Riemannian geometry for statistical estimation and learning: applications to remote sensing and M/EEG

Antoine Collas

OPIS seminar

Education and Research

- **2022 - Present: Postdoctoral Researcher in Machine Learning**
  Mind team (ex-Parietal) at Inria Saclay
  Work with Alexandre Gramfort, Rémi Flamary

- **2019-2022: PhD in Signal processing**
  SONDRA laboratory, CentraleSupélec, University of Paris-Saclay
  Directors: Jean-Philippe Ovarlez, Guillaume Ginolhac

Table of contents

Context

Context in remote sensing

Many **image time series** from the **Earth**: SAR, multi/hyper spectral imaging, ...

Objective

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

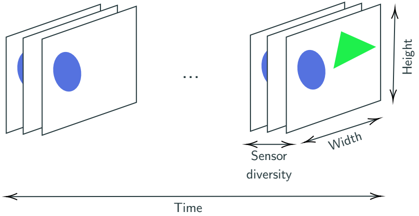


Figure 1: Multivariate image time series.

Applications

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

Context in neuroscience

Many **new datasets** are available in **neuroscience**: EEG, MEG, fMRI, ...

Objectives

- **Classify** brain signals into different **cognitive states** (sleep, wake, anesthesia, seizure, ...).
- **Regress** biomarkers (e.g. age) from brain signals.

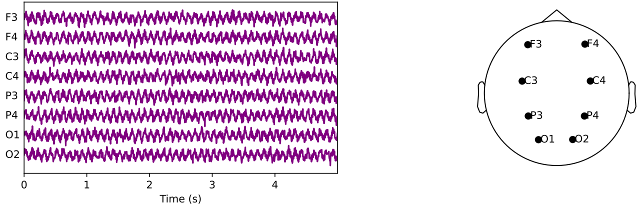


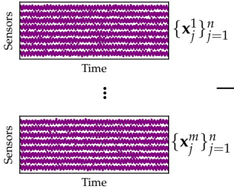
Figure 2: Multivariate EEG time series and the sensor locations.

Applications

Brain-computer interfaces, sleep monitoring, brain aging, ...

Classification and regression pipeline

Step 1: Data extraction



$\longrightarrow$

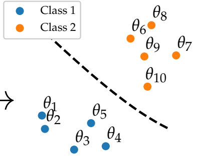
Step 2: Feature extraction

$\min_{\theta_i \in \mathcal{M}} \quad \mathcal{L}(\theta_i, \mathbf{x}_1^i, \cdots, \mathbf{x}_n^i)$

One θ_i characterizes
one batch of data to classify.

$\longrightarrow$

Step 3: Classification or regression



Assumption:

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

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: feature estimation

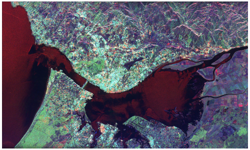


Figure 3: Example of a SAR image (from `nasa.gov`).

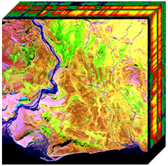


Figure 4: Example of a hyperspectral image (from `nasa.gov`).

Problematics:

- heterogeneous data because of outliers in biosignals,
- high dimension of hyperspectral images and MEG.

Objectives:

- develop **robust estimators**,
- develop **regularized/structured estimators**.

Step 3: classification / regression

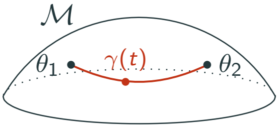


Figure 5: Divergence δ_γ: squared length of the curve γ.

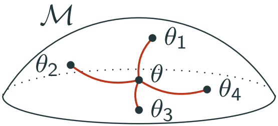


Figure 6: Center of mass of $\{\theta_i\}_{i=1}^{M}$.

Objectives:

Develop divergences that

- **respect the constraints** of $\mathcal{M}$,
- are **related to the chosen statistical distributions**.

