Classifying histograms using information geometry of beta distributions

Alice Le Brigant
SAMM - Université Paris 1

10ème Biennale Française des Mathématiques Appliquées et Industrielles
21-25 juin 2021

References

A. L., N. Guigui, S. Rebbah, S. Puechmorel,
Classifying histograms of medical data using information geometry of beta distributions.
arXiv preprint arXiv :2006.04511 (2020).

A. L., S. Preston, S. Puechmorel,
Fisher-Rao geometry of Dirichlet distributions.
Differential Geometry and its Applications 74 (2021) : 101702.

Table of contents

Table of contents

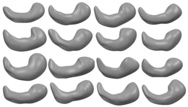
Non Euclidean data

Statistics have been developed for Euclidean data

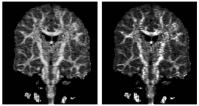
$$x_1, \ldots, x_n \in \mathbb{R}^d, \quad \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

In many applications, the data of interest do not belong to a vector space

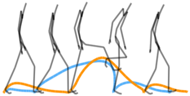
- Geostatistics $\rightarrow$ spherical data
- Diffusion tensor images $\rightarrow$ SPD matrices
- Character animations $\rightarrow$ rotation matrices
- Images $\rightarrow$ histograms, probability distributions
- Shapes of curves, surfaces $\rightarrow$ immersions, submanifolds



Fletcher et al. 2004



Fletcher et al. 2006



Celledoni et al. 2016

Riemannian geometric statistics

All these objects can be seen as elements of differentiable manifolds

- Open sets of vector spaces (SPD matrices, immersions)
- Embedded submanifolds (sphere)

Riemannian geometric statistics

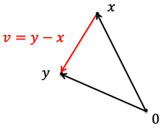
All these objects can be seen as elements of differentiable manifolds

- Open sets of vector spaces (SPD matrices, immersions)
- Embedded submanifolds (sphere)

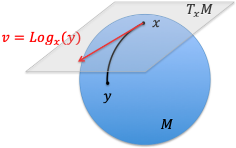
To do statistics on these objects, one can use Riemannian geometry.

- Euclidean scalar product $\rightarrow$ Riemannian metric
- straight lines $\rightarrow$ geodesics
- Euclidean distance $\rightarrow$ geodesic distance
- addition / subtraction $\rightarrow$ Riemannian exponential / logarithm

$$
\begin{aligned}
x + v \quad &\rightarrow \quad \mathrm{Exp}_x(v) \\
x - y \quad &\rightarrow \quad \mathrm{Log}_x y
\end{aligned}
$$



Euclidean space



Manifold

Riemannian geometric statistics

Riemannian geometry is a convenient framework to generalize usual statistical notions and data processing algorithms.

Riemannian geometric statistics

Riemannian geometry is a convenient framework to generalize usual statistical notions and data processing algorithms.

Example : the **Fréchet mean** $x_m = \underset{y \in M}{\text{argmin}} \frac{1}{n} \sum_{i=1}^{n} d(x_i, y)^2.$

- In general, it is not unique
- Unicity under certain conditions on injectivity radius of the data manifold
- Or for Hadamard manifolds (geodesically complete and negative curvature)
- In a Riemannian manifold, it can be computed via a gradient descent based on the Riemannian Logarithm

$$F(y) = \frac{1}{n} \sum_{i=1}^{n} d(x_i, y)^2, \qquad \nabla_y F \propto - \sum_{i=1}^{n} \log_y(x_i).$$

Riemannian geometric statistics

Riemannian geometry is a convenient framework to generalize usual statistical notions and data processing algorithms.

Example : the **Fréchet mean** $x_m = \underset{y \in M}{\text{argmin}} \frac{1}{n} \sum_{i=1}^{n} d(x_i, y)^2.$

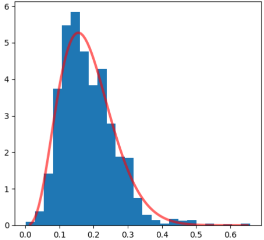
- In general, it is not unique
- Unicity under certain conditions on injectivity radius of the data manifold
- Or for Hadamard manifolds (geodesically complete and negative curvature)
- In a Riemannian manifold, it can be computed via a gradient descent based on the Riemannian Logarithm

$$F(y) = \frac{1}{n} \sum_{i=1}^{n} d(x_i, y)^2, \qquad \nabla_y F \propto - \sum_{i=1}^{n} \log_y(x_i).$$

Knowledge of the **data manifold geometry** (existence of geodesics, curvature...) is important to apply geometric statistics on it.

