Estimation and classification of location and covariance matrix using Riemannian geometry: application to remote sensing

Antoine Collas

Postdoc in the Mind team at Inria Saclay, supervised by Alexandre Gramfort (Meta).

PhD done at SONDRA lab, CentraleSupélec, Univ. Paris-Saclay in 2019-2022.

Table of contents

Some context around remote sensing

Context

In recent years, many image time series have been taken from the **earth** with different technologies: **SAR, multi/hyper spectral imaging**, ...

Objective

To segment semantically these data using **sensor diversity** (spectral bands, polarization...), and **spatial** and/or **temporal** informations.

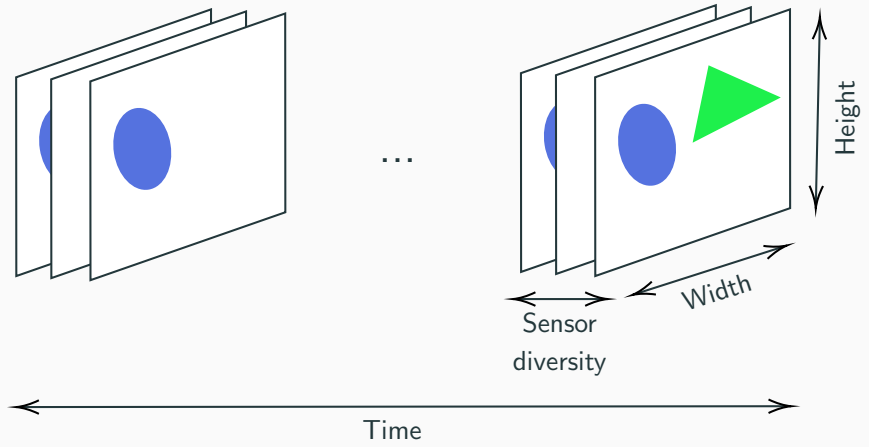


Figure 1: Multivariate image time series.

Applications

Activity monitoring, land cover mapping, crop type mapping, disaster assessment ...

Example of multi-spectral time series

Breizhcrops dataset[1]:

- more than 600 000 crop time series across the whole Brittany,
- 13 spectral bands, 9 classes.

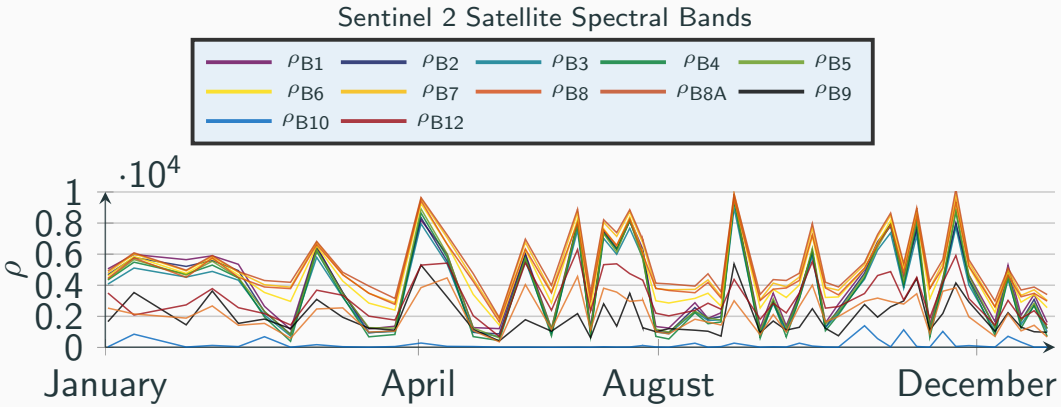


Figure 2: Reflectances ρ of a time series of **meadows**.

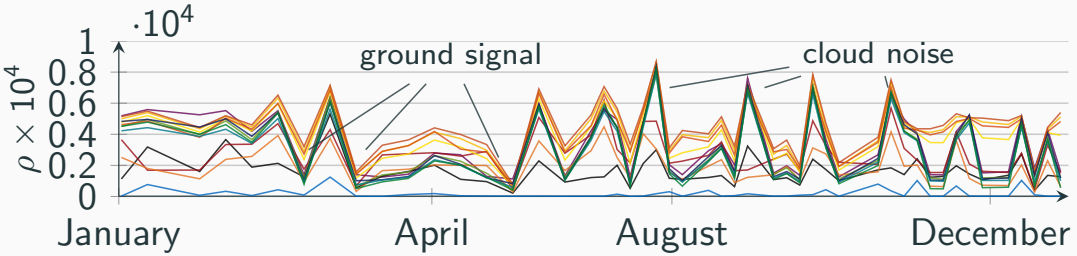


Figure 3: Reflectances ρ of a time series of **corn**.

[1] https://breizhcrops.org/

Example of a hyperspectral image

Indian pines dataset:

- 145×145 pixels, 200 spectral bands,
- 16 classes (corn, grass, wood, ...).

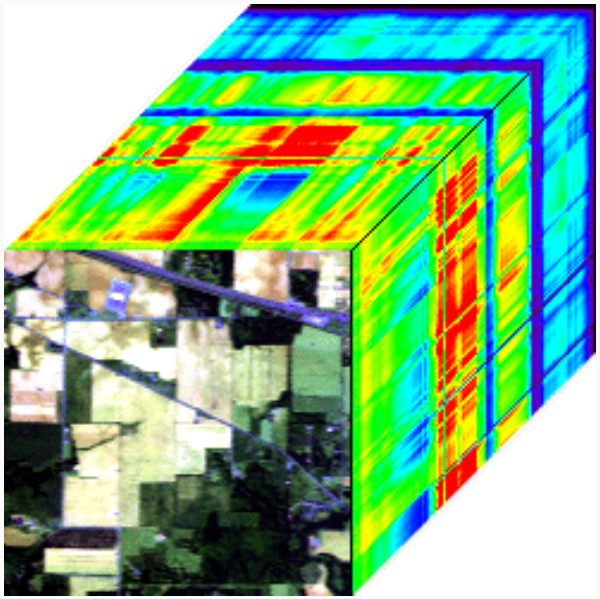


Figure 4: Raw image.

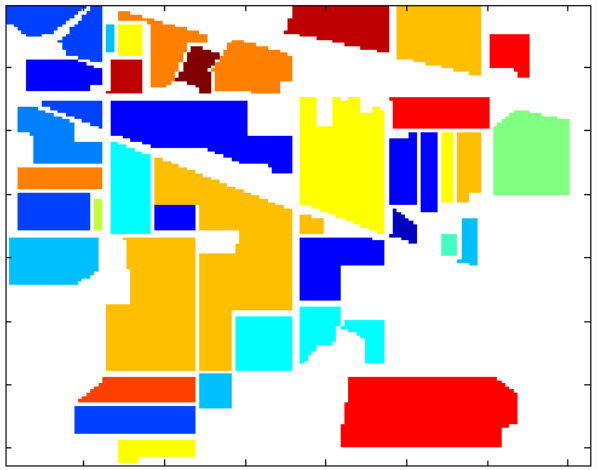


Figure 5: Ground truth, one color = one class.

Clustering/classification pipeline

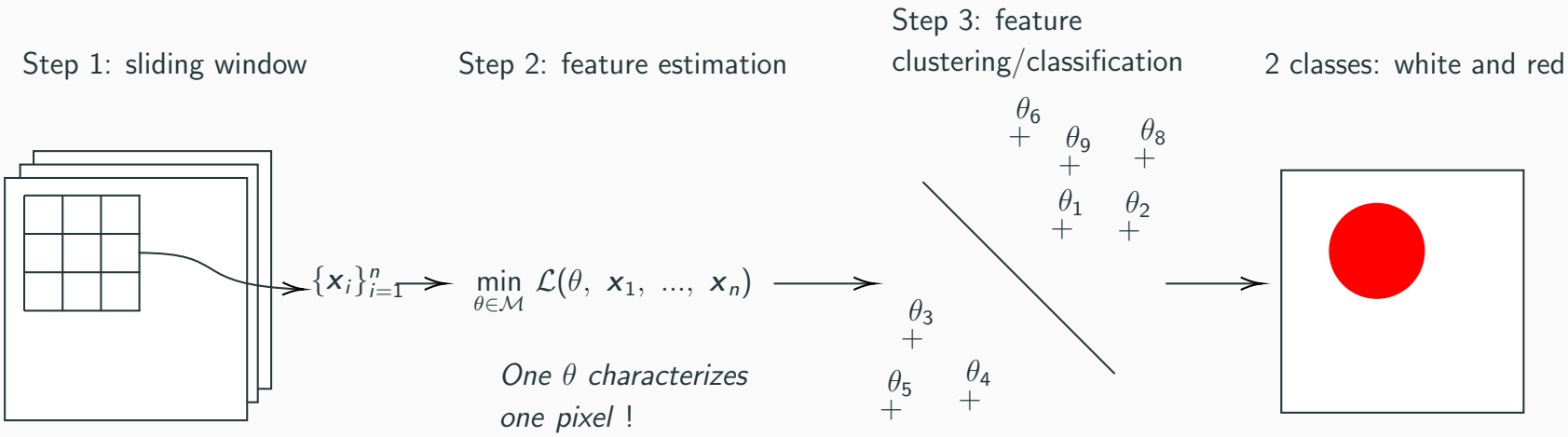


Figure 6: Clustering/classification pipeline.

Assumption:

$$\boldsymbol{x} \sim f(.;\theta), \text{ a probability density function, } \theta \in \mathcal{M}$$

Examples of θ:

$\theta = \boldsymbol{\Sigma}$ a covariance matrix, $\theta = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ a vector and a covariance matrix,
$\theta = (\{\tau_i\}, \boldsymbol{U})$ a scalar and an orthogonal matrix...

$\mathcal{M}$ can be constrained !

Step 2: objectives for feature estimation

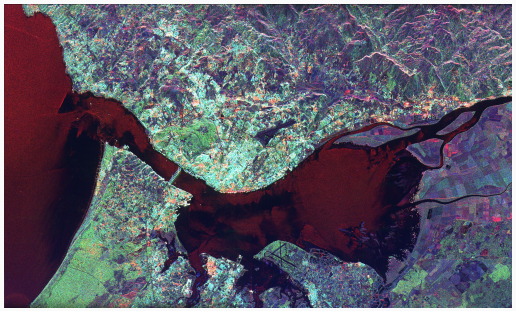


Figure 7: Example of a SAR image (from `nasa.gov`).

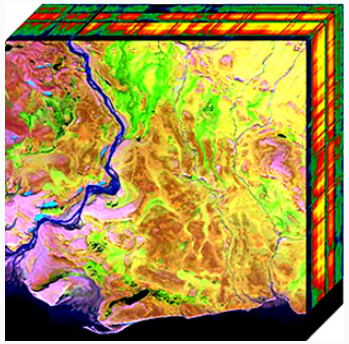


Figure 8: Example of a hyperspectral image (from `nasa.gov`).

Objectives:

- develop **robust estimators**, *i.e.* estimators that work well with non Gaussian or heterogeneous data because of the high resolution of images,
- develop **regularized/structured estimators**, *i.e.* estimators that handle the high dimension of hyperspectral images.

Step 3: objectives for clustering/classification

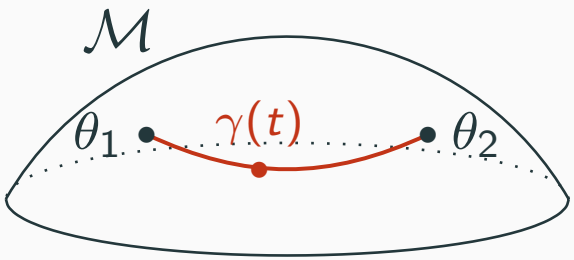


Figure 9: Divergence δ_γ: squared length of the curve γ.

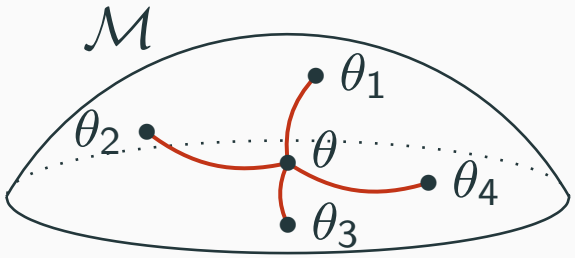


Figure 10: Center of mass of $\{\theta_i\}_{i=1}^M$.

Objectives:

develop divergences that

- **respect the constraints** of $\mathcal{M}$,
- are **related to the chosen statistical distributions**,
- are robust to **mislabelization** and **distribution shifts** between train and test sets.

Clustering/classification pipeline and Riemannian geometry

Random variable: $\boldsymbol{x} \sim f(.;\theta)$, $\theta \in \mathcal{M}$

Step 2: maximum likelihood estimation

$$\underset{\theta \in \mathcal{M}}{\text{minimize}}\ \mathcal{L}(\theta | \{\boldsymbol{x}_i\}_{i=1}^n) = -\log f(\{\boldsymbol{x}_i\}_{i=1}^n; \theta)$$

Step 3: given δ, center of mass of $\{\theta_i\}_{i=1}^M$

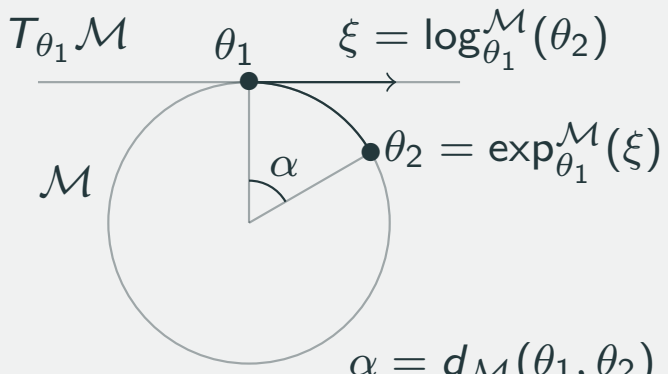
$$\underset{\theta \in \mathcal{M}}{\text{minimize}} \sum_i \delta(\theta, \theta_i)$$

Use of Riemannian geometry:

- optimization under constraints,
- "Fisher information metric" $\implies$ a canonical Riemannian manifold for the parameter space $\mathcal{M}$ (fast estimators, intrinsic Carmér-Rao bounds...),
- δ: squared Riemannian distance.

Riemannian geometry and problematics

What is a Riemannian manifold ?



Curvature induced by:

- constraints, *e.g.* the sphere: $\|\boldsymbol{x}\| = 1$,
- Riemannian metric, *e.g.* on $\mathcal{S}_p^{++}$: $\langle \boldsymbol{\xi}, \boldsymbol{\eta} \rangle_{\boldsymbol{\Sigma}}^{\mathcal{S}_p^{++}} = \mathrm{Tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\xi}\boldsymbol{\Sigma}^{-1}\boldsymbol{\eta})$.