Classification pipeline and Riemannian geometry

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

Step 2: maximum likelihood estimation

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

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

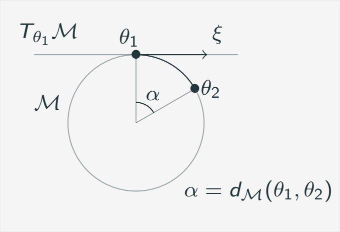
$$\underset{\theta \in \mathcal{M}}{\text{minimize}} \sum_i \delta(\theta, \theta_i)$$

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,
- **distance**: $d_{\mathcal{M}}(\theta_1, \theta_2) =$ length of γ.

Examples of $\mathcal{M}$: $\mathbb{R}^{p \times k}$, the sphere S^{p-1}, symmetric positive definite matrices $\mathcal{S}_p^{++}$, orthonormal $k-$frames $\mathrm{St}_{p,k}$, low-rank matrices, ...

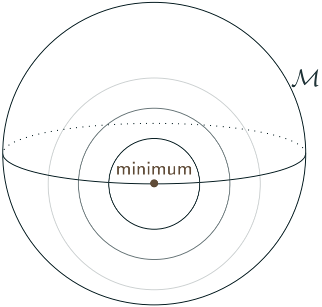
Nicolas Boumal. *An introduction to optimization on smooth manifolds.* Cambridge University Press, 2023

Optimization on a manifold

Optimization

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

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

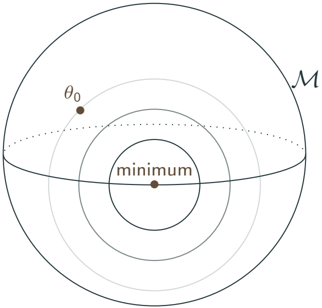


Optimization on a manifold

Optimization

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

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

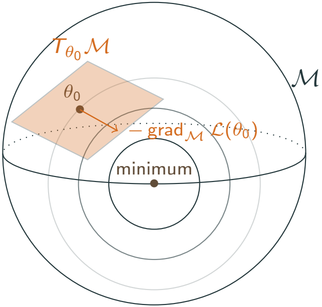


Optimization on a manifold

Optimization

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

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

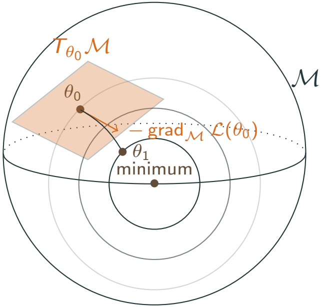


Optimization on a manifold

Optimization

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

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



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

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

where

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

is the Fisher information matrix.

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

Existing work: centered Gaussian

A well known geometry:

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

with the Fisher information metric:

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

Induced pipeline

Step 2:

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

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

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

Riemannian gradient descent to solve:

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

Alexandre Barachant et al. "Multiclass Brain–Computer Interface Classification by Riemannian Geometry". In: *IEEE Transactions on Biomedical Engineering* 59.4 (2012), pp. 920–928

Problematics

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

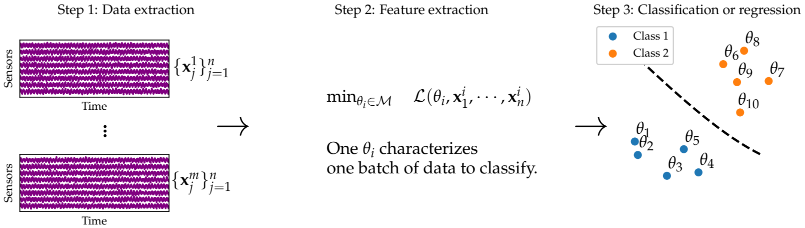
- $\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{\mu}, \tau_i \boldsymbol{\Sigma})$ for non-centered data and robustness,
- $\boldsymbol{x}_i \sim \mathcal{N}(\mathbf{0}, \tau_i \boldsymbol{U}\boldsymbol{U}^T + \boldsymbol{I}_p)$ for high dimensional data and robustness.

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 ?

Estimation and classification of non centered and heteroscedastic data

Non-centered mixtures of scaled Gaussian distributions



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.