Histogram data

Here we focus on histogram data.



- Each data point is a histogram of bounded values, e.g. measurements corresponding to a patient
- Lack of a common representation space : a correspondence between measurements across patients may not exist
- Bounded values : a beta distribution is fitted to each histogram
- Idea : do geometric statistics on dataset of beta distributions

Table of contents

Information geometry and the Fisher-Rao distance

- Information geometry : geometric approach to parametric statistics based on the **Fisher information**. Given a parametric model $\{p_\theta \mu, \theta \in \Theta\}$,

$$I(\theta) = \mathbb{E}_\theta[\partial_\theta \ell_\theta(X) \cdot \partial_\theta \ell_\theta(X)^t], \quad \ell_\theta = \log p_\theta.$$

Information geometry and the Fisher-Rao distance

- Information geometry : geometric approach to parametric statistics based on the **Fisher information**. Given a parametric model $\{p_\theta \mu, \theta \in \Theta\}$,

$$I(\theta) = \mathbb{E}_\theta[\partial_\theta \ell_\theta(X) \cdot \partial_\theta \ell_\theta(X)^t], \quad \ell_\theta = \log p_\theta.$$

- Rao (1945) : the Fisher information can be used to define a **Riemannian metric** on the parameter space Θ, i.e. a local scalar product on each tangent space

$$\langle u, v\rangle_\theta = {}^t u I(\theta) v, \quad u, v \in T_\theta\Theta \simeq \mathbb{R}^d.$$

Information geometry and the Fisher-Rao distance

- Information geometry : geometric approach to parametric statistics based on the **Fisher information**. Given a parametric model $\{p_\theta \mu, \theta \in \Theta\}$,

$$I(\theta) = \mathbb{E}_\theta[\partial_\theta \ell_\theta(X) \cdot \partial_\theta \ell_\theta(X)^t], \quad \ell_\theta = \log p_\theta.$$

- Rao (1945) : the Fisher information can be used to define a **Riemannian metric** on the parameter space Θ, i.e. a local scalar product on each tangent space

$$\langle u, v \rangle_\theta = {}^t u I(\theta) v, \quad u, v \in T_\theta\Theta \simeq \mathbb{R}^d.$$

- The induced *geodesic distance* is called the **Fisher-Rao distance**

$$\mathrm{dist}(\theta_0, \theta_1) = \inf_{\theta(0)=\theta_0, \theta(1)=\theta_1} \ell(\theta) \quad \text{where} \quad \ell(\theta) = \int_0^1 \|\theta'(t)\|_{\theta(t)} dt.$$

Information geometry and the Fisher-Rao distance

- Information geometry : geometric approach to parametric statistics based on the **Fisher information**. Given a parametric model $\{p_\theta \mu, \theta \in \Theta\}$,

$$I(\theta) = \mathbb{E}_\theta[\partial_\theta \ell_\theta(X) \cdot \partial_\theta \ell_\theta(X)^t], \quad \ell_\theta = \log p_\theta.$$

- Rao (1945) : the Fisher information can be used to define a **Riemannian metric** on the parameter space Θ, i.e. a local scalar product on each tangent space

$$\langle u, v \rangle_\theta = {}^t u I(\theta) v, \quad u, v \in T_\theta \Theta \simeq \mathbb{R}^d.$$

- The induced *geodesic distance* is called the **Fisher-Rao distance**

$$\mathrm{dist}(\theta_0, \theta_1) = \inf_{\theta(0)=\theta_0, \theta(1)=\theta_1} \ell(\theta) \quad \text{where} \quad \ell(\theta) = \int_0^1 \|\theta'(t)\|_{\theta(t)} dt.$$

- The Fisher-Rao metric is **invariant** to
  - diffeomorphic parametrization change $\varphi : \theta \mapsto \tilde{\theta}$.
  - transformation of the statistical model by a sufficient statistic (Chentsov).