Some geometric tools:

- **tangent space** $T_\theta\mathcal{M}$ (vector space): linearization of $\mathcal{M}$ at $\theta \in \mathcal{M}$,
- **Riemannian metric** $\langle ., . \rangle_\theta^{\mathcal{M}}$: inner product on $T_\theta\mathcal{M}$,
- **geodesic** γ: curve on $\mathcal{M}$ with zero acceleration,
- **exponential map**: $\exp_\theta^{\mathcal{M}}(\xi) = \gamma(1)$ with $\gamma(0) = \theta$, $\dot{\gamma}(0) = \xi$,
- **logarithmic map**: $\log_{\theta_1}^{\mathcal{M}}(\theta_2) = \xi$ with $\exp_{\theta_1}^{\mathcal{M}}(\xi) = \theta_2$,
- **distance**: $d_{\mathcal{M}}(\theta_1, \theta_2) = \left\| \log_{\theta_1}^{\mathcal{M}}(\theta_2) \right\|_{\theta_1}^{\mathcal{M}}$.

N. Boumal, "An introduction to optimization on smooth manifolds"

What is a Riemannian manifold ?

Examples of $\mathcal{M}$

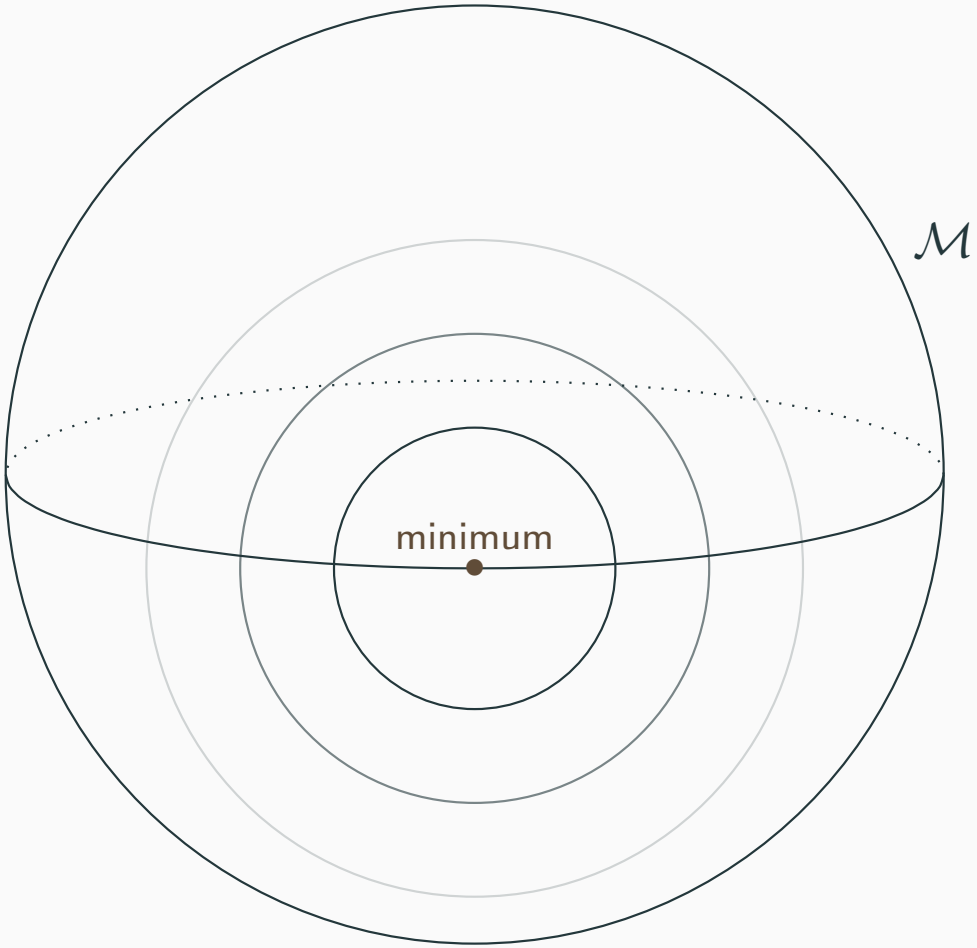
- linear spaces: $\mathbb{R}^{p\times k}$, $\mathcal{S}_p = \{\boldsymbol{X} \in \mathbb{R}^{p\times p} : \boldsymbol{X}^T = \boldsymbol{X}\}$,
- norm constraints: $\mathcal{S}^{p^2-1} = \{\boldsymbol{X} \in \mathbb{R}^{p\times p} : \|\boldsymbol{X}\|_F = 1\}$,
- positivity constraints: $\mathcal{S}_p^{++} = \{\boldsymbol{\Sigma} \in \mathcal{S}_p : \forall \boldsymbol{x} \neq \boldsymbol{0} \in \mathbb{R}^p,\ \boldsymbol{x}^T\boldsymbol{\Sigma}\boldsymbol{x} > 0\}$,
- positivity and scale constraints: $\mathcal{SS}_p^{++} = \{\boldsymbol{\Sigma} \in \mathcal{S}_p^{++} : |\boldsymbol{\Sigma}| = 1\}$,
- orthogonality constraints: $\mathrm{St}_{p,k} = \{\boldsymbol{U} \in \mathbb{R}^{p\times k} : \boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}_k\}$,
- quotient spaces: $\mathrm{Gr}_{p,k} = \{\{\boldsymbol{UO} : \boldsymbol{O} \in \mathcal{O}_k\} : \boldsymbol{U} \in \mathrm{St}_{p,k}\}$.

Optimization

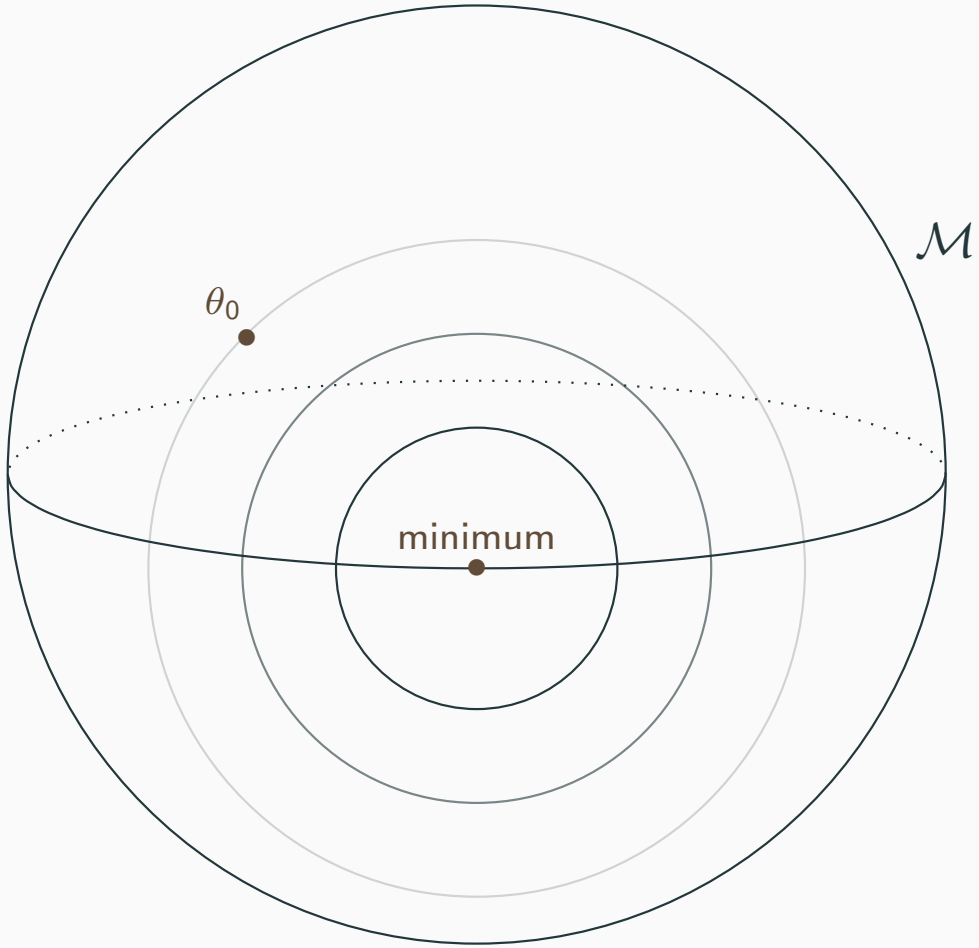
$f : \mathcal{M} \to \mathbb{R}$, smooth

$$\underset{\theta\in\mathcal{M}}{\text{minimize}}\ f(\theta)$$

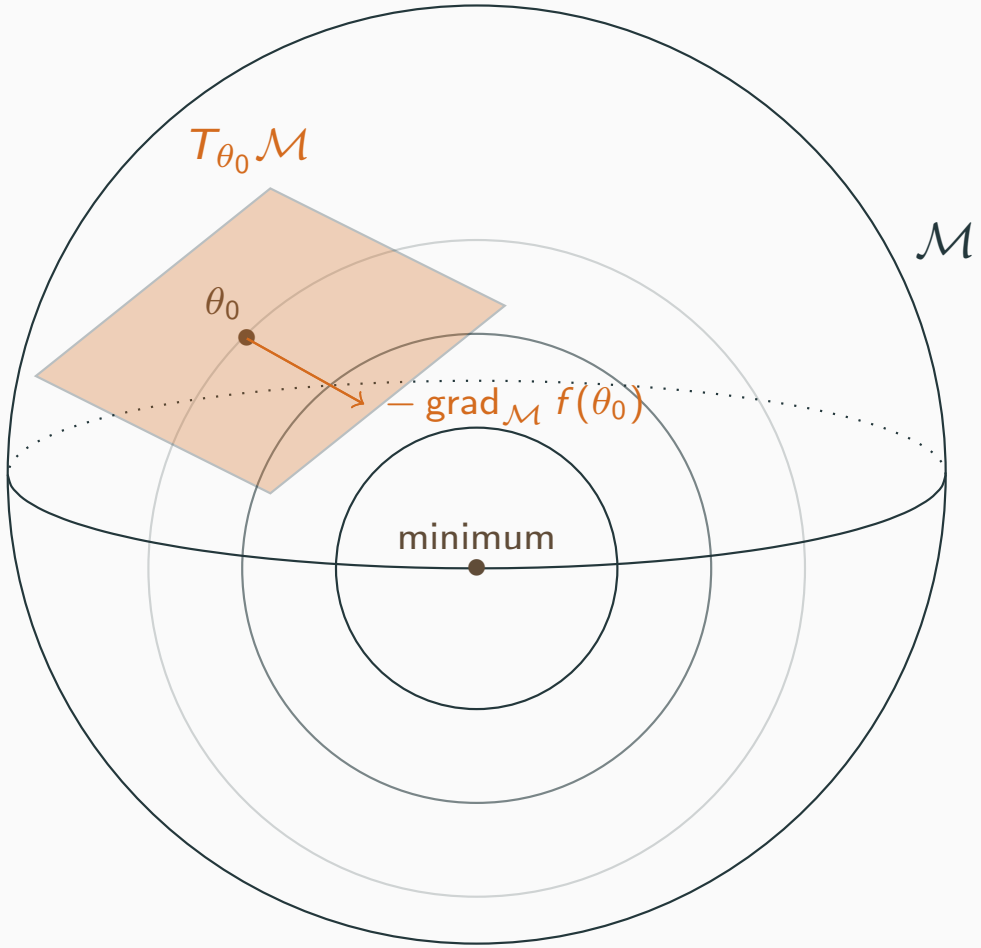
Optimization on a manifold



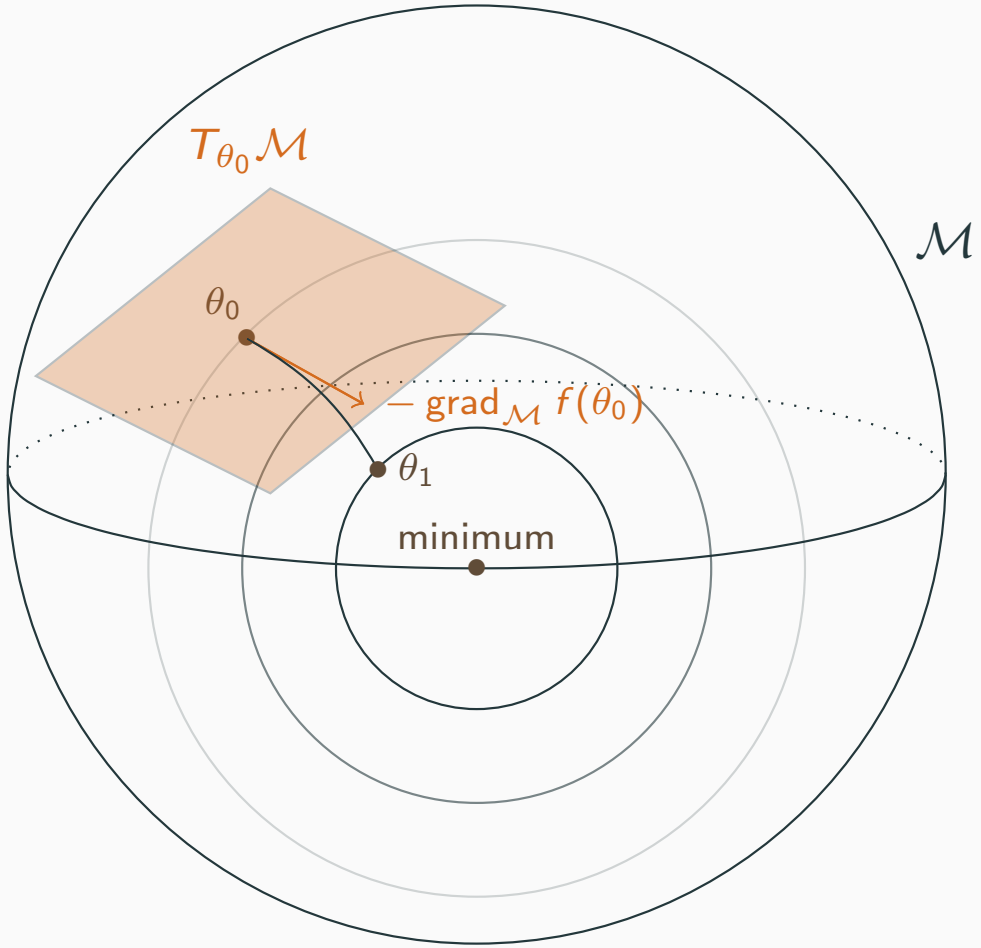
Optimization on a manifold



Optimization on a manifold



Optimization on a manifold



Fisher information metric

Random variable, negative log-likelihood

$$\boldsymbol{x} \sim f(.;\theta), \quad \theta \in \mathcal{M}$$

$$\mathcal{L}(\theta|\boldsymbol{x}) = -\log f(\boldsymbol{x};\theta)$$

Fisher information metric

$$\begin{aligned}\langle \xi, \eta \rangle_\theta^{\mathsf{FIM}} &= \mathbb{E}_{\boldsymbol{x}\sim f(.;\theta)}\left[\mathrm{D}^2 \mathcal{L}(\theta|\boldsymbol{x})[\xi,\eta]\right] \\ &= \mathrm{vec}(\xi)^T I(\theta)\mathrm{vec}(\eta)\end{aligned}$$

where

$$I(\theta) = \mathbb{E}_{\boldsymbol{x}\sim f(.;\theta)}\left[\mathrm{Hess}\,\mathcal{L}(\theta|\boldsymbol{x})\right] \in \mathcal{S}_p^{++}$$

is the Fisher information matrix.