A. Collas et al. "Riemannian optimization for non-centered mixture of scaled Gaussian distributions" in IEEE Trans. on Signal Processing 2023

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: $\prod_{i=1}^{n} \tau_i = 1$

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

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

A. Collas et al. "Riemannian optimization for non-centered mixture of scaled Gaussian distributions" in IEEE Trans. on Signal Processing 2023

Parameter space with a product metric

Product metric

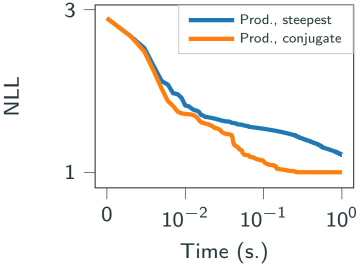
Let $\xi = (\boldsymbol{\xi_\mu}, \boldsymbol{\xi_\Sigma}, \boldsymbol{\xi_\tau})$, $\eta = (\boldsymbol{\eta_\mu}, \boldsymbol{\eta_\Sigma}, \boldsymbol{\eta_\tau})$ in the tangent space,

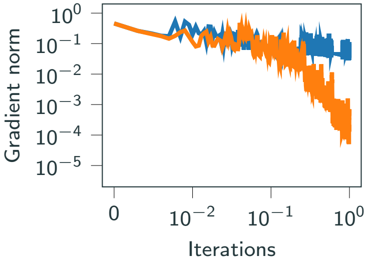
$$\langle \xi, \eta \rangle_\theta^{\mathcal{M}_{p,n}^{\text{Prod.}}} = \boldsymbol{\xi}_\mu^T \boldsymbol{\eta}_\mu + \text{Tr}(\boldsymbol{\Sigma}^{-1} \boldsymbol{\xi}_\Sigma \boldsymbol{\Sigma}^{-1} \boldsymbol{\eta}_\Sigma) + (\boldsymbol{\xi}_\tau \odot \boldsymbol{\tau}^{\odot -1})^T (\boldsymbol{\eta}_\tau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

where $\odot$ is the elementwise operator.

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

Slow in practice ...





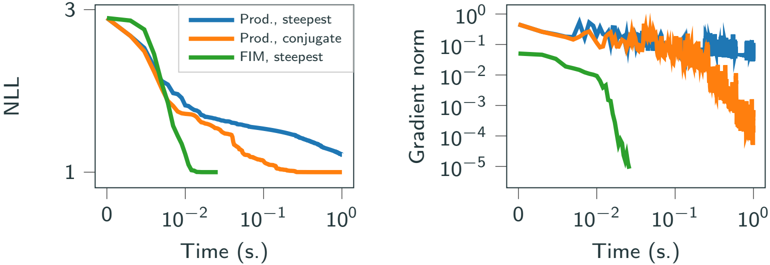
Fisher information metric of NC-MSGs

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

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

Derivation of the Riemannian gradient and a second order retraction.

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



Parameter estimation: existence

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

$$\mathcal{L}\left(\theta^{(\ell+1)}\right) < \mathcal{L}\left(\theta^{(\ell)}\right) \quad \text{and} \quad \theta^{(\ell)} \xrightarrow[\ell \to +\infty]{} \partial\theta$$

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

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 $\mathcal{M}_{p,n}$.

Example:

$$\mathcal{R}_\kappa(\theta) = \sum_{i,j}\left(\left(\tau_i\lambda_j\right)^{-1} - \kappa^{-1}\right)^2$$

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

Classification

KL divergence between NC-MSGs

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

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

Minimization of the KL variance:

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

Done with a Riemannian gradient descent.

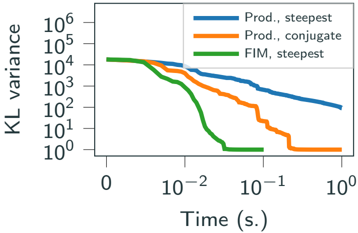


Figure 9: KL variance vs. iterations with $p = 10$, $n = 150$ and $M = 2$.