Information geometry and the Fisher-Rao distance

- **Geodesics** are local minimizers of the geodesic distance and curves $t \mapsto \gamma(t)$ with zero acceleration
$$\nabla_{\dot{\gamma}}\dot{\gamma} = 0$$
where ∇ is the *Levi-Civita connection* and allows to take "intrinsic" derivatives of vector fields $\rightarrow$ geodesics are **solutions of an ODE**.

Information geometry and the Fisher-Rao distance

- **Geodesics** are local minimizers of the geodesic distance and curves $t \mapsto \gamma(t)$ with zero acceleration

$$\nabla_{\dot\gamma}\dot\gamma = 0$$

where ∇ is the *Levi-Civita connection* and allows to take "intrinsic" derivatives of vector fields $\rightarrow$ geodesics are **solutions of an ODE**.

- If the manifold is complete, any two points can be linked by a minimizing geodesic (Hopf-Rinov theorem), i.e.

$$\mathrm{dist}(\theta_0, \theta_1) = \ell(\gamma) \quad \text{with} \quad \gamma \quad \text{s.t.} \quad \nabla_{\dot\gamma}\dot\gamma = 0.$$

Information geometry and the Fisher-Rao distance

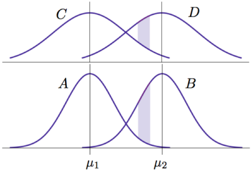
- **Geodesics** are local minimizers of the geodesic distance and curves $t \mapsto \gamma(t)$ with zero acceleration
$$\nabla_{\dot\gamma}\dot\gamma = 0$$
where ∇ is the *Levi-Civita connection* and allows to take "intrinsic" derivatives of vector fields $\rightarrow$ geodesics are **solutions of an ODE**.

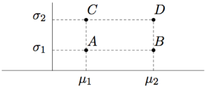
- If the manifold is complete, any two points can be linked by a minimizing geodesic (Hopf-Rinov theorem), i.e.
$$\mathrm{dist}(\theta_0, \theta_1) = \ell(\gamma) \quad \text{with} \quad \gamma \quad \text{s.t.} \quad \nabla_{\dot\gamma}\dot\gamma = 0.$$

- If the manifold has **negative curvature** and is simply connected, the minimizing geodesic is unique (Cartan-Hadamard theorem).

Examples

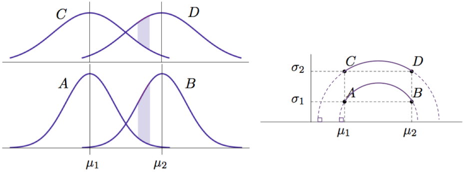
- **Univariate normal distributions** (Atkinson & Mitchell 1981)
  $p_\theta = \mathcal{N}(\mu, \sigma)$, then $\Theta = \mathbb{R} \times \mathbb{R}_+^*$ equipped with the Fisher-Rao metric is the Poincaré half-plane. Geodesics are half-circles orthogonal to the x-axis.





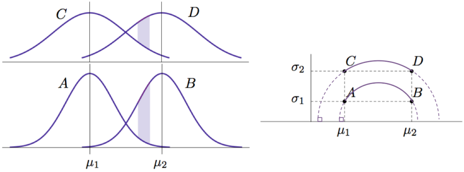
Examples

- **Univariate normal distributions** (Atkinson & Mitchell 1981)
  $p_\theta = \mathcal{N}(\mu, \sigma)$, then $\Theta = \mathbb{R} \times \mathbb{R}_+^*$ equipped with the Fisher-Rao metric is the Poincaré half-plane. Geodesics are half-circles orthogonal to the x-axis.



Examples

- **Univariate normal distributions** (Atkinson & Mitchell 1981)
  $p_\theta = \mathcal{N}(\mu, \sigma)$, then $\Theta = \mathbb{R} \times \mathbb{R}_+^*$ equipped with the Fisher-Rao metric is the Poincaré half-plane. Geodesics are half-circles orthogonal to the x-axis.