(Set of constraints, Fisher information metric) = a Riemannian manifold

Existing work: centered Gaussian

A well known geometry:

$$\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}), \quad \boldsymbol{\Sigma} \in \mathcal{S}_p^{++}$$

with the Fisher information metric:

$$\langle \boldsymbol{\xi}, \boldsymbol{\eta} \rangle_{\boldsymbol{\Sigma}}^{\text{FIM}} = \text{Tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\xi}\boldsymbol{\Sigma}^{-1}\boldsymbol{\eta}\right).$$

Induced pipeline

Step 2:

$$\hat{\boldsymbol{\Sigma}}_{\text{SCM}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^T.$$

Step 3: geodesic distance on $\mathcal{S}_p^{++}$

$$d_{\mathcal{S}_p^{++}}(\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2) = \left\| \log\left(\boldsymbol{\Sigma}_1^{-\frac{1}{2}}\boldsymbol{\Sigma}_2\boldsymbol{\Sigma}_1^{-\frac{1}{2}}\right)\right\|_2.$$

Riemannian gradient descent to solve:

$$\underset{\boldsymbol{\Sigma}\in\mathcal{S}_p^{++}}{\text{minimize}} \sum_i d^2_{\mathcal{S}_p^{++}}(\boldsymbol{\Sigma}, \boldsymbol{\Sigma}_i).$$

Performant and standard pipeline in EEG/MEG signals classification !

A. Barachant et al., "Multi-class Brain Computer Interface Classification by Riemannian Geometry", IEEE Transactions on Biomedical Engineering

Problematics

Go beyond $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$

- $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for non-centered data,
- $\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{\mu}, \tau_i \boldsymbol{\Sigma})$ for non-centered data and robustness,
- $\boldsymbol{x}_i \sim \mathcal{N}(\mathbf{0}, \tau_i \boldsymbol{U}\boldsymbol{U}^T + \boldsymbol{I}_p)$ for high dimensional data and robustness.

Problems

- Existence of maximum likelihood estimators ?
- Not always closed form estimators: how to get fast iterative algo. ?
- Not always closed form expression of the Riemannian distance: what to do ?
- How to get fast estimators of centers of mass ?

Geodesic triangles for machine learning

Geodesic triangles for machine learning

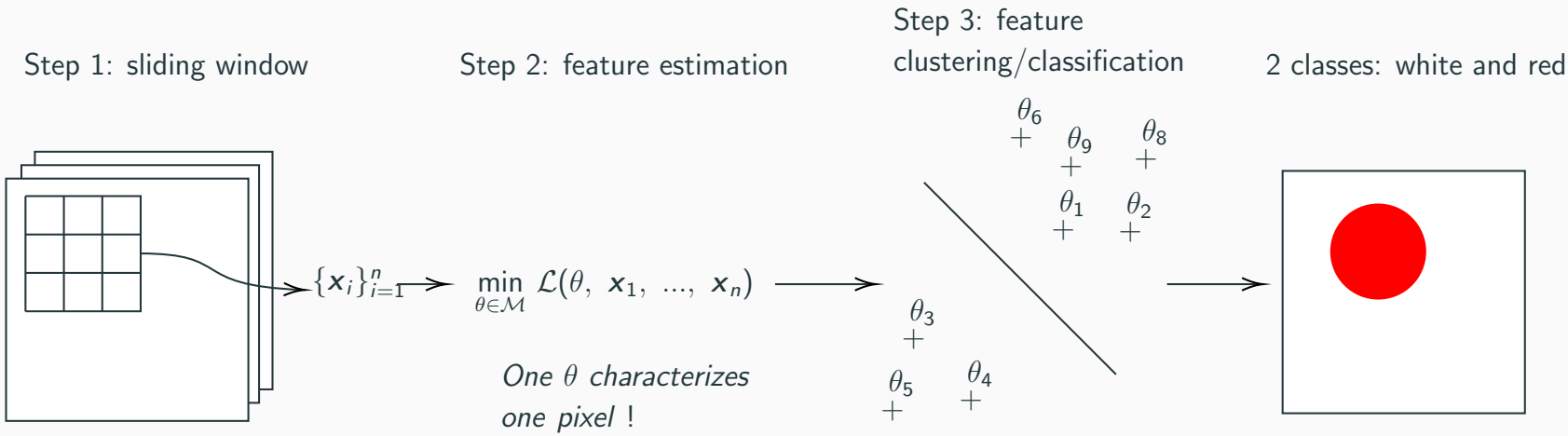


Figure 11: Clustering/classification pipeline.

Non-centered multivariate Gaussian distributions

Let $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n \in \mathbb{R}^p$ distributed as $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu} \in \mathbb{R}^p$, $\boldsymbol{\Sigma} \in \mathcal{S}_p^{++}$.

Goal: classify $\theta = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Geodesic triangles for machine learning

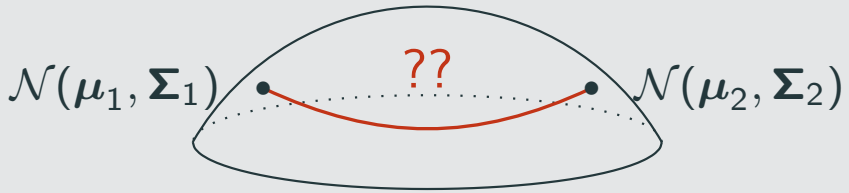
Riemannian geometry of Gaussian distributions

Space of $\theta = (\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in \mathbb{R}^p \times \mathcal{S}_p^{++}$ with the Fisher information metric:
$\forall \xi = \left(\boldsymbol{\xi}_{\boldsymbol{\mu}}, \boldsymbol{\xi}_{\boldsymbol{\Sigma}}\right), \eta = \left(\boldsymbol{\eta}_{\boldsymbol{\mu}}, \boldsymbol{\eta}_{\boldsymbol{\Sigma}}\right)$ in the tangent space

$$\langle \xi, \eta \rangle_\theta^{\mathsf{FIM}} = \boldsymbol{\xi}_{\boldsymbol{\mu}}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\eta}_{\boldsymbol{\mu}} + \frac{1}{2} \operatorname{Tr}\left(\boldsymbol{\Sigma}^{-1} \boldsymbol{\xi}_{\boldsymbol{\Sigma}} \boldsymbol{\Sigma}^{-1} \boldsymbol{\eta}_{\boldsymbol{\Sigma}}\right).$$

Problem

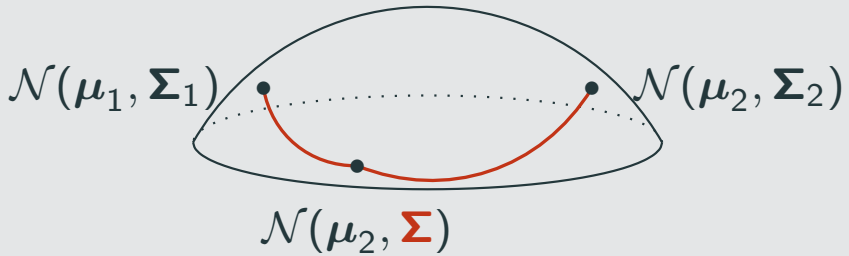
This Riemannian geometry is not fully known...



M. Calvo and J. M. Oller, "An explicit solution of information geodesic equations for the multivariate normal model", Statistics & Risk Modeling

Geodesic triangles for machine learning

Solution: use of geodesic triangles



Divergence δ:
squared arc length of the path between $(\mu_1, \mathbf{\Sigma}_1)$ and $(\mu_2, \mathbf{\Sigma}_2)$.

$$\delta_c : \quad (\mu_1, \mathbf{\Sigma}_1) \to (\mu_2, c\mathbf{\Sigma}_1) \to (\mu_2, \mathbf{\Sigma}_2), \qquad \forall c > 0$$

$$\delta_\perp : \quad (\mu_1, \mathbf{\Sigma}_1) \to (\mu_2, \mathbf{\Sigma}_1 + \Delta\mu\Delta\mu^T) \to (\mu_2, \mathbf{\Sigma}_2), \quad \Delta\mu = \mu_2 - \mu_1$$

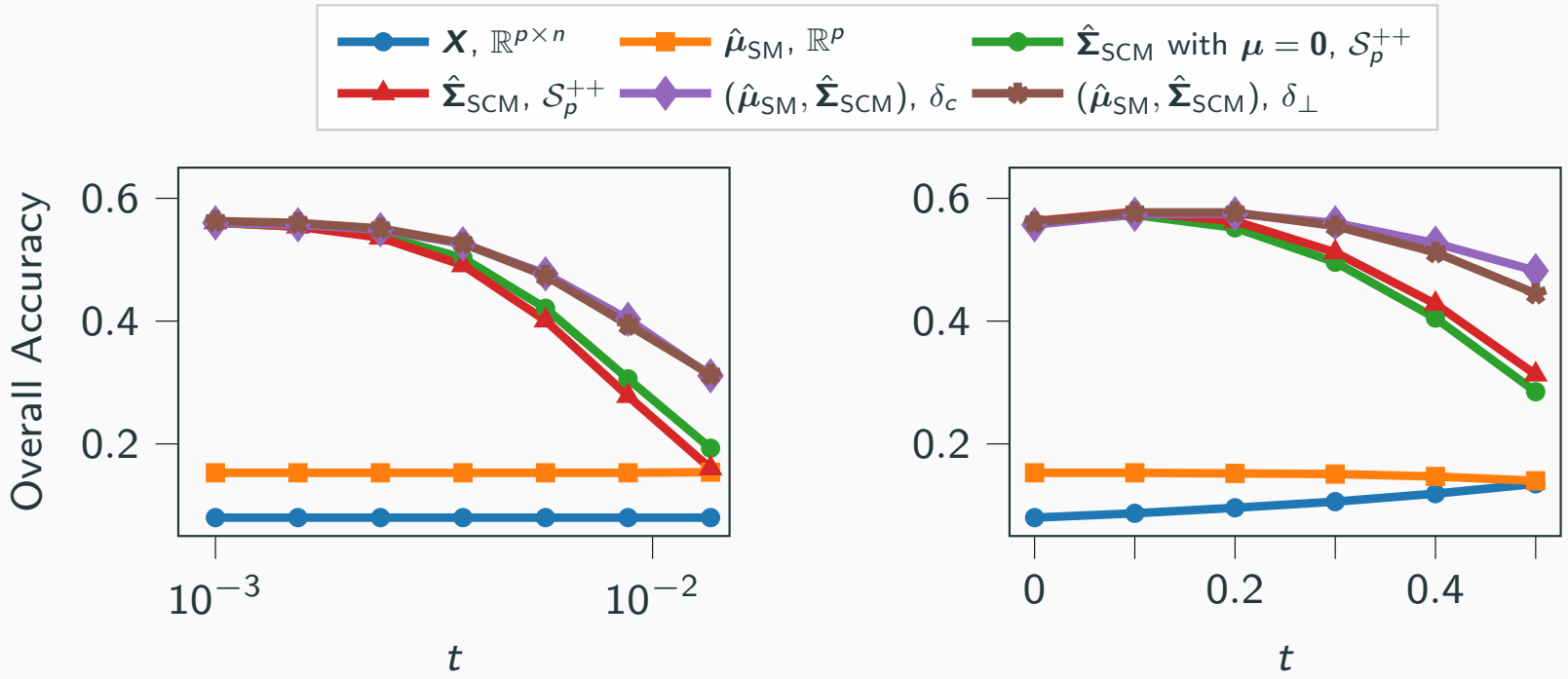
Riemannian center of mass $(\mu, \mathbf{\Sigma})$ of $\{(\mu_i, \mathbf{\Sigma}_i)\}$

$$(\mu, \mathbf{\Sigma}) = \underset{(\mu,\mathbf{\Sigma})\in\mathbb{R}^p\times\mathcal{S}_p^{++}}{\arg\min} \sum_i \delta_{\{c,\perp\}} \left((\mu, \mathbf{\Sigma}), (\mu_i, \mathbf{\Sigma}_i)\right)$$

Computed with a Riemannian gradient descent on $\left(\mathbb{R}^p \times \mathcal{S}_p^{++}, \langle ., . \rangle_.^{\text{FIM}}\right)$.

Application to the Breizhcrops dataset

Parameter estimation + classification with a *Nearest centroïd classifier*



(a) Rotation transformation:
$\boldsymbol{x}_i \mapsto \boldsymbol{Q}(t)^T \boldsymbol{x}_i$ with $\boldsymbol{Q}(0) = \boldsymbol{I}_p$

(b) Scale transformation:
$\boldsymbol{x}_i \mapsto (1-t)\boldsymbol{x}_i$

Figure 12: "Overall Accuracy" versus the parameter t of two transformations applied to the test set of the *Breizhcrops* dataset.

