+ such that:*

$$Tr\left(V(t^+)\left(\left.\frac{\delta f(\Sigma(t))}{\delta t}\right|_{t^+}\right)V(t^+)^T M\Lambda^{t^+}\log\Lambda M^T\right)=0.$$

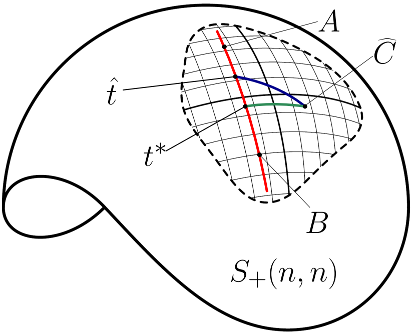
- *Notation:*
  - $X(t)=\widehat{C}^{-\frac{1}{2}}A_1^{\frac{1}{2}}U\Lambda^t U^T A_1^{\frac{1}{2}}\widehat{C}^{-\frac{1}{2}}$,
  - $X(t)=V(t)\Sigma(t)V(t)^T$, *a proper eigenvalue decomposition,*
  - $M=\widehat{C}^{-\frac{1}{2}}A^{\frac{1}{2}}U$.

Properties of the proposed optimization problems (I/II)

Lemma (Uniqueness of the solution)

The aforementioned problems are respectively concave and convex, thus:

1. *There exists a unique $\hat{t}$ that maximizes the likelihood $p_X(y^{(1)}, \ldots, y^{(q)}|t)$.*
2. *There exists a unique t^* that minimizes the distance $d(\varphi_{A \to B}(t), \widehat{C})$.*

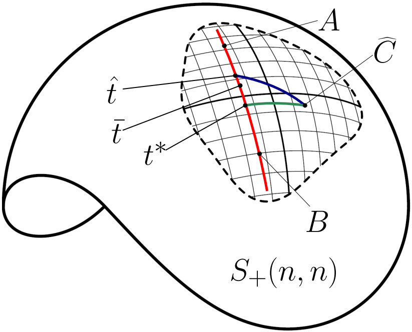


Properties of the proposed optimization problems (II/II)

Lemma (Idempotence of the projection)

If $\widehat{C} \in \varphi_{A \to B}(t)$, then:

1. *There exists a unique $\bar{t}$ such that either (i) $(\lambda_k^{(A,B)})^{\bar{t}} = \lambda_k^{(A,\widehat{C})}$, or (ii) $(\lambda_k^{(B,A)})^{\bar{t}} = \lambda_k^{(B,\widehat{C})}$, for all $k = 1, 2 \ldots n$.*
2. *Moreover, $\bar{t} = t^* = \hat{t}$ and $\widehat{C} = \varphi_{A \to B}(\bar{t})$.*



Solution to the minimization problem

Result 1 (Differential geometrical solution of covariance estimation)

1. If $\widehat{C} \in \varphi_{A \to B}(t)$, then:

$$t^* = \frac{\sum_{k=1}^{n} \log(\lambda_k^{\widehat{C}}) - \sum_{k=1}^{n} \log(\lambda_k^{A})}{\sum_{k=1}^{n} \log(\lambda_k^{B}) - \sum_{k=1}^{n} \log(\lambda_k^{A})},$$

   solves the minimization problem, where λ_k^A, λ_k^B, and $\lambda_k^{\widehat{C}}$, are the k-th eigenvalues of A, B, and $\widehat{C}$, respectively.

2. This expression also holds when $A = \alpha B$, for any positive real α.
3. Otherwise, t^* is the solution of:

$$Tr(\log_m(\Lambda^{t^*} \widehat{C}^{-\frac{1}{2}} A \widehat{C}^{-\frac{1}{2}}) \log_m(\Lambda)) = 0.$$

4. In all cases, the solution is unique.
5. The aforementioned t^* minimizes the Fisher information metric if data assumed to be normally distributed with known mean.