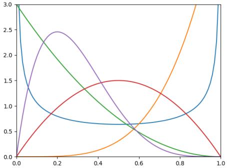


- Others :
  - Special cases of multivariate normal distributions (Atkinson & Mitchell 1981, Skovgaard 1984)
  - Gamma distributions (Lauritzen 1987, Arwini & Dodsen 2008)
  - Power inverse Gaussian distributions (Zhang, Sun, Zhong, 2007)
  - Von Mises-Fisher and location-scale models (Said et al. 2019)
  - Generalized Gamma distributions (Rebbah, Nicol, Puechmorel 2019)

Fisher-Rao metric on beta distributions

- Parameter space is $\Theta = \{(x,y), x > 0, y > 0\}$.
- The beta p.d.f. of parameter $\theta \in \Theta$ is

$$p_\theta(u) = \frac{\Gamma(x+y)}{\Gamma(x)\Gamma(y)} u^{x-1}(1-u)^{y-1}, \quad \forall u \in [0,1].$$



- The Fisher-Rao metric is defined for any $v = (v_x, v_y) \in T_{(x,y)}\Theta$ by

$$\|v\|^2 = \psi'(x)v_x^2 + \psi'(y)v_y^2 - \psi'(x+y)(v_x+v_y)^2,$$

where $\psi = \Gamma'/\Gamma$ and ψ' are the digamma and trigamma functions.

- Parameter space Θ + Fisher-Rao metric = **beta manifold**.

Fisher-Rao metric on beta distributions

- The geodesics $t \mapsto (x(t), y(t))$ can be found numerically by solving the ODE

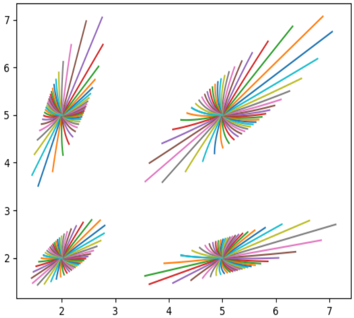
$$\ddot{x} + a(x,y)\dot{x}^2 + b(x,y)\dot{x}\dot{y} + c(x,y)\dot{y}^2 = 0$$
$$\ddot{y} + a(y,x)\dot{y}^2 + b(y,x)\dot{x}\dot{y} + c(y,x)\dot{x}^2 = 0$$

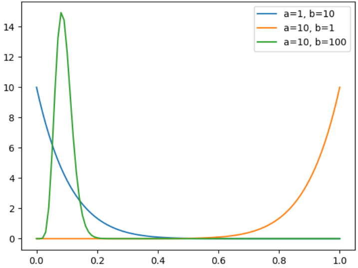
  with

$$a(x,y) = \frac{\psi''(x)\psi'(y) - \psi''(x)\psi'(x+y) - \psi'(y)\psi''(x+y)}{2d(x,y)},$$
$$b(x,y) = -\frac{\psi''(x+y)\psi'(y)}{d(x,y)},$$
$$c(x,y) = \frac{\psi''(y)\psi'(x+y) - \psi'(y)\psi''(x+y)}{2d(x,y)},$$
$$d(x,y) = \psi'(x)\psi'(y) - \psi'(x+y)(\psi'(x) + \psi'(y)).$$

- Initial value problem : fix $x(0), y(0), \dot{x}(0), \dot{y}(0)$
  Boundary value problem : fix $x(0), y(0), x(1), y(1)$.

Fisher-Rao metric on beta distributions





	Euclidean	Fisher-Rao
$\mathrm{dist}((1,10),(10,1))$	12,7	4,2
$\mathrm{dist}((1,10),(10,100))$	99,4	1,8

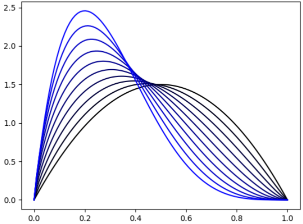
Fisher-Rao metric on beta distributions

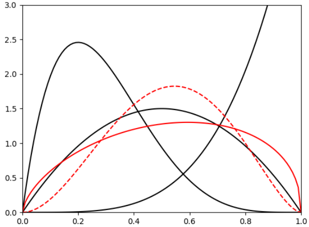
Computation of geodesics allows to compute

- distances
- interpolations
- barycenters

Questions :