Breizhcrops dataset

Breizhcrops dataset[1]:

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

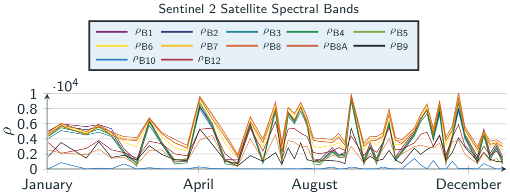


Figure 10: Reflectances ρ of a time series of **meadows**.

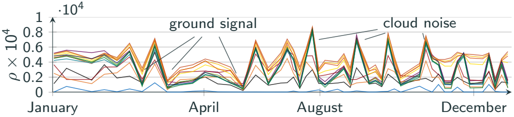
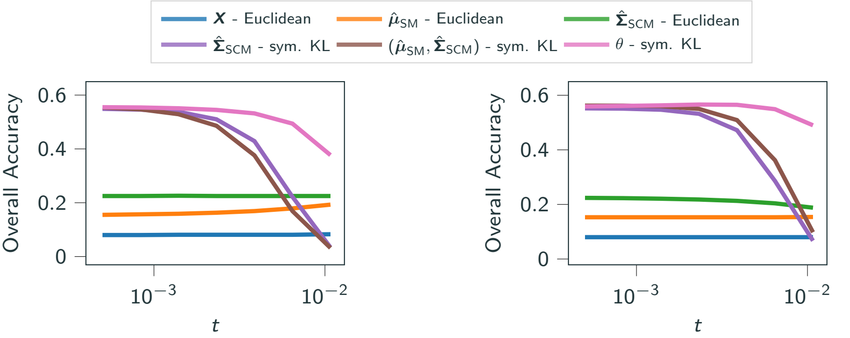


Figure 11: Reflectances ρ of a time series of **corn**.

[1] https://breizhcrops.org/

Application to the Breizhcrops dataset

Parameter estimation + classification with a *Nearest centroïd classifier*



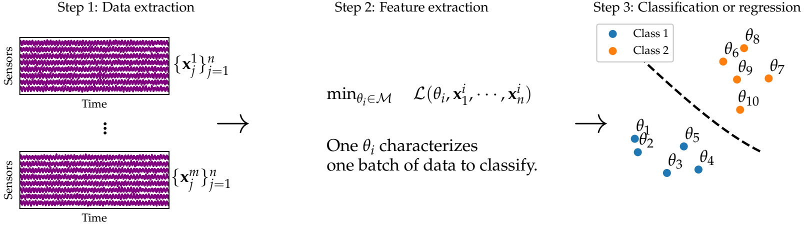
(a) Mean transformation: $\boldsymbol{x}_i \mapsto \boldsymbol{x}_i + \boldsymbol{\mu}(t)$ with $\boldsymbol{\mu}(0) = \boldsymbol{0}$

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

Figure 12: "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 $\beta = 10^{-11}$.

Probabilistic PCA from heteroscedastic signals

Study of a “low rank” statistical model



Statistical model

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

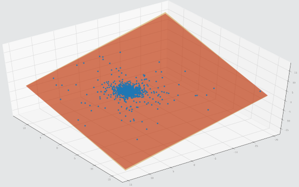
$$\boldsymbol{x}_i \sim \mathcal{N}(\mathbf{0}, \tau_i \boldsymbol{U}\boldsymbol{U}^T + \boldsymbol{I}_p)$$

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 $\boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_k) \perp \boldsymbol{n} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_p)$, $\boldsymbol{\tau} \in (\mathbb{R}_*^+)^n$, and $\boldsymbol{U} \in \mathbb{R}^{p\times k}$ s.t. $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}_k$.

