Bayesian Inference on Local Distributions of Functions and Multi-dimensional Curves with Spherical HMC Sampling

Anis Fradi[1,2] and Chafik Samir[1]

[1]CNRS-LIMOS (UMR 6158), UCA, Clermont-Ferrand, France
[2]Faculty of Sciences of Monastir, Monastir, Tunisia

Introduction

What?

- A Bayesian clustering of uni-variate functions and multi-dimensional curves.
- A GMM with an unknown reparamatrization for each cluster to be estimated.

How?

- Reducing the complexity of reparametrization functions when dealing with the Hilbert sphere.
- The spherical HMC sampling for spherical constraint distributions.

Discussion and conclusion

- We proposed a novel Bayesian clustering of uni-variate functions and multidimensional curves.
- The proposed model was tested on multiple simulated and real datasets.
- Several benefits of our proposal compared to the state-of-the-art methods.
- This work remains valid for other models, e.g., curve registration, regression and classification.

Perspectives

- This work can be extended for more complex domains for new aspects of manifold learning, e.g., surfaces.

Publication

- A. Fradi and C. Samir, "Bayesian Cluster Analysis for Registration and Clustering Homogeneous Subgroups in Multidimensional Functional Data", Communication in Statistics, 2020.

Illustration of our framework

- Observations: $\mathbf{N}$ curves $\beta_1, \dots, \beta_N$ with $\beta_i : \mathbf{I} = [0,1] \to \mathbb{R}^d$; $d \geq 1$.
- Goal: Assign each curve to one of $\mathbf{K}$ clusters with $\mathbf{K} << \mathbf{N}$.
- Notations:
  - $\mathbf{T} = (t_1, \dots, t_n)$ denote a discretization of $\mathbf{I}$ and $\beta_i \circ \mathbf{F}(\mathbf{T})$ that of β_i.
  - $\mathbf{F}$ is a reparametrization, identified with a cumulative density function, belonging to

$$\mathcal{F} = \left\{ \mathbf{F} : \mathbf{I} \to \mathbf{I} \,|\, \mathbf{F}(0) = 0,\ \mathbf{F}(1) = 1, \text{ and } \dot{\mathbf{F}} \text{ is nonnegative} \right\}$$

- Remark: The shape space of curves endowed with the elastic metric is nonlinear $\to$ Each β_i is represented by its q-function (SRVF) $\mathbf{q}_i(\mathbf{t}) = \dot{\beta}_i(\mathbf{t}) / \sqrt{||\dot{\beta}_i(\mathbf{t})||_2} \to \beta_i \circ \mathbf{F}$ is then represented by $\mathbf{q}_i^*(\mathbf{t}) = \sqrt{\dot{\mathbf{F}}(\mathbf{t})}\mathbf{q}_i(\mathbf{F}(\mathbf{t}))$.

Problem reformulation and main contributions

- $\mathcal{F}$ is a nonlinear group of diffeomorphisms for the composition operation with no natural metric $\to$ Optimizing a cost function over $\mathbf{F} \in \mathcal{F}$ is computationally intractable.
- $\mathcal{F}$ is isometrically mapped to the Hilbert upper-hemisphere

$$\mathcal{H} = \left\{ \psi \equiv \sqrt{\dot{\mathbf{F}}} \in \mathbb{L}^2(\mathbf{I}) \,|\, \psi \text{ is nonnegative, and } ||\psi||_{\mathbb{L}^2} = 1 \right\}$$

- Although $\mathcal{H}$ is more simple than $\mathcal{F}$, it remains infinite-dimensional $\to$ Finite empirical approximation.
- If $\psi(\mathbf{t}) \sim \mathcal{GP}\big(0, \mathbf{c}(\mathbf{t}, \mathbf{s})\big)$, the truncated Karhunen-Loève expansion of $\mathbf{F}$ is

$$\mathbf{F}_m(\mathbf{t}) = \sum_{j=1}^{m} a_j^2 \int_0^t \phi_j^2(s)ds + 2\sum_{j=1}^{m}\sum_{r=j+1}^{m} a_j a_r \int_0^t \phi_j(s)\phi_r(s)ds \quad \text{with} \quad a_j \overset{ind}{\sim} \mathcal{N}(0, \lambda_j)$$

- Estimating the unknown reparametrization $\mathbf{F}^k$ for $\mathbf{k}$-th cluster, $\mathbf{k} = 1, \dots, \mathbf{K}$, is necessary before clustering.
- Estimating $\mathbf{F}_m^k$ is equivalent to estimating $\mathbf{A}^k = (a_1^k, \dots, a_m^k) \in \mathcal{S}^{m-1}$.