- Are the geodesics always defined ? Unique ?
- **What is the curvature of the beta manifold ?**



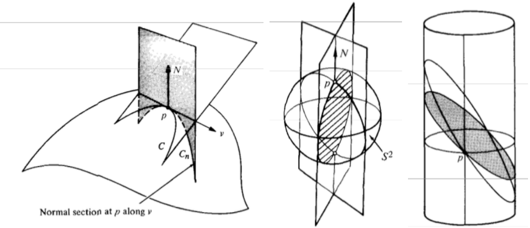


Gauss curvature of a surface $S \subset \mathbb{R}^3$

- The normal to S defines the *Gauss map* $N : S \to \mathbb{R}^3$.
- The *shape operator* $\Sigma_p(v) = -dN_p(v)$ is symmetric and defines the *second fundamental form*

$$\mathrm{II}_p(v) = \langle \Sigma_p(v), v \rangle.$$

  It is the signed curvature of the normal section of S at p in direction v.
- The *principal curvatures* k_1 and k_2 are the extremal eigenvalues of the *shape operator* Σ_p
- *Mean (extrinsic) curvature* : $\frac{k_1+k_2}{2}$ vs *Gauss (intrinsic) curvature* : $k_1 k_2$



Sectional curvature of a Riemannian manifold M

- **Sectional curvature** of a plane $\sigma \subset T_pM$ spanned by vectors e_1, e_2

$$K(\sigma) = K(e_1, e_2) = \frac{\langle R(e_1, e_2)e_1, e_2 \rangle}{\langle e_1, e_1 \rangle \langle e_2, e_2 \rangle - \langle e_1, e_2 \rangle^2}$$

where $R(U,V)W = \nabla_U \nabla_V W - \nabla_V \nabla_U W - \nabla_{[U,V]} W$ is the curvature tensor.

Sectional curvature of a Riemannian manifold M

- **Sectional curvature** of a plane $\sigma \subset T_pM$ spanned by vectors e_1, e_2

$$K(\sigma) = K(e_1, e_2) = \frac{\langle R(e_1, e_2)e_1, e_2\rangle}{\langle e_1, e_1\rangle\langle e_2, e_2\rangle - \langle e_1, e_2\rangle^2}$$

where $R(U,V)W = \nabla_U\nabla_V W - \nabla_V\nabla_U W - \nabla_{[U,V]}W$ is the curvature tensor.

- When M is a **hypersurface** in $\bar{M}$, then the Gauss map $p \mapsto N(p)$ is defined and

$$K(U,V) - \bar{K}(U,V) = \frac{\langle\Sigma(U), U\rangle\langle\Sigma(V), V\rangle - \langle\Sigma(U), V\rangle^2}{\langle U, U\rangle\langle V, V\rangle - \langle U, V\rangle^2}.$$

where $\Sigma_p(v) = -\bar{\nabla}_v N$ is the *shape operator*.

Sectional curvature of a Riemannian manifold M

- **Sectional curvature** of a plane $\sigma \subset T_pM$ spanned by vectors e_1, e_2

$$K(\sigma) = K(e_1, e_2) = \frac{\langle R(e_1, e_2)e_1, e_2\rangle}{\langle e_1, e_1\rangle\langle e_2, e_2\rangle - \langle e_1, e_2\rangle^2}$$

where $R(U,V)W = \nabla_U\nabla_V W - \nabla_V\nabla_U W - \nabla_{[U,V]}W$ is the curvature tensor.

- When M is a **hypersurface** in $\bar{M}$, then the Gauss map $p \mapsto N(p)$ is defined and

$$K(U,V) - \bar{K}(U,V) = \frac{\langle\Sigma(U), U\rangle\langle\Sigma(V), V\rangle - \langle\Sigma(U), V\rangle^2}{\langle U, U\rangle\langle V, V\rangle - \langle U, V\rangle^2}.$$

where $\Sigma_p(v) = -\bar{\nabla}_v N$ is the *shape operator*.