Maximum likelihood estimation

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

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

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

A. Collas et al. “Probabilistic PCA From Heteroscedastic Signals: Geometric Framework and Application to Clustering” in IEEE Trans. on Signal Processing 2021

Study of a "low rank" statistical model: estimation

Fisher information metric

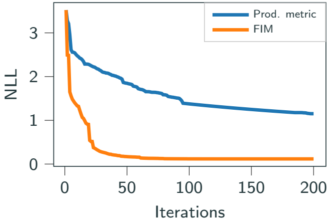
$\forall \xi = (\boldsymbol{\xi_U}, \boldsymbol{\xi_\tau}), \eta = (\boldsymbol{\eta_U}, \boldsymbol{\eta_\tau})$ in the tangent space

$$\langle \xi, \eta \rangle_\theta^{\mathrm{FIM}} = 2nc_{\boldsymbol{\tau}} \operatorname{Tr}\left(\boldsymbol{\xi}_{\boldsymbol{U}}^T \boldsymbol{\eta}_{\boldsymbol{U}}\right) + k \left(\boldsymbol{\xi_\tau} \odot (\mathbf{1} + \boldsymbol{\tau})^{\odot -1}\right)^T \left(\boldsymbol{\eta_\tau} \odot (\mathbf{1} + \boldsymbol{\tau})^{\odot -1}\right),$$

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

Derivation of the Riemannian gradient and of a retraction.

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



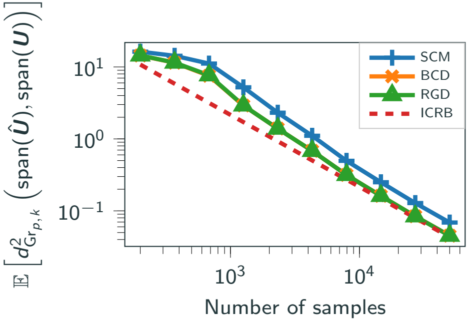
Study of a "low rank" statistical model: bounds

Intrinsic Cramér-Rao bounds

Study of the performance through intrinsic Cramér-Rao bounds:

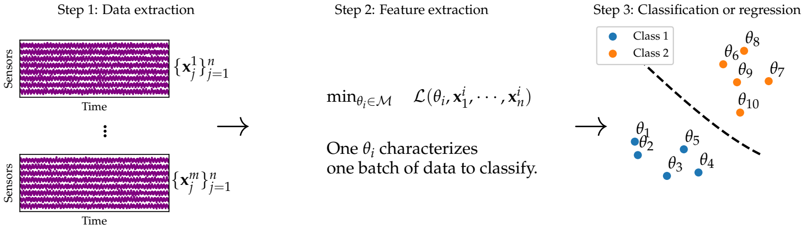
$$\overbrace{\mathbb{E}[d^2_{\mathrm{Gr}_{p,k}}(\mathrm{span}(\hat{\boldsymbol{U}}), \mathrm{span}(\boldsymbol{U}))]}^{\text{subspace estimation error}} \geq \frac{(p-k)k}{nc_{\boldsymbol{\tau}}} \approx \frac{(p-k)k}{n \times \mathrm{SNR}}$$

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



Aligning M/EEG data to enhance predictive regression modeling

Regression from M/EEG data



Statistical model

$\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n \in \mathbb{R}^p$,

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

Goal:

- regression from $\theta = \boldsymbol{\Sigma}$,
- correct distribution shifts between domains.

A. Mellot, **A. Collas** et al. "Harmonizing and aligning M/EEG datasets with covariance-based techniques to enhance predictive regression modeling" in Imaging Neuroscience MIT Press 2023

Generative model for regression with M/EEG

Linear instantaneous mixing model (from Maxwell's equations)

Signal $\boldsymbol{h}(t) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$:

$$\underbrace{\boldsymbol{h}(t)}_{\text{observed signal}} = \boldsymbol{A} \underbrace{\boldsymbol{\eta}(t)}_{\text{latent sources}}$$

Covariance matrix:

$$\boldsymbol{\Sigma} = \mathbb{E}_t \left[\boldsymbol{h}(t)\boldsymbol{h}(t)\right] = \boldsymbol{A} \operatorname{diag}(\boldsymbol{p}) \boldsymbol{A}^\top$$

with $\boldsymbol{p} = \operatorname{Var}(\boldsymbol{\eta}(t))$.