Estimation and classification of non centered and heteroscedastic data

Non-centered mixtures of scaled Gaussian distributions

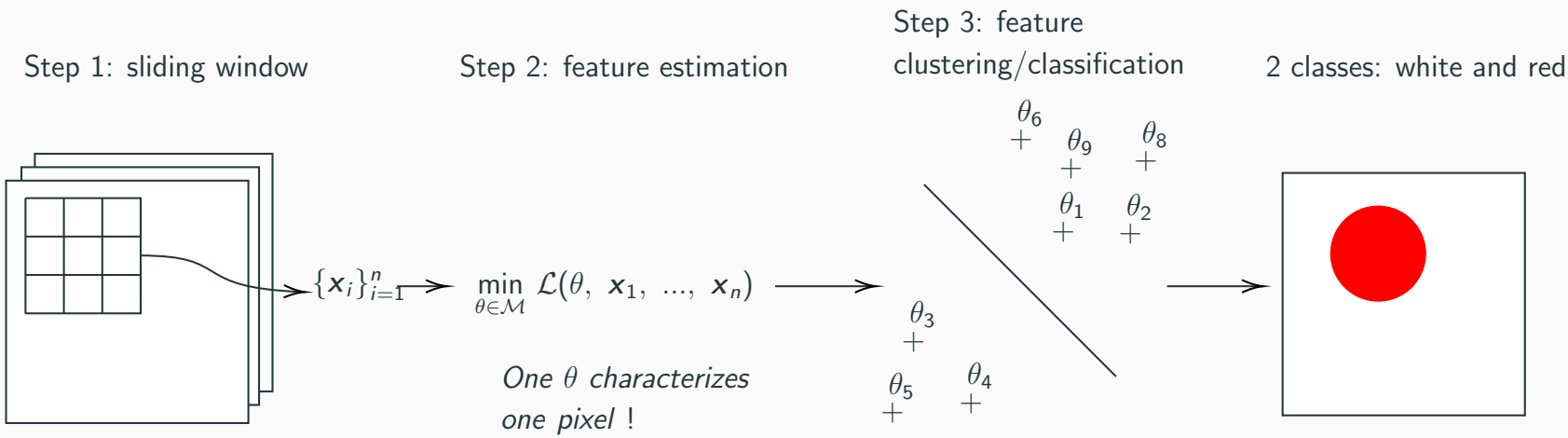


Figure 13: Clustering/classification pipeline.

Non-centered mixtures of scaled Gaussian distributions (NC-MSGs)

Let $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n \in \mathbb{R}^p$ distributed as $\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{\mu}, \tau_i \boldsymbol{\Sigma})$ with $\boldsymbol{\mu} \in \mathbb{R}^p$, $\boldsymbol{\Sigma} \in \mathcal{S}_p^{++}$, and $\boldsymbol{\tau} \in (\mathbb{R}_*^+)^n$.
Goal: estimate and classify $\theta = (\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\tau})$.

Interesting when data are heteroscedastic (*e.g.* time series) and/or contain outliers.

Parameter space and cost functions

Parameter space: location, scatter matrix, and textures

$$\mathcal{M}_{p,n} = \mathbb{R}^p \times \mathcal{S}_p^{++} \times \mathcal{S}(\mathbb{R}_*^+)^n$$

where

$$\mathcal{S}(\mathbb{R}_*^+)^n = \left\{ \boldsymbol{\tau} \in (\mathbb{R}_*^+)^n : \prod_{i=1}^n \tau_i = 1 \right\}$$

- Positivity constraints: $\boldsymbol{\Sigma} \succ \mathbf{0}$, $\tau_i > 0$
- Scale constraint: $\prod_{i=1}^n \tau_i = 1$

Need generic optimization algorithms on $\mathcal{M}_{p,n}$.

Parameter estimation

Minimization of a regularized negative log-likelihood (NLL), $\beta \geq 0$

$$\underset{\theta \in \mathcal{M}_{p,n}}{\text{minimize}} \; \mathcal{L}\left(\theta | \{\boldsymbol{x}_i\}_{i=1}^n\right) + \beta \mathcal{R}_\kappa(\theta)$$

Center of mass estimation

Averaging parameters $\{\theta_i\}_{i=1}^M$ with a to be defined divergence δ

$$\underset{\theta \in \mathcal{M}_{p,n}}{\text{minimize}} \; \frac{1}{M} \sum_{i=1}^M \delta(\theta, \theta_i)$$

Parameter space with a product metric

Product metric

Let $\xi = (\boldsymbol{\xi_\mu}, \boldsymbol{\xi_\Sigma}, \boldsymbol{\xi_\tau})$, $\eta = (\boldsymbol{\eta_\mu}, \boldsymbol{\eta_\Sigma}, \boldsymbol{\eta_\tau})$ in the tangent space,

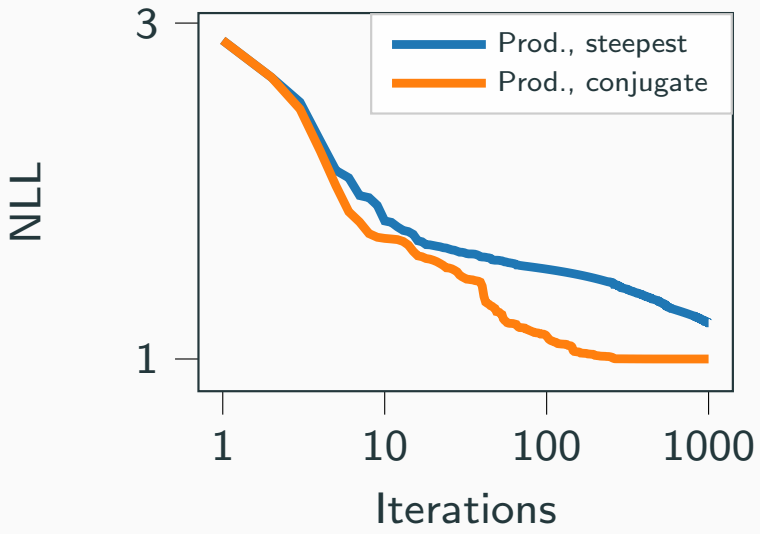
$$\langle \xi, \eta \rangle_\theta^{\mathcal{M}_{p,n}^{\text{Prod.}}} = \boldsymbol{\xi_\mu}^T \boldsymbol{\eta_\mu} + \text{Tr}(\boldsymbol{\Sigma}^{-1} \boldsymbol{\xi_\Sigma} \boldsymbol{\Sigma}^{-1} \boldsymbol{\eta_\Sigma}) + (\boldsymbol{\xi_\tau} \odot \boldsymbol{\tau}^{\odot -1})^T (\boldsymbol{\eta_\tau} \odot \boldsymbol{\tau}^{\odot -1})$$

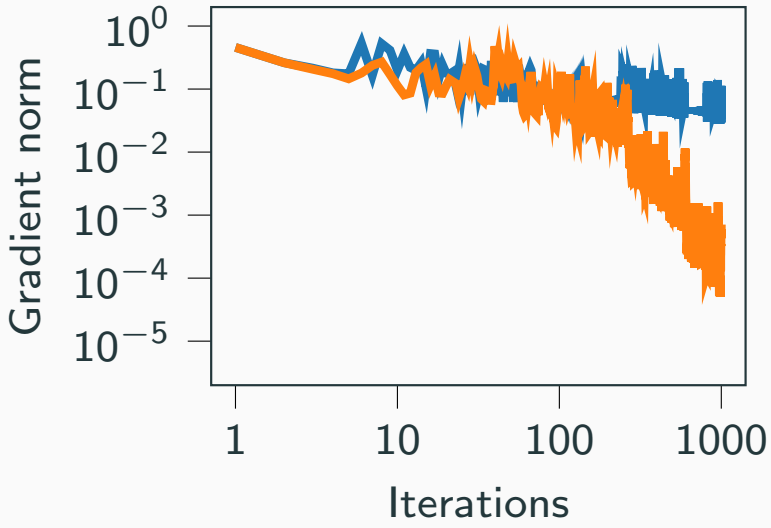
where $\odot$ is the elementwise operator.

Product manifold $\implies$ easy derivation of the geometric tools

$\implies$ Riemannian gradient descent and conjugate gradient on $\left(\mathcal{M}_{p,n}, \langle ., . \rangle_.^{\mathcal{M}_{p,n}^{\text{Prod.}}}\right)$.

Slow in practice ...





Parameter space with the Fisher information metric

Fisher information metric of NC-MSGs

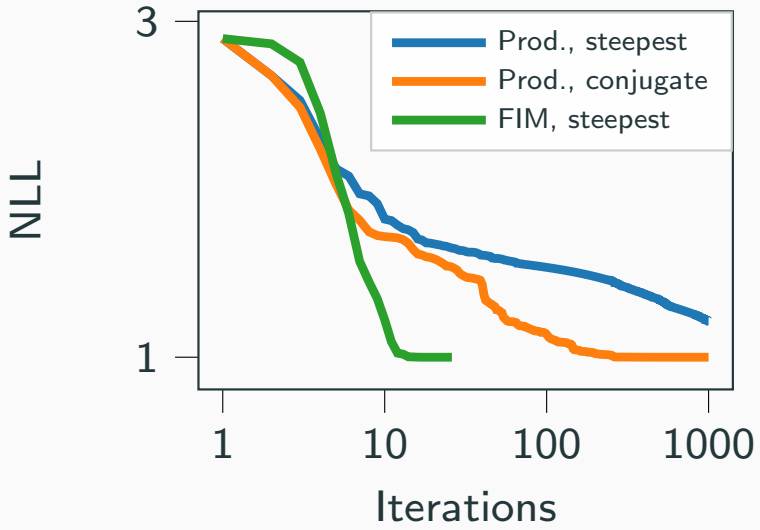
Let $\xi = (\xi_\mu, \xi_\Sigma, \xi_\tau)$, $\eta = (\eta_\mu, \eta_\Sigma, \eta_\tau)$ in the tangent space,

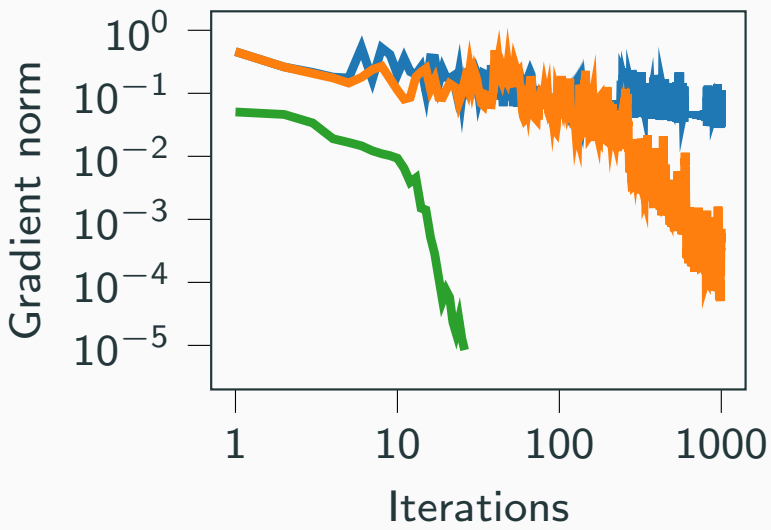
$$\langle \xi, \eta \rangle_\theta^{\mathcal{M}_{p,n}^{\text{FIM}}} = \sum_{i=1}^{n} \frac{1}{\tau_i} \xi_\mu^T \Sigma^{-1} \eta_\mu + \frac{n}{2} \operatorname{Tr}(\Sigma^{-1} \xi_\Sigma \Sigma^{-1} \eta_\Sigma) + \frac{p}{2} (\xi_\tau \odot \tau^{\odot -1})^T (\eta_\tau \odot \tau^{\odot -1})$$

Most geometric tools remain unknown like geodesics, distance ...

But, derivation of the Riemannian gradient and a second order retraction.

$\implies$ Riemannian gradient descent on $\left(\mathcal{M}_{p,n}, \langle ., . \rangle_.^{\mathcal{M}_{p,n}^{\text{FIM}}}\right)$.





Simulated data

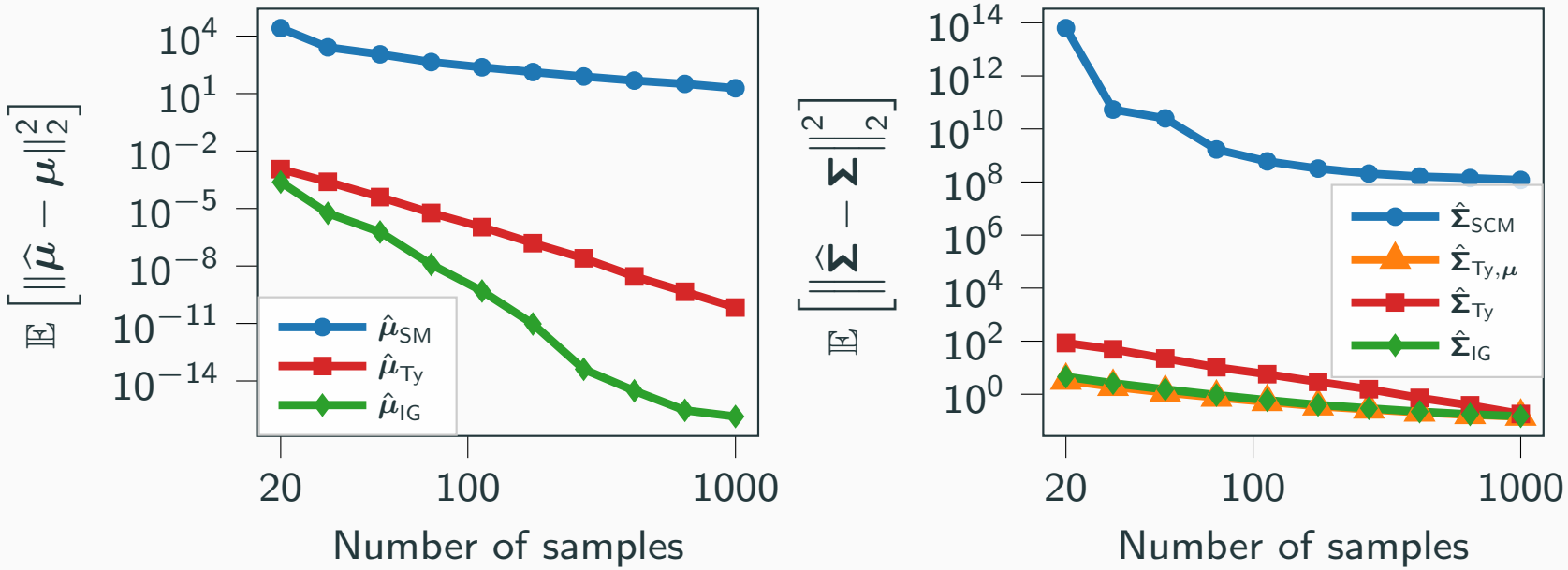


Figure 16: MSE over 2000 simulated sets $\{\boldsymbol{x}_i\}_{i=1}^n \subset \mathbb{R}^{10}$ versus the number samples $\boldsymbol{x}_i$ for the considered estimators $\hat{\boldsymbol{\mu}} \in \{\hat{\boldsymbol{\mu}}_{\text{SM}}, \hat{\boldsymbol{\mu}}_{\text{Ty}}, \hat{\boldsymbol{\mu}}_{\text{IG}}\}$ and $\hat{\boldsymbol{\Sigma}} \in \{\hat{\boldsymbol{\Sigma}}_{\text{SCM}}, \hat{\boldsymbol{\Sigma}}_{\text{Ty},\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}_{\text{Ty}}, \hat{\boldsymbol{\Sigma}}_{\text{IG}}\}$. The proposed estimators are $\hat{\boldsymbol{\mu}}_{\text{IG}}$ and $\hat{\boldsymbol{\Sigma}}_{\text{IG}}$.