- If $\bar{K} = 0$ and for an o.n.b. of eigenvectors $e_1, \ldots, e_n$ of Σ_p associated to the eigenvalues $k_1, \ldots, k_n$ (*principal curvatures*), we get

$$K(e_i, e_j) = k_i k_j.$$

Sectional curvature of the beta manifold

- Only one sectional curvature $K(\theta)$ at each point $\theta = (x, y)$ $(\dim = 2)$

Sectional curvature of the beta manifold

- Only one sectional curvature $K(\theta)$ at each point $\theta = (x, y)$ $(\dim = 2)$
- At point (x, y), the sectional curvature can be computed from coordinate vectors e_x, e_y :

$$K(x,y) = \frac{\psi''(x)\psi''(y)\psi''(x+y)}{4d(x,y)^2}\left(\frac{\psi'(x)}{\psi''(x)} + \frac{\psi'(y)}{\psi''(y)} - \frac{\psi'(x+y)}{\psi''(x+y)}\right)$$

where $d(x, y)$ is the determinant of the metric matrix.

Sectional curvature of the beta manifold

- Only one sectional curvature $K(\theta)$ at each point $\theta = (x, y)$ $(\dim = 2)$
- At point (x, y), the sectional curvature can be computed from coordinate vectors e_x, e_y :

$$K(x,y) = \frac{\psi''(x)\psi''(y)\psi''(x+y)}{4d(x,y)^2}\left(\frac{\psi'(x)}{\psi''(x)} + \frac{\psi'(y)}{\psi''(y)} - \frac{\psi'(x+y)}{\psi''(x+y)}\right)$$

  where $d(x, y)$ is the determinant of the metric matrix.
- $\psi'' \leq 0$ and (Yang, 2017) ψ'/ψ'' is subadditive $\Rightarrow K \leq 0$

Sectional curvature of the beta manifold

- Only one sectional curvature $K(\theta)$ at each point $\theta = (x, y)$ $(\dim = 2)$
- At point (x, y), the sectional curvature can be computed from coordinate vectors e_x, e_y :

$$K(x,y) = \frac{\psi''(x)\psi''(y)\psi''(x+y)}{4d(x,y)^2}\left(\frac{\psi'(x)}{\psi''(x)} + \frac{\psi'(y)}{\psi''(y)} - \frac{\psi'(x+y)}{\psi''(x+y)}\right)$$

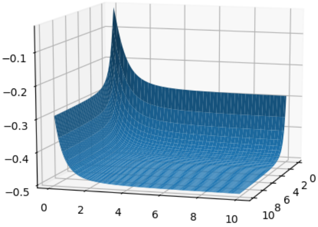
  where $d(x, y)$ is the determinant of the metric matrix.
- $\psi'' \leq 0$ and (Yang, 2017) ψ'/ψ'' is subadditive $\Rightarrow K \leq 0$

Theorem

The beta manifold has everywhere negative sectional curvature.

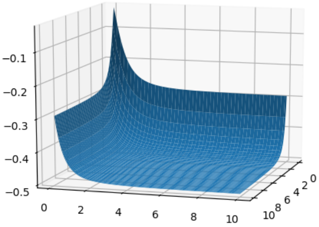
Sectional curvature of the beta manifold

- Conjecture : $K \geq -1/2$



Sectional curvature of the beta manifold

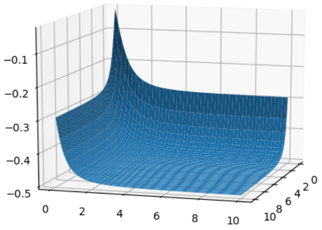
- Conjecture : $K \geq -1/2$



- One can show that the beta manifold is complete, and simply connected, so it is a **Hadamard manifold**
  $\rightarrow$ existence and unicity of minimizing geodesics
  $\rightarrow$ existence and unicity of the Fréchet mean.

Sectional curvature of the beta manifold

- Conjecture : $K \geq -1/2$



- One can show that the beta manifold is complete, and simply connected, so it is a **Hadamard manifold**
  $\rightarrow$ existence and unicity of minimizing geodesics
  $\rightarrow$ existence and unicity of the Fréchet mean.

- Algorithms based on distance and barycenter computations are well defined on the beta manifold.