Regression model, $(\boldsymbol{\Sigma}_i, y_i)_{i=1}^m$

If $\exists \boldsymbol{\beta} \in \mathbb{R}^p$ s.t.

$$y_i = \boldsymbol{\beta}^\top \log(\boldsymbol{p}_i) + \varepsilon_i$$

then $\exists \boldsymbol{\beta}' \in \mathbb{R}^{p(p+1)/2}$ s.t.

$$y_i = \boldsymbol{\beta}'^\top \operatorname{vec}\left(\underbrace{\log(\overline{\boldsymbol{\Sigma}}^{-\frac{1}{2}} \boldsymbol{\Sigma}_i \overline{\boldsymbol{\Sigma}}^{-\frac{1}{2}})}_{\in T_{\boldsymbol{I}} \mathcal{S}_p^{++}} \right) + \varepsilon_i$$

where $\overline{\boldsymbol{\Sigma}}$ is the Riemannian mean of $\{\boldsymbol{\Sigma}_i\}_{i=1}^m$.

David Sabbagh et al. "Manifold-regression to predict from MEG/EEG brain signals without source modeling". In: *Advances in Neural Information Processing Systems* 32 (2019)

A. Mellot, **A. Collas** et al. "Harmonizing and aligning M/EEG datasets with covariance-based techniques to enhance predictive regression modeling" in Imaging Neuroscience MIT Press 2023

Statistics on the $\mathcal{S}_p^{++}$ manifold

Gaussian distribution on $\mathcal{S}_p^{++}$ and normalization

$$f\left(\boldsymbol{\Sigma}; \overline{\boldsymbol{\Sigma}}, \sigma^2\right) = \frac{1}{Z(\sigma)} \exp\left(-\frac{d^2_{\mathcal{S}_p^{++}}(\boldsymbol{\Sigma}, \overline{\boldsymbol{\Sigma}})}{2\sigma^2}\right)$$

with $Z(\sigma)$ the normalization constant.

Recenter-rescale operator: $\phi_{\overline{\boldsymbol{\Sigma}},\sigma^2}(\boldsymbol{\Sigma}) = \left(\overline{\boldsymbol{\Sigma}}^{-\frac{1}{2}} \boldsymbol{\Sigma} \overline{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\right)^{\frac{1}{\sigma}}$.

Salem Said et al. "Riemannian Gaussian Distributions on the Space of Symmetric Positive Definite Matrices". In: *IEEE Transactions on Information Theory* 63.4 (2017), pp. 2153–2170

Estimation with $(\boldsymbol{\Sigma}_i)_{i=1}^n \sim f\left(.; \overline{\boldsymbol{\Sigma}}, \sigma^2\right)$

$$\hat{\overline{\boldsymbol{\Sigma}}} = \underset{\boldsymbol{\Sigma} \in \mathcal{S}_p^{++}}{\arg\min} \frac{1}{n} \sum_{i=1}^n d^2_{\mathcal{S}_p^{++}}(\boldsymbol{\Sigma}, \boldsymbol{\Sigma}_i), \qquad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n d^2_{\mathcal{S}_p^{++}}(\hat{\overline{\boldsymbol{\Sigma}}}, \boldsymbol{\Sigma}_i)$$

Domain adaptation: for $\mathcal{D} \in \{\mathcal{S}, \mathcal{T}\}$

$$\boldsymbol{\Sigma}_i^{\mathcal{D}} \leftarrow \phi_{\hat{\overline{\boldsymbol{\Sigma}}}^{\mathcal{D}}, (\hat{\sigma}^2)^{\mathcal{D}}}\left(\boldsymbol{\Sigma}_i^{\mathcal{D}}\right)$$

A. Mellot, **A. Collas** et al. "Harmonizing and aligning M/EEG datasets with covariance-based techniques to enhance predictive regression modeling" in Imaging Neuroscience MIT Press 2023

Results on MEG data

Brain age prediction on the Cam-CAN dataset:

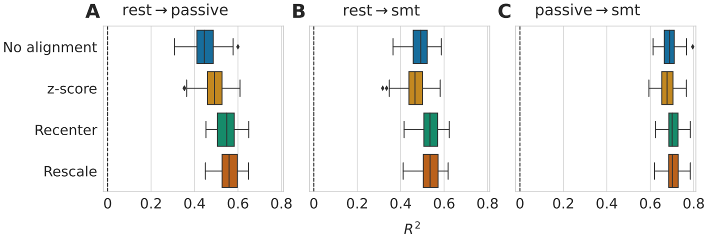
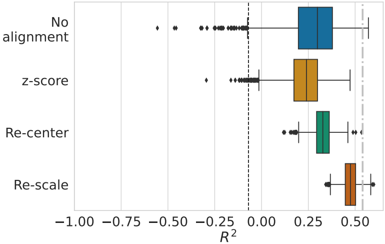


Figure 13: R^2 score on the Cam-CAN dataset (MEG), $n = 646$, 306 channels reduced to $p = 65$ after PCA and age range of $18 - 89$ years old.

A. Mellot, **A. Collas** et al. "Harmonizing and aligning M/EEG datasets with covariance-based techniques to enhance predictive regression modeling" in Imaging Neuroscience MIT Press 2023

Results on EEG datasets: LEMON → TUAB

Brain age prediction on the LEMON → TUAB datasets, regression on supervised SPoC components: $\mathrm{diag}(\log(\boldsymbol{W}_{\mathrm{SPoC}} \boldsymbol{\Sigma}_i \boldsymbol{W}_{\mathrm{SPoC}}^\top))$.



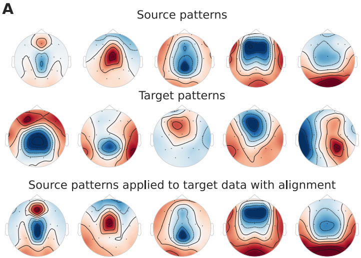


Figure 14: Left: R^2 score on LEMON ($n = 1385$) → TUAB ($n = 213$) (EEG), and $p = 15$ after PCA. Dashed line is the R^2 score of a cross-validation on target dataset. Right: topomaps of the SPoC patterns.

Many other results in the paper: simulations, rotation corrections, ...

A. Mellot, **A. Collas** et al. "Harmonizing and aligning M/EEG datasets with covariance-based techniques to enhance predictive regression modeling" in Imaging Neuroscience MIT Press 2023

Robust Geometric Metric Learning

Robust Geometric Metric Learning (RGML)

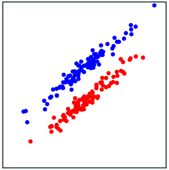
Let be a classification problem with K classes.

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.



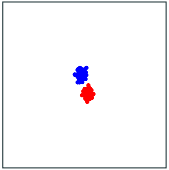
$\{\boldsymbol{x}_i\}$

Metric learning as covariance estimation

Proposed minimization problem:

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

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



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

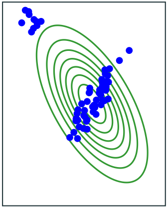
Robust Geometric Metric Learning (RGML)

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

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

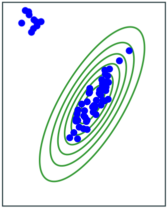
$$\text{minimized for } \boldsymbol{A}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} \boldsymbol{s}_{ki} \boldsymbol{s}_{ki}^T$$



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)