D. E. Tyler, "A Distribution-Free *M*-Estimator of Multivariate Scatter", Ann. Statist.

Parameter estimation

NLL neglecting terms not depending on θ:

$$\mathcal{L}(\theta|\{\boldsymbol{x}_i\}_{i=1}^n) = \frac{1}{2}\sum_{i=1}^{n}\left[\log|\tau_i\boldsymbol{\Sigma}| + \frac{(\boldsymbol{x}_i-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu})}{\tau_i}\right].$$

Existence of the maximum likelihood estimator ?

Observation of sequences $(\theta^{(\ell)})_\ell$ such that

$$\mathcal{L}\left(\theta^{(\ell+1)}|\{\boldsymbol{x}_i\}_{i=1}^n\right) < \mathcal{L}\left(\theta^{(\ell)}|\{\boldsymbol{x}_i\}_{i=1}^n\right) \quad \text{and} \quad \theta^{(\ell)} \xrightarrow[\ell\to+\infty]{} \partial\theta$$

where $\partial\theta$ is a border of $\mathcal{M}_{p,n}$ (*e.g.* $\tau_i = 0$).

The existence of a minimum in $\mathcal{M}_{p,n}$ depends on $\{\boldsymbol{x}_i\}_{i=1}^n$!!

Parameter estimation

Existence of a regularized maximum likelihood estimator

Under some assumptions on $\mathcal{R}_\kappa$ and $\beta > 0$, the regularized NLL

$$\theta \mapsto \mathcal{L}\left(\theta | \{\boldsymbol{x}_i\}_{i=1}^n\right) + \beta \mathcal{R}_\kappa(\theta),$$

admits a minimum in $\mathcal{M}_{p,n}$.

Name	$\mathcal{R}_\kappa$
L1 penalty	$\left\|(\mathrm{diag}(\boldsymbol{\tau}) \otimes \boldsymbol{\Sigma})^{-1} - \kappa^{-1} \boldsymbol{I}_{n\times p}\right\|_1 = \sum_{i,j} \left\|\left(\tau_i \lambda_j\right)^{-1} - \kappa^{-1}\right\|$
L2 penalty	$\left\|(\mathrm{diag}(\boldsymbol{\tau}) \otimes \boldsymbol{\Sigma})^{-1} - \kappa^{-1} \boldsymbol{I}_{n\times p}\right\|_2^2 = \sum_{i,j} \left(\left(\tau_i \lambda_j\right)^{-1} - \kappa^{-1}\right)^2$
Bures-Wasserstein squared distance	$d_{\mathrm{BW}}^2\left((\mathrm{diag}(\boldsymbol{\tau}) \otimes \boldsymbol{\Sigma})^{-1}, \kappa^{-1} \boldsymbol{I}_{n\times p}\right) = \sum_{i,j} \left(\left(\tau_i \lambda_j\right)^{-\frac{1}{2}} - \kappa^{-\frac{1}{2}}\right)^2$
Gaussian KL divergence	$\delta_{\mathrm{KL}}(\kappa \boldsymbol{I}_{n\times p}, \mathrm{diag}(\boldsymbol{\tau}) \otimes \boldsymbol{\Sigma}) = \frac{1}{2}\left[\sum_{i,j} \left(\kappa \left(\tau_i \lambda_j\right)^{-1} + \log\left(\tau_i \lambda_j\right)\right) - np(1 + \log(\kappa))\right]$

where λ_j are the eigenvalues of $\boldsymbol{\Sigma}$.

Classification

KL divergence between NC-MSGs

$$\delta_{\mathsf{KL}}(\theta_1, \theta_2) = \frac{1}{2}\left[\sum_{i=1}^{n} \frac{\tau_{1,i}}{\tau_{2,i}} \operatorname{Tr}\left(\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1\right) + \sum_{i=1}^{n} \frac{1}{\tau_{2,i}} \Delta\boldsymbol{\mu}^T \boldsymbol{\Sigma}_2^{-1} \Delta\boldsymbol{\mu} + n \log\left(\frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|}\right) - np\right]$$

with $\Delta\boldsymbol{\mu} = \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$.

Symmetrization:

$$\delta_{\mathcal{M}_{p,n}}(\theta_1, \theta_2) = \frac{1}{2}\left(\delta_{\mathsf{KL}}(\theta_1, \theta_2) + \delta_{\mathsf{KL}}(\theta_2, \theta_1)\right)$$

Riemannian center of mass

$$\underset{\theta \in \mathcal{M}_{p,n}}{\text{minimize}} \ \frac{1}{M} \sum_{i=1}^{M} \delta_{\mathcal{M}_{p,n}}(\theta, \theta_i)$$

Riemannian gradient descent on $\left(\mathcal{M}_{p,n}, \langle ., . \rangle_{.}^{\mathcal{M}_{p,n}^{\mathsf{FIM}}}\right)$.

Simulated data

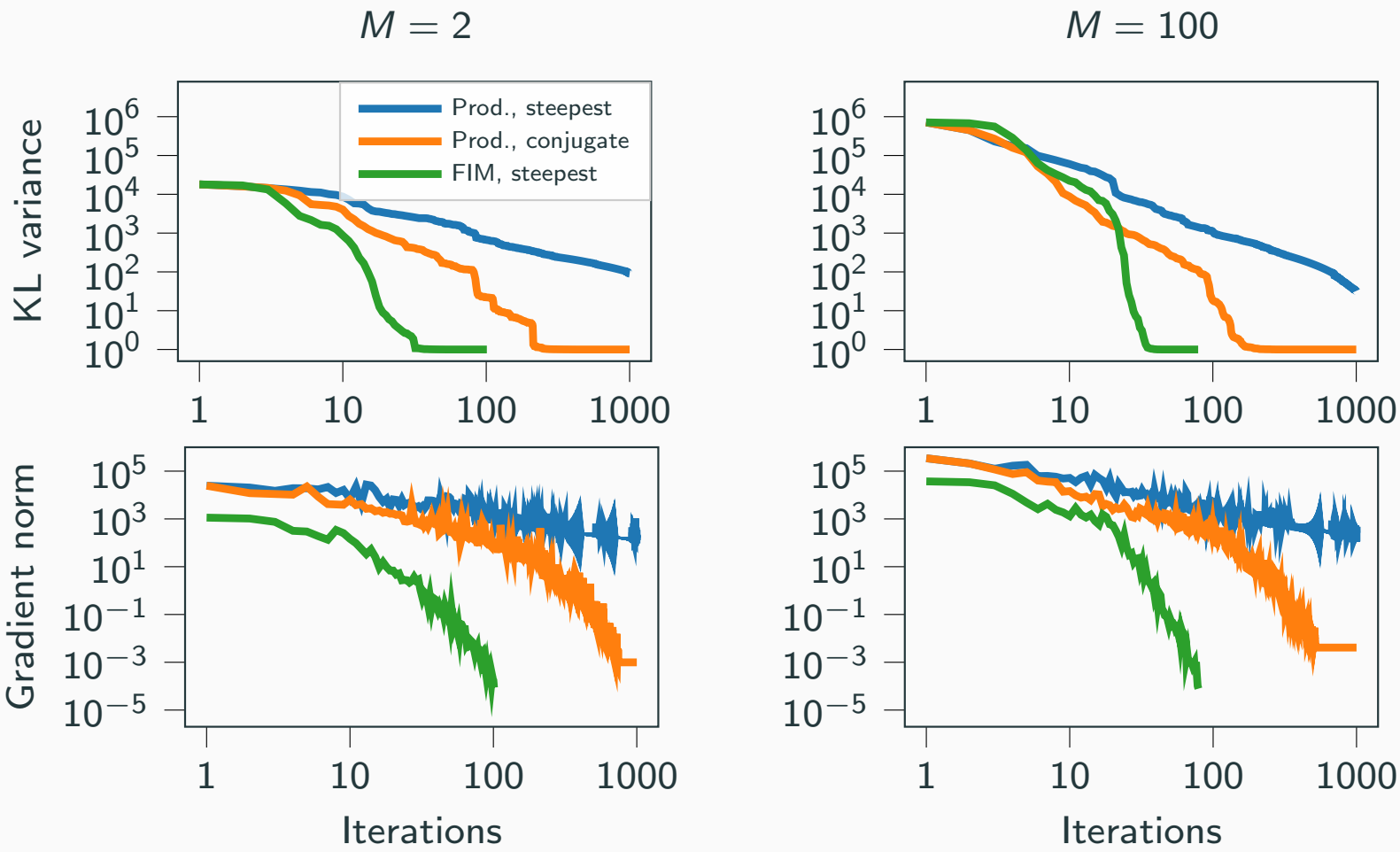
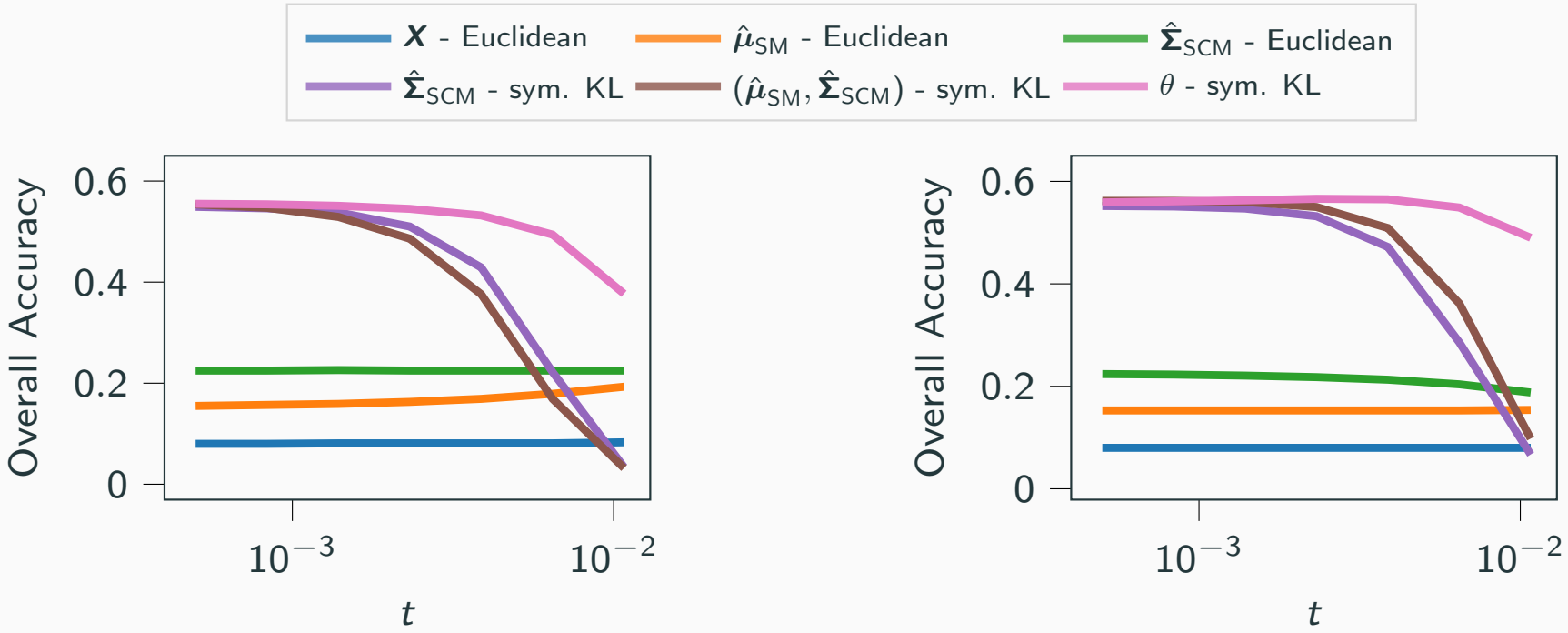


Figure 17: KL variance and its gradient norm versus iterations of the optimizers. The dimensions of the parameter space are $p = 10$ and $n = 150$. Two different numbers of points M are considered: 2 in the left column and 100 in the right one.

Application to the Breizhcrops dataset

Parameter estimation + classification with a *Nearest centroïd classifier*



(a) Mean transformation:
$\boldsymbol{x}_i \mapsto \boldsymbol{x}_i + \boldsymbol{\mu}(t)$ with $\boldsymbol{\mu}(0) = \boldsymbol{0}$

(b) Rotation transformation:
$\boldsymbol{x}_i \mapsto \boldsymbol{Q}(t)^T \boldsymbol{x}_i$ with $\boldsymbol{Q}(0) = \boldsymbol{I}_p$

Figure 18: "Overall Accuracy" metric versus the parameter t associated with transformations applied to the test set. The proposed *Nearest centroïd classifier* is "θ - sym. KL". The regularization is the L2 penalty and $\beta = 10^{-11}$.

Github: `https://github.com/antoinecollas/optim_compound`

Probabilistic PCA from heteroscedastic signals

Study of a "low rank" statistical model

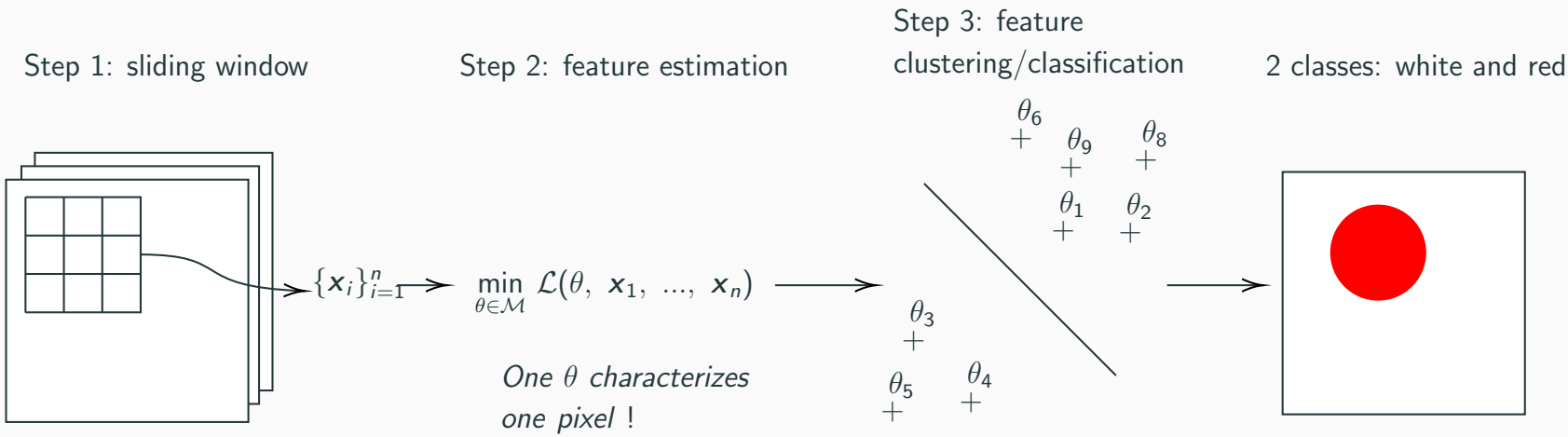


Figure 19: Clustering/classification pipeline.