Solution to the maximization problem

Result 2 (Maximum likelihood solution of covariance estimation)

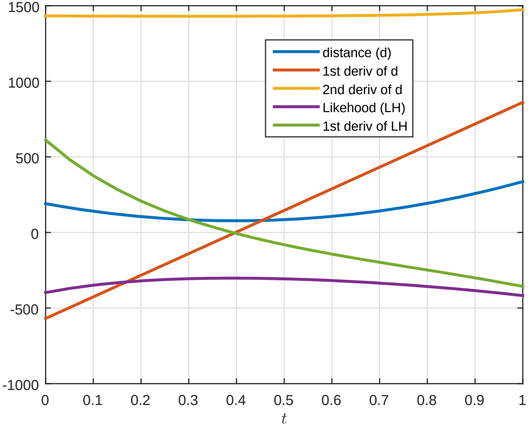
1. Refer to the preceding. If $\widehat{C} \in \varphi_{A \to B}(t)$, then the solution in Result 1 continues to hold and $\hat{t} = t^*$.
2. Otherwise, $\hat{t}$ is the solution of:

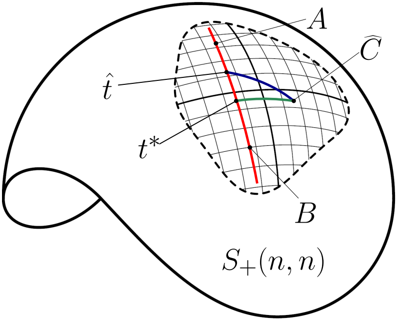
$$Tr(\widehat{C} A^{-\frac{1}{2}} U \Lambda^{-\hat{t}} \log_m(\Lambda) U^T A^{-\frac{1}{2}} - \log_m(\Lambda)) = 0.$$

3. In all cases, the solution is unique.
4. The aforementioned $\hat{t}$ minimizes the Kullback-Leibler divergence if data assumed to be normally distributed with known mean.

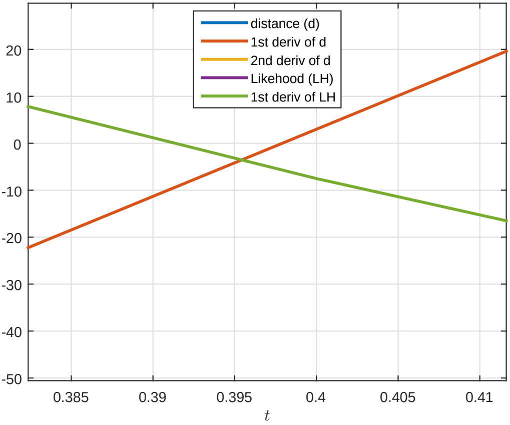
Results in a toy problem

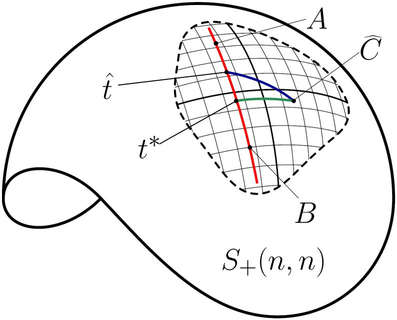
- Illustration of the cost functions in the maximization and minimization problems in a toy example:





- Illustration of the cost functions in the maximization and minimization problems in a toy example:





Pros and cons of our approach

- **Advantages of using geodesic as covariance function**
  - Possibility to use empirical covariance matrices to define richer parametric families of covariance functions.
  - Covariances offer more flexibility for problem-specific tailoring than classical parametric families of covariance kernels.
  - Works properly as a non-stationary covariance kernel.
- **Advantages of minimizing distance vs maximizing likelihood**
  - Do not require to specify a distribution for the data.
  - Minimizing distance is the natural way in differential geometry.
  - It also minimizes Fisher information metric, which is an intrinsic property in inference.
- **Disadvantages**
  - Impossibility to recover the covariance generating kernel.
  - The covariance matrix must be full rank.
  - We require prior knowledge of the problem.

Conclusions and limitations

- **Contributions**
  - Devised a covariance function that follows naturally from the data.
  - Proposed a differential geometrical approach to covariance estimation.
- **Limitations**
  - Computational cost is of the same order than maximizing the likelihood.
  - Our covariance function is already discretized, as opposed to the classic covariance kernels.
- **Further research**
  - Devise a multi-variate covariance function.
  - Generalize for low rank covariance matrices.
  - Compute error bounds.