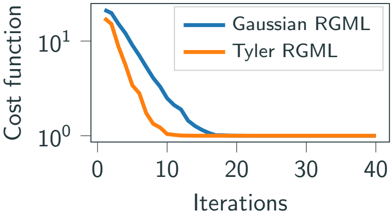
Riemannian metric

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

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

$\Longrightarrow$ strongly geodesically convexity of the minimization problem

$\Longrightarrow$ the Riemannian gradient descent is fast



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

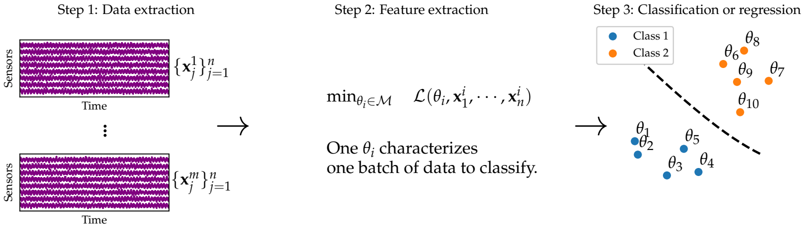
	Wine				Vehicle				Iris			
	$p = 13$, $n = 178$, $K = 3$				$p = 18$, $n = 846$, $K = 4$				$p = 4$, $n = 150$, $K = 3$			
Method	Mislabeling rate				Mislabeling rate				Mislabeling rate			
	0%	5%	10%	15%	0%	5%	10%	15%	0%	5%	10%	15%
Euclidean	30.12	30.40	31.40	32.40	38.27	38.58	39.46	40.35	3.93	4.47	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`

Open source software and conclusions

Open source software



pyCovariance: github.com/antoinecollas/pyCovariance

- _FeatureArray: custom data structure to store batch of points of product manifolds,
- implements statistical manifolds from this presentation,
- automatic computation of Riemannian centers of mass using exp/log or autodiff
- *K-means++* and *Nearest centroïd classifier* on any Riemannian manifolds,
- 15K lines of code, 96% of test coverage.

Open source software

pyManopt: github.com/pymanopt/pymanopt

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

Provide f smooth, choose a Riemannian manifold $\mathcal{M}$, and pyManopt does the rest !

Geomstats: information geometry module github.com/geomstats/geomstats

Choose a statistical manifold $\mathcal{M}$ (or give a p.d.f. !), and Geomstats does the rest: geodesics, log, exp, barycenter, leaning: K-means, KNN, PCA, etc...

A. Le Brigant, J. Deschamps, **A. Collas** and N. Miolane, "Parametric information geometry with the package Geomstats" ACM Transactions on Mathematical Software 2023.

References

References

Barachant, Alexandre, Stéphane Bonnet, Marco Congedo, and Christian Jutten. "Multiclass Brain–Computer Interface Classification by Riemannian Geometry". In: *IEEE Transactions on Biomedical Engineering* 59.4 (2012), pp. 920–928.

Boumal, Nicolas. *An introduction to optimization on smooth manifolds*. Cambridge University Press, 2023.

Collas, **Antoine**, Florent Bouchard, Arnaud Breloy, Guillaume Ginolhac, Chengfang Ren, and Jean-Philippe Ovarlez. "Probabilistic PCA from heteroscedastic signals: geometric framework and application to clustering". In: *IEEE Transactions on Signal Processing* 69 (2021), pp. 6546–6560.

Collas, **Antoine**, Arnaud Breloy, Chengfang Ren, Guillaume Ginolhac, and Jean-Philippe Ovarlez. "Riemannian optimization for non-centered mixture of scaled Gaussian distributions". In: *IEEE Transactions on Signal Processing* (2023).

Le Brigant, Alice, Jules Deschamps, **Antoine Collas**, and Nina Miolane. "Parametric information geometry with the package Geomstats". In: *ACM Transactions on Mathematical Software* (2022).

Mellot, Apolline, **Antoine Collas**, Pedro L. C. Rodrigues, Denis Engemann, and Alexandre Gramfort. "Harmonizing and aligning M/EEG datasets with covariance-based techniques to enhance predictive regression modeling". In: *Imaging Neuroscience* (Nov. 2023). ISSN: 2837-6056. DOI: 10.1162/imag_a_00040.

Sabbagh, David, Pierre Ablin, Gaël Varoquaux, Alexandre Gramfort, and Denis A Engemann. "Manifold-regression to predict from MEG/EEG brain signals without source modeling". In: *Advances in Neural Information Processing Systems* 32 (2019).

Said, Salem, Lionel Bombrun, Yannick Berthoumieu, and Jonathan H. Manton. "Riemannian Gaussian Distributions on the Space of Symmetric Positive Definite Matrices". In: *IEEE Transactions on Information Theory* 63.4 (2017), pp. 2153–2170.

Riemannian geometry for statistical estimation and learning: applications to remote sensing and M/EEG

Antoine Collas

OPIS seminar