Statistical model

$\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n \in \mathbb{R}^p$, $\forall k < p$:

$$\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{0}, \tau_i \boldsymbol{U}\boldsymbol{U}^T + \boldsymbol{I}_p)$$

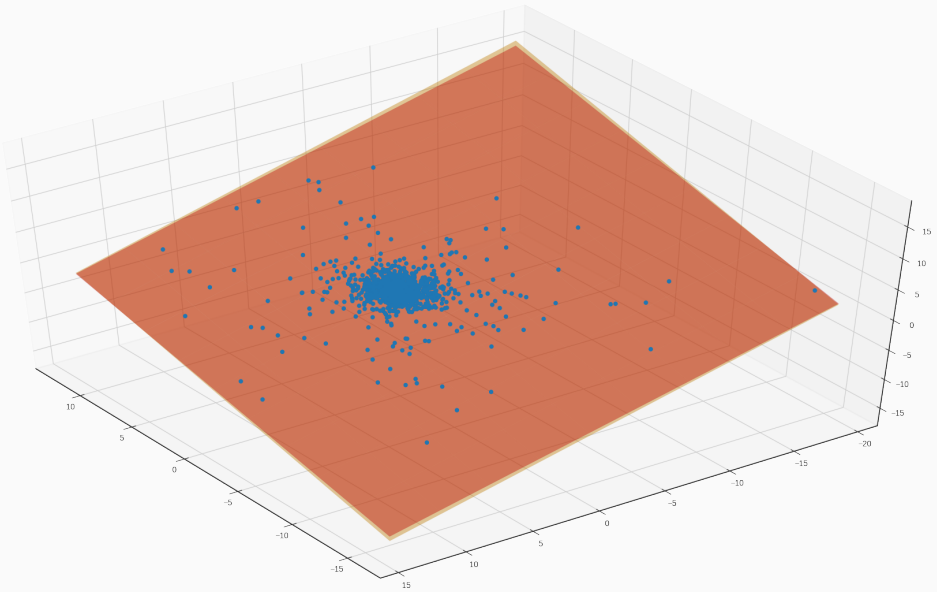
with $\tau_i > 0$ and $\boldsymbol{U} \in \mathbb{R}^{p \times k}$ is an orthogonal basis ($\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}_k$).
Goal: estimate and classify $\theta = (\boldsymbol{U}, \boldsymbol{\tau})$.

Study of a "low rank" statistical model

Statistical model

$$\underbrace{\boldsymbol{x}_i}_{\in\mathbb{R}^p} \overset{d}{=} \underbrace{\sqrt{\tau_i}\boldsymbol{U}\boldsymbol{g}}_{\text{signal}\in\text{span}(\boldsymbol{U})} + \underbrace{\boldsymbol{n}}_{\text{noise}\in\mathbb{R}^p}$$

where $\boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_k)$ and $\boldsymbol{n} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_p)$ are independent, $\boldsymbol{\tau} \in (\mathbb{R}_*^+)^n$, and $\boldsymbol{U} \in \mathbb{R}^{p\times k}$ is an orthogonal basis ($\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}_k$).



Study of a "low rank" statistical model: estimation

Maximum likelihood estimation

Minimization of the NLL with constraints, $\theta = (\boldsymbol{U}, \boldsymbol{\tau})$

- $\boldsymbol{U} \in \mathrm{Gr}_{p,k}$: orthogonal basis of the subspace (and thus invariant by rotation !)
- $\boldsymbol{\tau} \in (\mathbb{R}_*^+)^n$: positivity constraints

$$\underset{\theta \in \mathrm{Gr}_{p,k} \times (\mathbb{R}_*^+)^n}{\text{minimize}} \; \mathcal{L}(\theta | \, \{\boldsymbol{x}_i\}_{i=1}^n)$$

Study of a "low rank" statistical model: estimation

Fisher information metric

$\forall \xi = (\xi_U, \xi_\tau), \eta = (\eta_U, \eta_\tau)$ in the tangent space

$$\langle \xi, \eta \rangle_\theta^{\text{FIM}} = 2nc_\tau \operatorname{Tr}\left(\xi_U^T \eta_U\right) + k\left(\xi_\tau \odot (1+\tau)^{\odot -1}\right)^T \left(\eta_\tau \odot (1+\tau)^{\odot -1}\right),$$

where $c_\tau = \frac{1}{n}\sum_{i=1}^{n} \frac{\tau_i^2}{1+\tau_i}$.

Derivation of the Riemannian gradient and of a retraction.

To minimize the NLL: Riemannian gradient descent on $\left(\text{Gr}_{p,k} \times (\mathbb{R}_*^+)^n, \langle ., . \rangle_.^{\text{FIM}}\right)$.

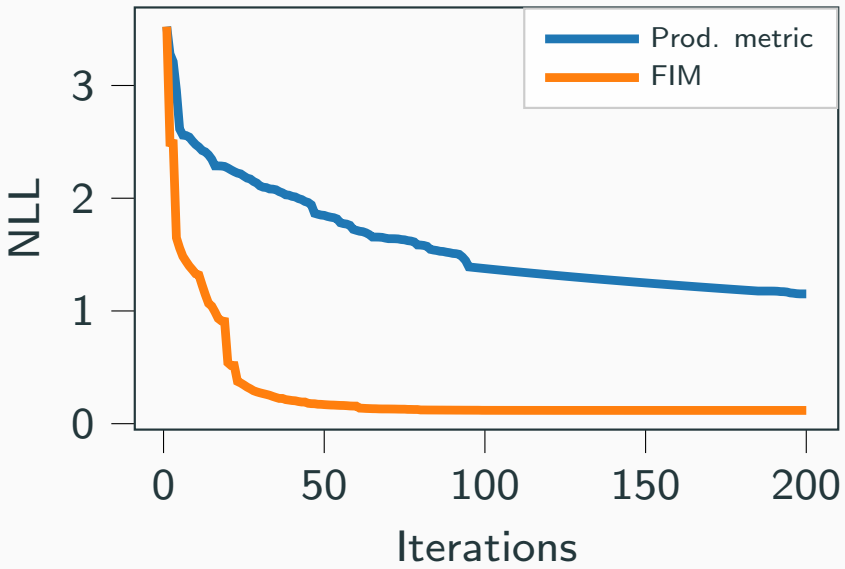


Figure 20: NLL versus the iterations.

Study of a "low rank" statistical model: bounds

Intrinsic Cramér-Rao bounds

Study of the performance through intrinsic Cramér-Rao bounds:

$$\overbrace{\mathbb{E}[d^2_{\mathrm{Gr}_{p,k}}(\mathrm{span}(\hat{\boldsymbol{U}}),\mathrm{span}(\boldsymbol{U}))]}^{\text{subspace estimation error}} \geq \frac{(p-k)k}{nc_{\boldsymbol{\tau}}} \approx \frac{(p-k)k}{n \times \mathrm{SNR}}$$

$$\underbrace{\mathbb{E}[d^2_{(\mathbb{R}^+_*)^n}(\hat{\boldsymbol{\tau}},\boldsymbol{\tau})]}_{\text{texture estimation error}} \geq \frac{1}{k}\sum_{i=1}^{n}\frac{(1+\tau_i)^2}{\tau_i^2}$$

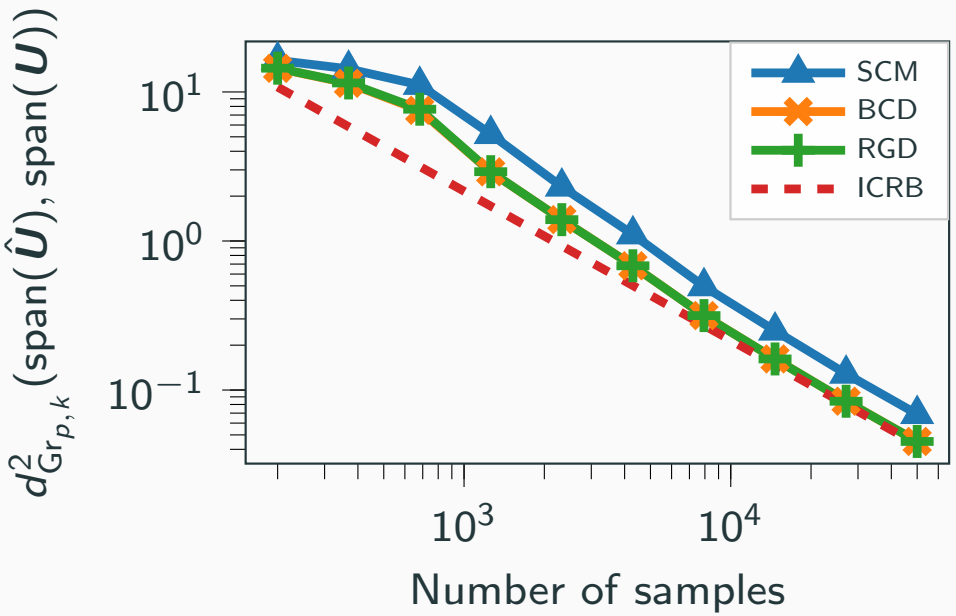


Figure 21: Mean squared error versus the number of simulated data.

Study of a "low rank" statistical model: K-means++

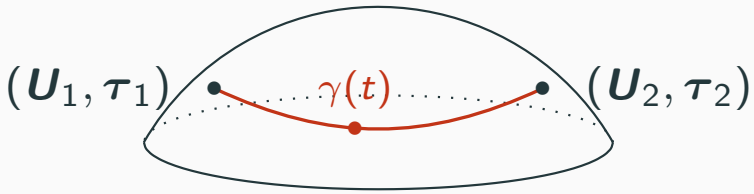


Figure 22: Distance.

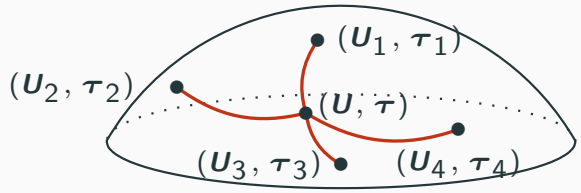


Figure 23: Center of mass (U, τ).

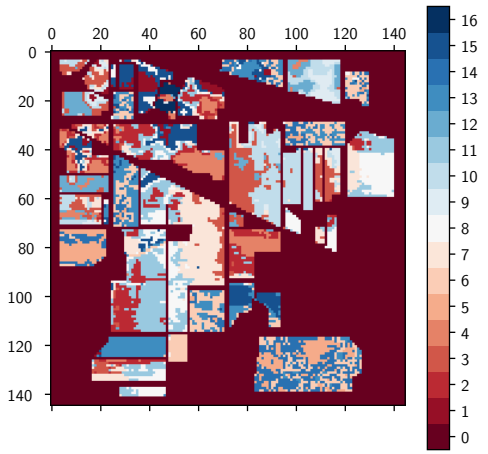


Figure 24:
Euclidean *K-means++*:
OA = 31.2%.

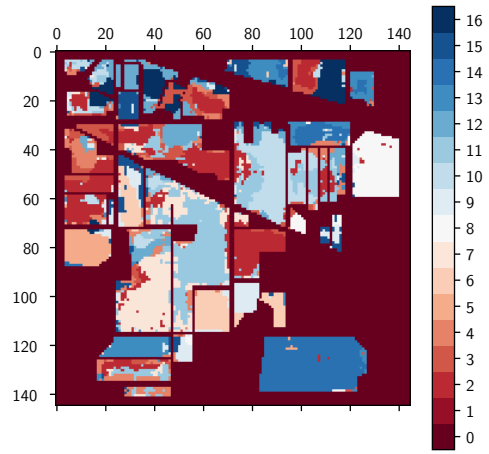


Figure 25:
Proposed *K-means++*:
OA = 47.2%.

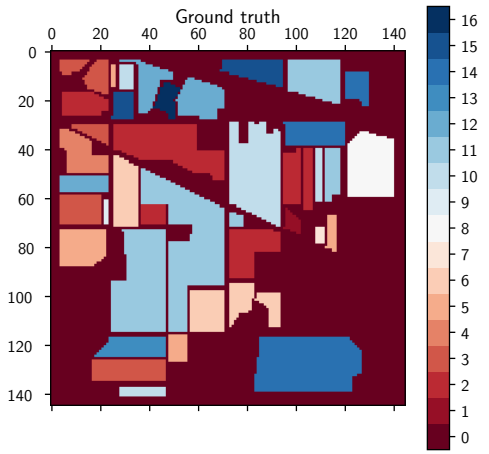


Figure 26: Ground truth.

Github: `https://github.com/antoinecollas/pyCovariance`

Conclusions

Conclusions

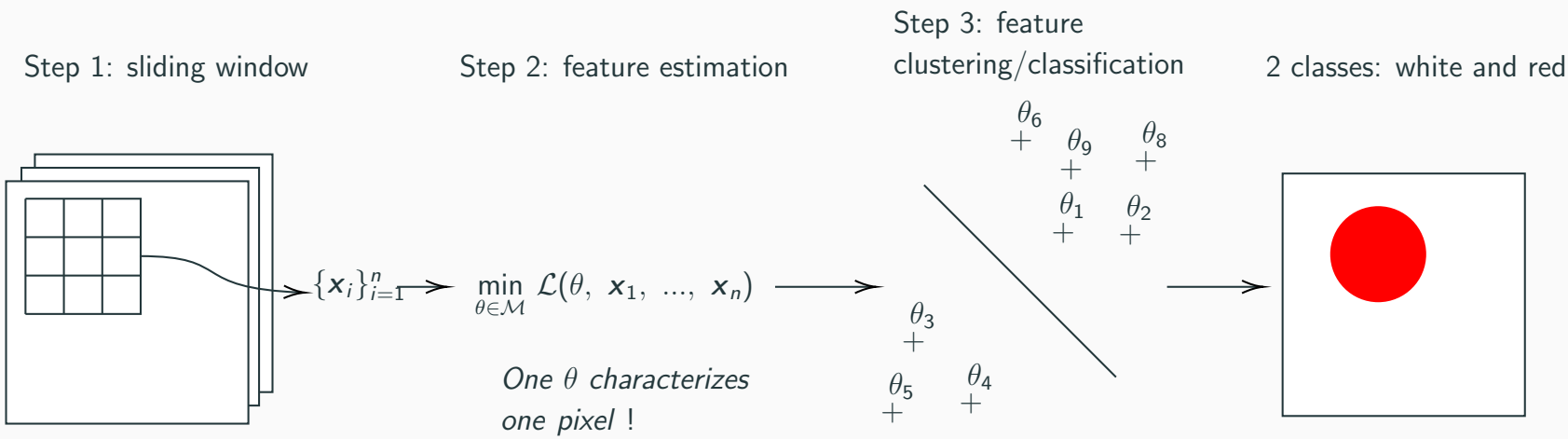


Figure 27: Clustering/classification pipeline.