Table of contents

Histograms of medical data

Example 1 : Area strain (AS) of the right ventricle

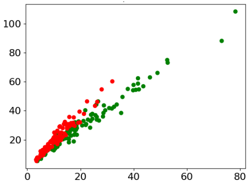
- Goal : observe local stretching of the area of the right ventricle (RV) to assess cardiac function
- Data : area change of each triangular cell of a 3d mesh of the RV
- This gives a distribution of n values in $[p, q]$ with constants p, q.
- Renormalize to obtain values in $[0, 1]$.

Example 2 : Cortical thickness (CTh) maps

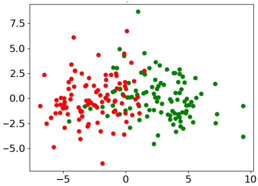
- Goal : measure brain atrophy for the early diagnosis of Alzheimer's disease
- Data : n cortical thickness (CTh) MRI measures along the whole cortical ribbon
- Each measure is made for a 1mm voxel
- Normalize values by maximal value among population

Methodology

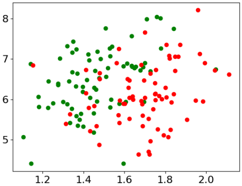
- Each dataset is divided in 2 classes : diseased and controls.
- 1 patient $\leftrightarrow$ 1 histogram $\leftrightarrow$ beta parameters (x, y)
- We perform both supervised (KNN, supervised Kmeans) and unsupervised (Kmeans) classification in the following representation spaces :
  - beta parameter space + Fisher-Rao metric
  - beta parameter space + Euclidean metric
  - original data space + Euclidean metric (only for AS)
  - 2-dimensional space obtained by PCA + Euclidean metric (only for AS)



AS beta representation



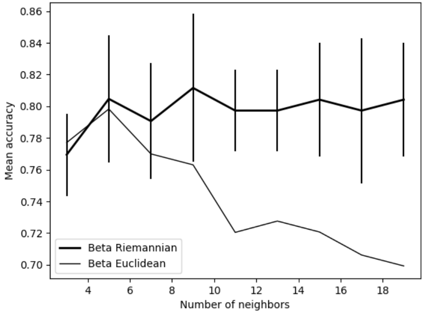
AS PCA representation



CTh beta representation

Results

CTh data



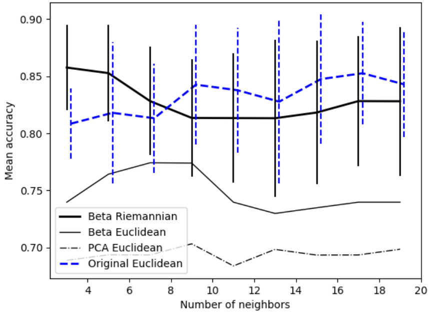
Mean accuracy of KNN on 5-fold cross-validation

	Beta	
	Euclidean	Riemannian
KNN	0.77 (0.05)	**0.79** (0.04)
SKM	0.66 (0.10)	**0.80** (0.05)
KM	0.61	**0.82**

Classification mean (and standard deviation) accuracy on 5-fold cross-validation

Results

AS data



Mean accuracy of KNN on 5-fold cross-validation

	Original Euclidean	PCA Euclidean	Beta Euclidean	Beta Riemannian
KNN	0.81 (0.05)	0.67 (0.32)	0.77 (0.09)	**0.83** (0.05)
SKM	**0.85** (0.03)	0.72 (0.28)	0.66 (0.06)	0.81 (0.06)
KM	**0.84**	**0.84**	0.60	0.80

Classification mean (and standard deviation) accuracy on 5-fold cross-validation

Table of contents

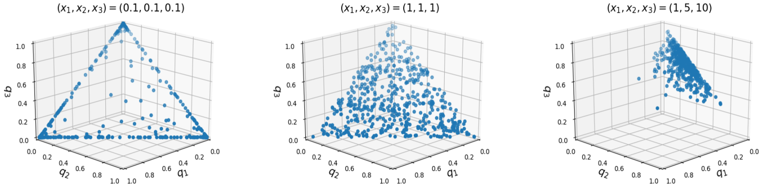
Dirichlet distributions

- Dirichlet distributions are probability distributions on the $(n-1)$-dimensional probability simplex

$$\Delta_n = \{q = (q_1, \ldots, q_n) \in \mathbb{R}^n, \sum_{i=1}^{n} q_i = 1, q_i \geq 0, i = 1, \ldots, n\}.$$