Interests of using **Riemannian geometry** in this pipeline:

- fast estimators, analysis of constrained estimation problems using intrinsic Cramér-Rao bounds,
- new classifiers robust to distribution shifts.

Applications on real datasets: *Indian pines* hyperspectral image, *Breizhcrops* multispectral times series, datasets from the *UCI* repository.

Estimation and classification of location and covariance matrix using Riemannian geometry: application to remote sensing

Antoine Collas

Postdoc in the Mind team at Inria Saclay, supervised by Alexandre Gramfort (Meta).

PhD done at SONDRA lab, CentraleSupélec, Univ. Paris-Saclay in 2019-2022.

Robust Geometric Metric Learning

Robust Geometric Metric Learning (RGML)

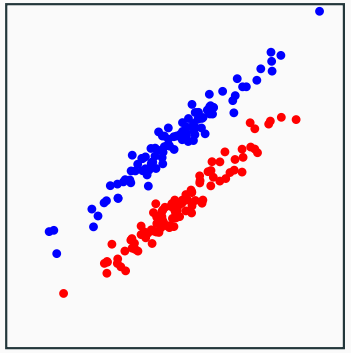
Let be a classification problem with K classes.

Metric learning

Find a *Mahalanobis* distance

$$d_{\boldsymbol{A}}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sqrt{(\boldsymbol{x}_i - \boldsymbol{x}_j)^T \boldsymbol{A}^{-1} (\boldsymbol{x}_i - \boldsymbol{x}_j)}$$

relevant for classification problems.



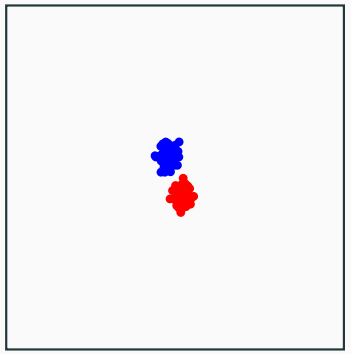
$\{\boldsymbol{x}_i\}$

Metric learning as covariance estimation

Proposed minimization problem, $\theta = (\boldsymbol{A}, \{\boldsymbol{A}_k\})$

$$\underset{\theta \in \left(\mathcal{S}_p^{++}\right)^{K+1}}{\text{minimize}} \quad \underbrace{\sum_{k=1}^{K} \pi_k \mathcal{L}_k(\boldsymbol{A}_k)}_{\text{negative log-likelihood}} + \lambda \underbrace{\sum_{k=1}^{K} \pi_k d^2_{\mathcal{S}_p^{++}}(\boldsymbol{A}, \boldsymbol{A}_k)}_{\substack{\text{cost function to compute} \\ \text{the center of mass of } \{\boldsymbol{A}_k\}}}$$

$\{\pi_k\}$ are the proportions of the classes and $\{\mathcal{L}_k\}$ are to be defined.



$\left\{\boldsymbol{A}^{-\frac{1}{2}} \boldsymbol{x}_i\right\}$

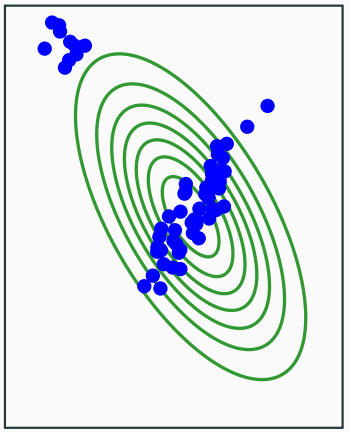
Robust Geometric Metric Learning (RGML)

Let $\boldsymbol{s}_{ki} = \boldsymbol{x}_l - \boldsymbol{x}_m$ where $\boldsymbol{x}_l, \boldsymbol{x}_m$ belong to the class k.

Gaussian negative log-likelihood

$$\mathcal{L}_{G,k}(\boldsymbol{A}_k) = \frac{1}{n_k} \sum_{i=1}^{n_k} \boldsymbol{s}_{ki}^T \boldsymbol{A}_k^{-1} \boldsymbol{s}_{ki} + \log |\boldsymbol{A}_k|$$

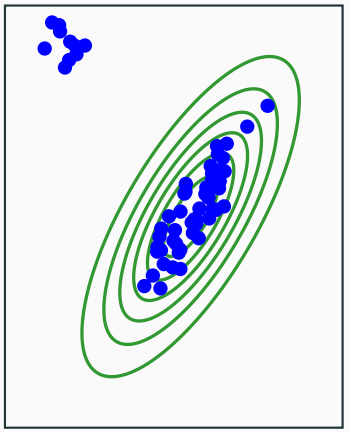
$$\text{minimized for } \boldsymbol{A}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} \boldsymbol{s}_{ki} \boldsymbol{s}_{ki}^T$$



Tyler cost function

$$\mathcal{L}_{T,k}(\boldsymbol{A}_k) = \frac{p}{n_k} \sum_{i=1}^{n_k} \log \left(\boldsymbol{s}_{ki}^T \boldsymbol{A}_k^{-1} \boldsymbol{s}_{ki} \right) + \log |\boldsymbol{A}_k|$$

$$\text{minimized for } \boldsymbol{A}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} \underbrace{\frac{p}{\boldsymbol{s}_{ki}^T \boldsymbol{A}_k^{-1} \boldsymbol{s}_{ki}}}_{\substack{\text{weight of} \\ \text{sample } \boldsymbol{s}_{ki}}} \boldsymbol{s}_{ki} \boldsymbol{s}_{ki}^T$$



Robust Geometric Metric Learning (RGML)

Riemannian metric

$\forall \xi = (\boldsymbol{\xi}, \{\boldsymbol{\xi}_k\}), \eta = (\boldsymbol{\eta}, \{\boldsymbol{\eta}_k\})$ in the tangent space

$$\langle \xi, \eta \rangle_\theta = \mathrm{Tr}\left(\boldsymbol{A}^{-1}\boldsymbol{\xi}\boldsymbol{A}^{-1}\boldsymbol{\eta}\right) + \sum_{k=1}^{K} \mathrm{Tr}\left(\boldsymbol{A}_k^{-1}\boldsymbol{\xi}_k\boldsymbol{A}_k^{-1}\boldsymbol{\eta}_k\right)$$

$\Longrightarrow$ strongly geodesically convexity of the minimization problem

$\Longrightarrow$ the Riemannian gradient descent is fast

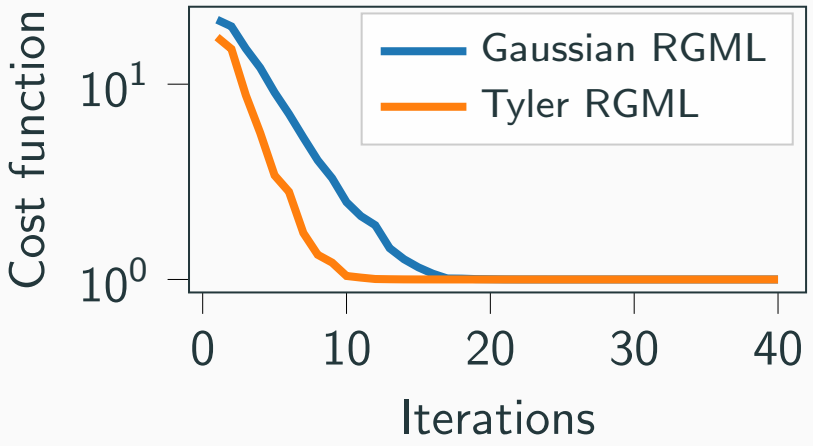


Figure 28: Cost function versus the iterations.

Robust Geometric Metric Learning (RGML)

RGML + *k-NN* on datasets from the UCI Machine Learning Repository

	Wine $p = 13$, $n = 178$, $K = 3$				Vehicle $p = 18$, $n = 846$, $K = 4$				Iris $p = 4$, $n = 150$, $K = 3$			
Method	Mislabeling rate				Mislabeling rate				Mislabeling rate			
	0%	5%	10%	15%	0%	5%	10%	15%	0%	5%	10%	15%
Euclidean	30.12	30.40	31.40	32.40	38.27	38.58	39.46	40.35	3.93	4.47	5.31	**6.70**
SCM	10.03	11.62	13.70	17.57	23.59	24.27	25.24	26.51	12.57	13.38	14.93	16.68
ITML - Identity	3.12	4.15	5.40	**7.74**	24.21	23.91	24.77	26.03	3.04	4.47	5.31	**6.70**
ITML - SCM	2.45	4.76	6.71	10.25	23.86	23.82	24.89	26.30	3.05	13.38	14.92	16.67
GMML	2.16	3.58	5.71	9.86	21.43	22.49	23.58	25.11	2.60	5.61	9.30	12.62
LMNN	4.27	6.47	7.83	9.86	20.96	24.23	26.28	28.89	3.53	9.59	11.19	12.22
Proposed - Gaussian	**2.07**	**2.93**	5.15	9.20	**19.76**	21.19	22.52	24.21	**2.47**	5.10	8.90	12.73
Proposed - Tyler	**2.12**	**2.90**	**4.51**	8.31	19.90	**20.96**	**22.11**	**23.58**	**2.48**	**2.96**	**4.65**	7.83

Table 1: Misclassification errors on 3 datasets: Wine, Vehicle and Iris. Mislabeling rate: percentage of labels randomly changed in the training set.

Github: `https://github.com/antoinecollas/robust_metric_learning`

Perspectives

Perspectives

- Beyond KL divergence between non-centered mixtures of scaled Gaussian distributions:
  - Wasserstein distance
  - Entropy-Regularized Optimal Transport

  Wasserstein distance between Gaussian distributions (from H. Janati et al., "Entropic Optimal Transport between Unbalanced Gaussian Measures has a Closed Form", Neurips 2020):

$$W_2^2\left(\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)\right) = \\ \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_2^2 + \mathrm{Tr}(\boldsymbol{\Sigma}_1) + \mathrm{Tr}(\boldsymbol{\Sigma}_2) - 2\,\mathrm{Tr}\left(\boldsymbol{\Sigma}_1^{\frac{1}{2}}\boldsymbol{\Sigma}_2\boldsymbol{\Sigma}_1^{\frac{1}{2}}\right)^{\frac{1}{2}}$$

- Apply *RGML* to more recent datasets (larger p, larger n, larger K): SGD, SGA, "low rank + identity" structure, ...

Perspectives

- Integrate developed estimators in a differentiable pipeline.

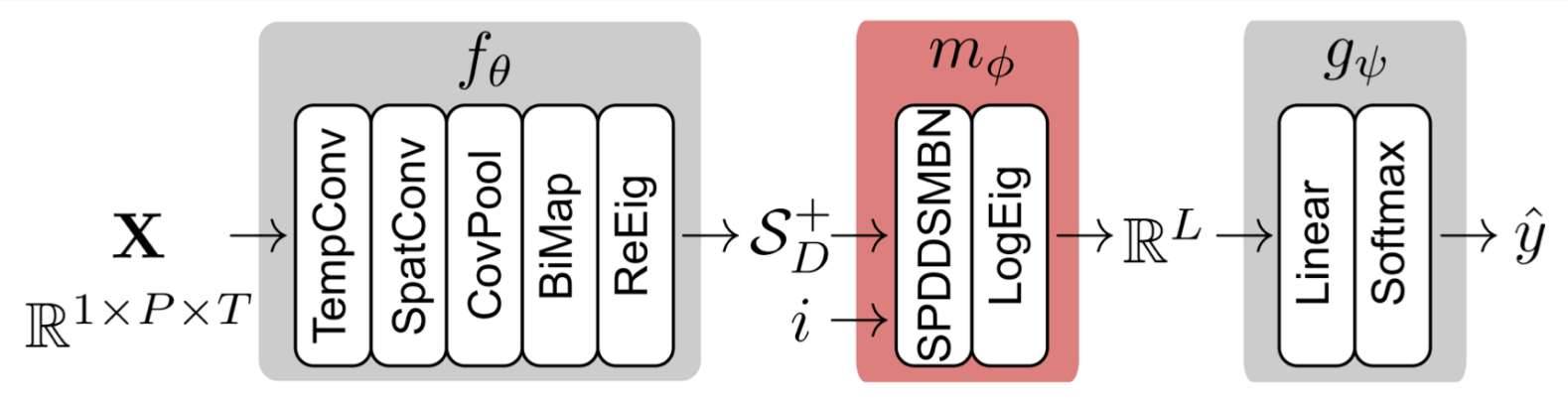


Figure 29: Tangent Space Mapping Network (from R. J. Kobler et al., "SPD domain-specific batch normalization to crack interpretable unsupervised domain adaptation in EEG", Neurips 2022)

- Domain adaptation: see if developed algorithms are robust to distributions shifts on real datasets; *e.g.* inter-session and -subject in EEG datasets.

Perspectives

Fisher information metric for optimization:

- Only a good preconditioner ?
- When second order retractions are useful ?

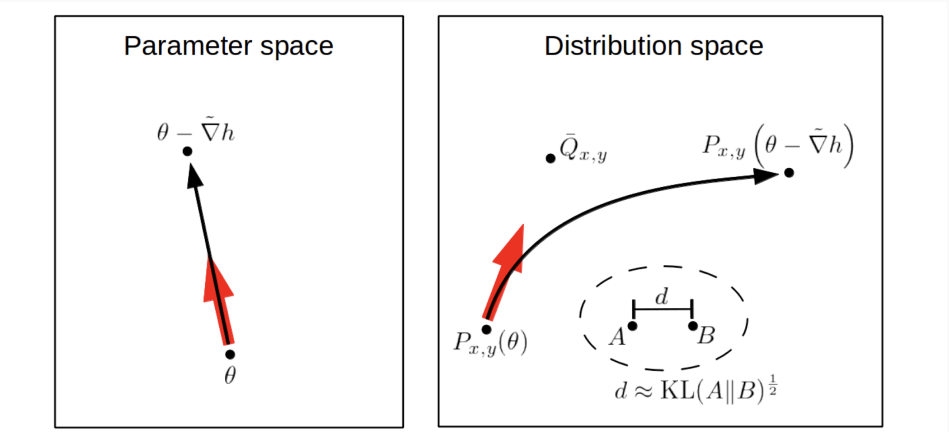


Figure 1: A typical situation encountered when performing large discrete updates in the original parameter space. The red arrow is the natural gradient direction (given by the vector $\tilde{\nabla} h$ in parameter space) and the black arrow is the path generated by taking $\theta - \alpha \tilde{\nabla} h$ for $\alpha \in [0, 1]$.

Figure 30: Euclidean retraction (from James Martens, "New Insights and Perspectives on the Natural Gradient Method", JMLR 2020)

- Iterates defined up to reparametrizations of the sample and parameter spaces:

$$\underset{\boldsymbol{\Sigma} \in \mathcal{S}_p^{++}}{\text{minimize}}\ f(\boldsymbol{\Sigma} |\ \{\boldsymbol{x}_i\}) = f(\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T |\ \{\boldsymbol{A}\boldsymbol{x}_i + \boldsymbol{\mu}\}) \quad (\boldsymbol{A} \text{ invertible})$$

$$\boldsymbol{\Sigma}^{(\ell+1)} = \exp^{\mathcal{S}_p^{++}}_{\boldsymbol{\Sigma}^{(\ell)}}(-\alpha \nabla f(\boldsymbol{\Sigma}^{(\ell)}))$$

$$\exp^{\mathcal{S}_p^{++}}_{\boldsymbol{A}\boldsymbol{\Sigma}^{(\ell)}\boldsymbol{A}^T}(-\alpha \nabla f(\boldsymbol{A}\boldsymbol{\Sigma}^{(\ell)}\boldsymbol{A}^T)) = \boldsymbol{A} \exp^{\mathcal{S}_p^{++}}_{\boldsymbol{\Sigma}}(-\alpha \nabla f(\boldsymbol{\Sigma}^{(\ell)}))\boldsymbol{A}^T = \boldsymbol{A}\boldsymbol{\Sigma}^{(\ell+1)}\boldsymbol{A}^T$$

References, publications and additional figures

Riemannian geometry, robust estimation and libraries

Optimization on Riemannian manifolds:

[8] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton, NJ: Princeton University Press, 2008, pp. xvi+224. ISBN: 978-0-691-13298-3.

[9] Nicolas Boumal. *An introduction to optimization on smooth manifolds*. 2022. URL: http://www.nicolasboumal.net/book.

Robust statistics:

[10] Ricardo Antonio Maronna. *Robust M-Estimators of Multivariate Location and Scatter*. 1976. DOI: 10.1214/aos/1176343347.

[11] David E. Tyler. *A Distribution-Free M-Estimator of Multivariate Scatter*. 1987. DOI: 10.1214/aos/1176350263.

[12] Esa Ollila et al. *Complex Elliptically Symmetric Distributions: Survey, New Results and Applications*. 2012. DOI: 10.1109/TSP.2012.2212433.

Software:

[13] N. Boumal et al. *Manopt, a Matlab Toolbox for Optimization on Manifolds*. 2014. URL: https://www.manopt.org.

[14] J. Townsend, N. Koep, and S. Weichwald. *Pymanopt: A Python Toolbox for Optimization on Manifolds Using Automatic Differentiation*. 2016.

Publications

Conferences:

[1] **A. Collas** et al. "A Tyler-Type Estimator of Location and Scatter Leveraging Riemannian Optimization". In: *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, Canada (Virtual)*. 2021.

[2] **A. Collas** et al. "On The Use of Geodesic Triangles Between Gaussian Distributions for Classification Problems". In: *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore*. 2022.

[3] **A. Collas** et al. "Robust Geometric Metric Learning". In: *2022 30th European Signal Processing Conference (EUSIPCO), Belgrade, Serbia*. 2022.

[4] **A. Collas** et al. "Apprentissage robuste de distance par géométrie riemannienne". In: *GRETSI 2022 XXVIIIème colloque, Nancy, France*. 2022.

Journals:

[5] A. Mian, **A**. **Collas** et al. "Robust Low-Rank Change Detection for Multivariate SAR Image Time Series". In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* (2020).

[6] **A. Collas** et al. "Probabilistic PCA From Heteroscedastic Signals: Geometric Framework and Application to Clustering". In: *IEEE Transactions on Signal Processing* (2021).

[7] **A. Collas** et al. "Riemannian optimization for non-centered mixture of scaled Gaussian distributions". In: *submitted to IEEE Transactions on Signal Processing* (2022).

Award:

2022: Best Student Paper Award at the EUSIPCO 2022 conference.

Study of a "low rank" statistical model: estimation

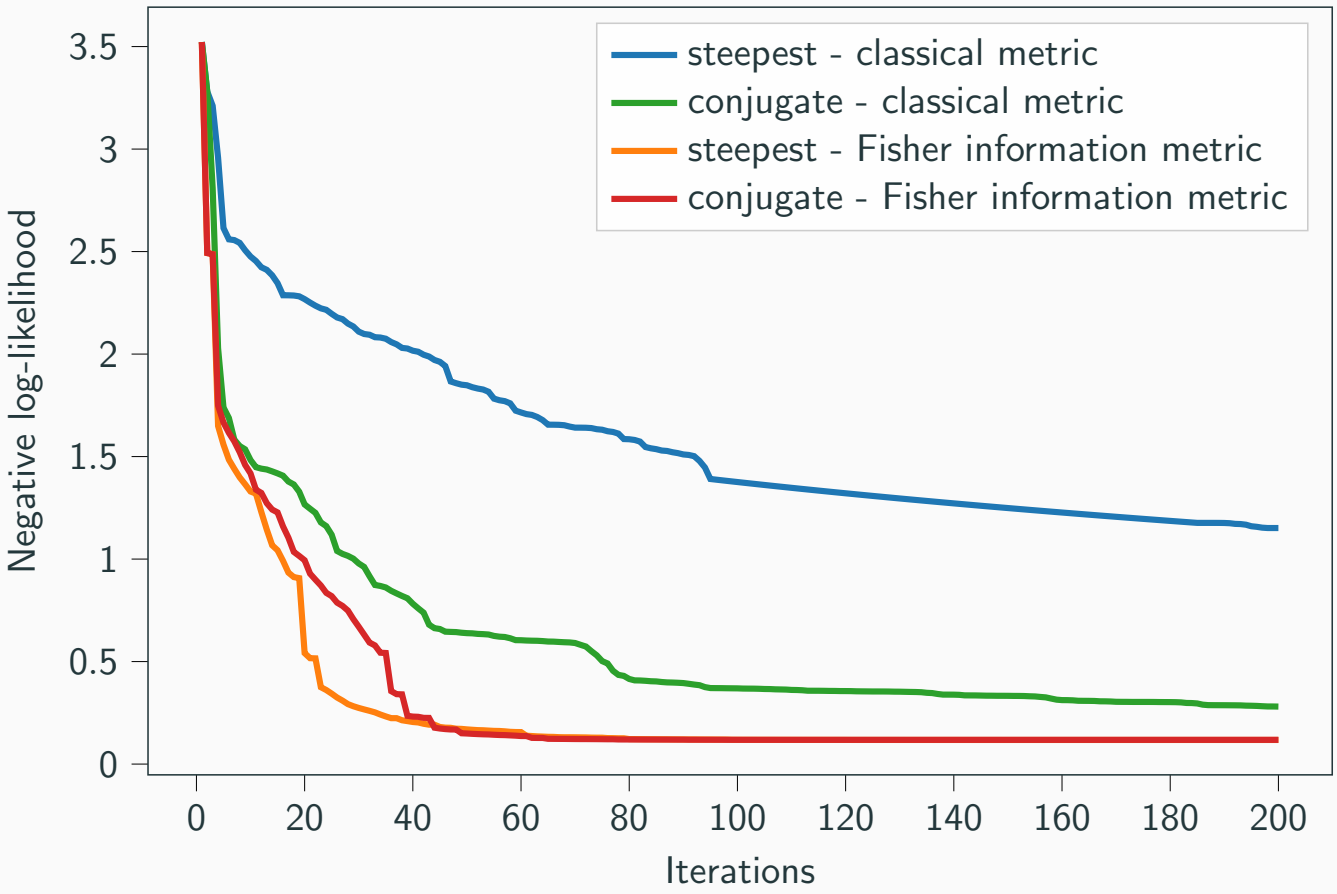


Figure 31: Negative log-likelihood versus the iterations.

NC-MSGs: application to the Breizhcrops dataset

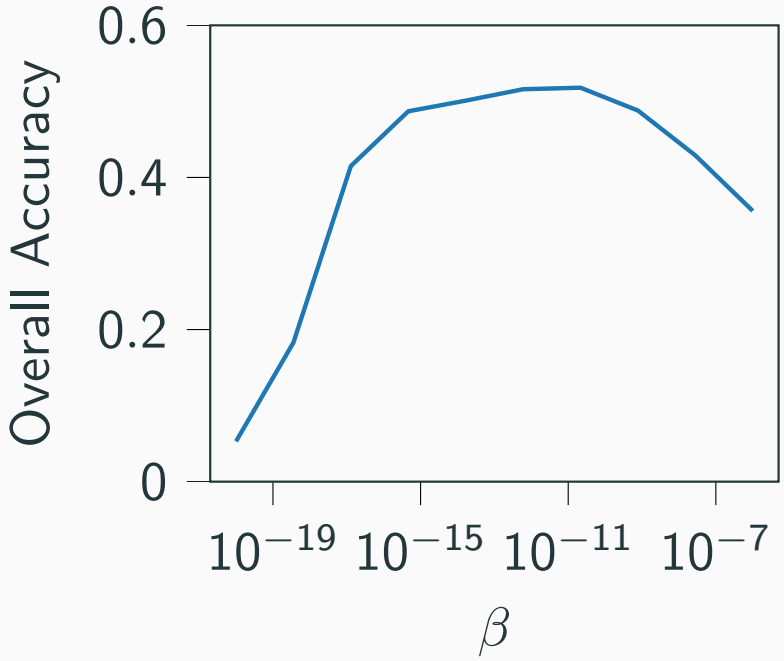


Figure 32: "Overall Accuracy" metric achieved by the proposed *Nearest centroïd classifier* on the *Breizhcrops* dataset versus β. The regularization is the L2 penalty.

NC-MSGs: numerical experiments for the estimation

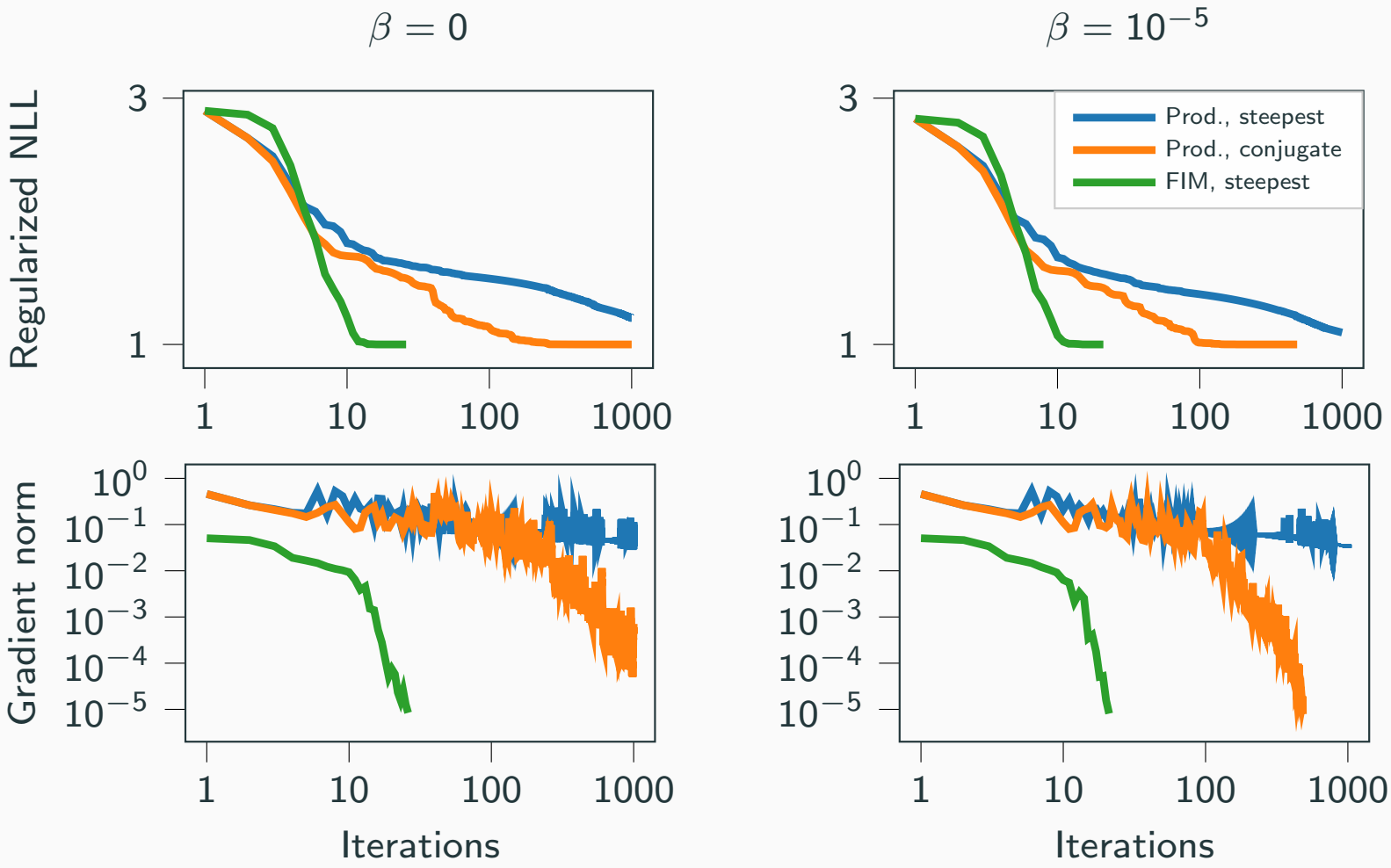


Figure 33: Regularized NLL and its gradient norm versus iterations of the optimizers. The chosen regularization is the L2 penalty. Each estimation is performed on $n = 150$ samples in $\mathbb{R}^{10}$ sampled from a NC-MSG.