- Parameter space is $\Theta = \{(x_1, \ldots, x_n), x_1, \ldots, x_n > 0\}$ and p.d.f. is

$$f_n(q|x_1, \ldots, x_n) = \frac{\Gamma(x_1 + \ldots + x_n)}{\Gamma(x_1) \ldots \Gamma(x_n)} q_1^{x_1 - 1} \ldots q_n^{x_n - 1}.$$



The Dirichlet manifold

- Notations : $f = \frac{1}{\psi'}$, $F = \frac{\psi'}{\psi''}$.
- The **Dirichlet manifold** is $\Theta = (\mathbb{R}_+^*)^n$ equipped with the Fisher-Rao metric

$$ds^2 = \frac{dx_1^2}{f(x_1)} + \cdots + \frac{dx_n^2}{f(x_n)} - \frac{(dx_1 + \cdots + dx_n)^2}{f(x_1 + \cdots + x_n)}$$

- The sectional curvature of the plane spanned by coordinate vectors e_i, e_j is

$$K_{ij}(x) = \frac{F(x_i)F(x_j)F(\sum x_k)}{4(f(\sum x_k) - \sum f(x_k))^2}\left(F(x_i) + F(x_j) - F\left(\sum x_k\right)\right)$$

  $F \leq 0$ is subadditive $\Rightarrow K_{ij} \leq 0$ everywhere.

- But in dimension $n > 2$, there are other sectional curvatures to consider !

The Dirichlet manifold

There is an isometric embedding

$$\Phi : M \to L^{n+1},$$
$$(x_1, \dots, x_n) \mapsto (\eta(x_1), \dots, \eta(x_n), \eta(x_1 + \dots + x_n))$$

with
$$\eta : \mathbb{R}_+ \to \mathbb{R}, \quad \eta(x) = \int_1^x \frac{dr}{\sqrt{f(r)}}$$

between M and the $(n+1)$-dimensional flat Minkowski space $L^{n+1} = (\mathbb{R}^{n+1}, ds_L^2)$, where

$$ds_L^2 = dy_1^2 + \dots + dy_n^2 - dy_{n+1}^2.$$

Proposition

$S = \Phi(M)$ is a codimension 1 submanifold (hypersurface) of L^{n+1} given by the graph of

$$y_{n+1} = \eta(\xi(y_1) + \dots + \xi(y_n)), \quad y_i > 0,$$

where $\xi = \eta^{-1}$. On this submanifold the metric is positive-definite and thus Riemannian.

Curvature of the Dirichlet manifold

Theorem

The Dirichlet manifold has everywhere negative sectional curvature.

Elements of proof :

- A basis of tangent vectors of T_yS is defined by

$$e_i = \frac{\partial}{\partial y_i} + \sqrt{\frac{f\circ\xi(y_i)}{f\circ\xi(y_{n+1})}}\frac{\partial}{\partial y_{n+1}}, \quad i = 1,\dots,n,$$

- The shape operator of $S = \Phi(M)$ has components

$$\langle \nabla_{e_i} N, e_j \rangle = \frac{1}{2\sqrt{f(t) - \sum_{\ell=1}^n f(x_\ell)}} \left(f'(x_i)\delta_{ij} - \frac{f'(t)}{f(t)}\sqrt{f(x_i)f(x_j)} \right).$$

- The second fundamental form $\mathrm{II}(U,V) = \langle -\nabla_U N, V \rangle$ is positive-definite.
- The sectional curvature is, for U, V tangent to the submanifold

$$K(U,V) = \bar{K}(U,V) - \frac{\mathrm{II}(U,U)\mathrm{II}(V,V) - \mathrm{II}(U,V)^2}{\langle U,U\rangle\langle V,V\rangle - \langle U,V\rangle^2},$$

The result follows from $\bar{K} = 0$ and the Cauchy-Schwarz inequality.

Conclusion

Future work :

- Bigger datasets
- Numerical efficiency

The code is available in Python library **geomstats** : computing and statistics on manifolds : `https://github.com/geomstats/geomstats`

Thank you for your attention !