Gaussian mixture model (GMM)

- Assuming that $\mathbf{q}_i^*(\mathbf{T}) | \mathbf{C}_i = \mathbf{k} \sim \mathcal{N}\big(\tilde{\mathbf{q}}^k(\mathbf{T}), \gamma^2\mathcal{I}\big)$ where we draw a class $\mathbf{C}_i$ under $\pi_k = \mathbf{p}(\mathbf{C}_i = \mathbf{k})$.
- The prior on $\mathbf{A}^k$ satisfies

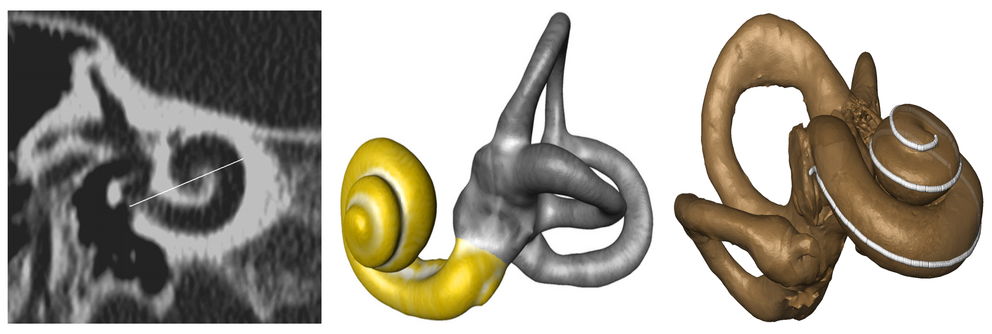
$$\mathbf{p}(\mathbf{A}^k) \propto \exp\Big(-\sum_{j=1}^{m} \frac{a_j^{k^2}}{2\lambda_j}\Big) \times \delta_{\mathbf{A}^k \in \mathcal{S}^{m-1}}$$

- Given $\mathbf{D} = \{\mathbf{q}_i\}_{i=1}^N$, the log-posterior of $\mathbf{A}^1, \dots, \mathbf{A}^K$ is

$$\log \mathbf{p}(\mathbf{A}^1, ..., \mathbf{A}^K | \mathbf{D}, \pi_1, ..., \pi_K, \tilde{\mathbf{q}}^1(\mathbf{T}), ..., \tilde{\mathbf{q}}^K(\mathbf{T}), \gamma^2) \propto \sum_{i=1}^{N} \log\Big(\sum_{k=1}^{K} \pi_k \exp\big(-\frac{1}{2\gamma^2}||\mathbf{q}_i^*(\mathbf{T}) - \tilde{\mathbf{q}}^k(\mathbf{T})||_2^2\big)\Big) - \frac{1}{2}\sum_{k=1}^{K}\sum_{j=1}^{m}\frac{a_j^{k^2}}{\lambda_j}$$

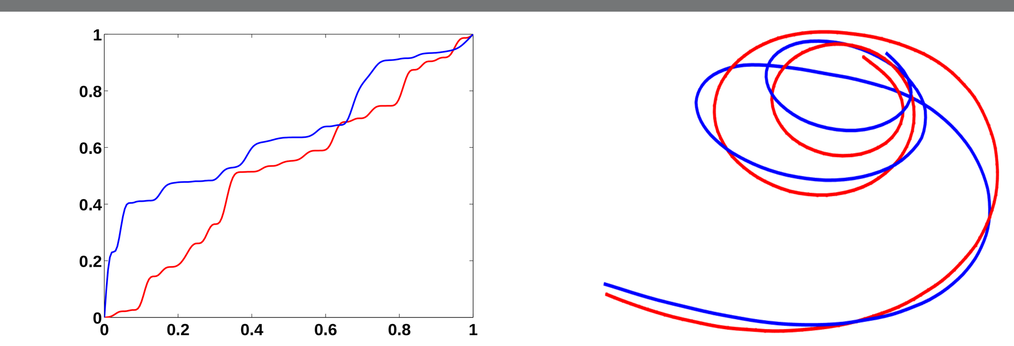
- Spherical HMC for the MAP of $\mathbf{A}^k$ with an extra Gibbs sampling for π_k, $\tilde{\mathbf{q}}^k(\mathbf{T})$ and γ^2.

Different cochlear shapes



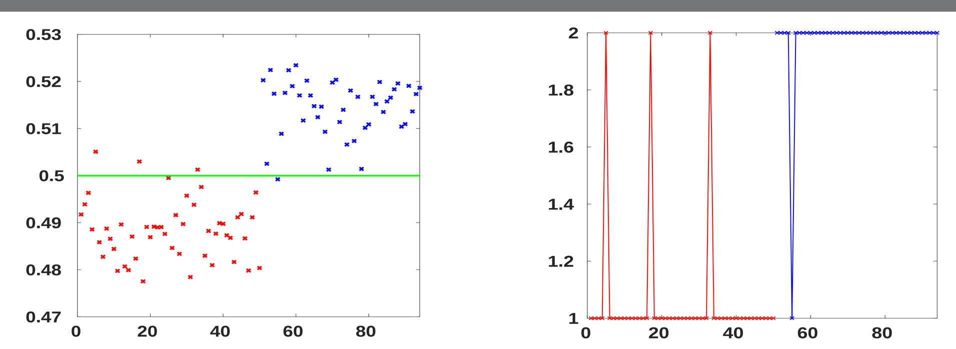
CT scan (left), surface without the curve (middle), and with the extracted curve (right).

female (red) and male (blue)



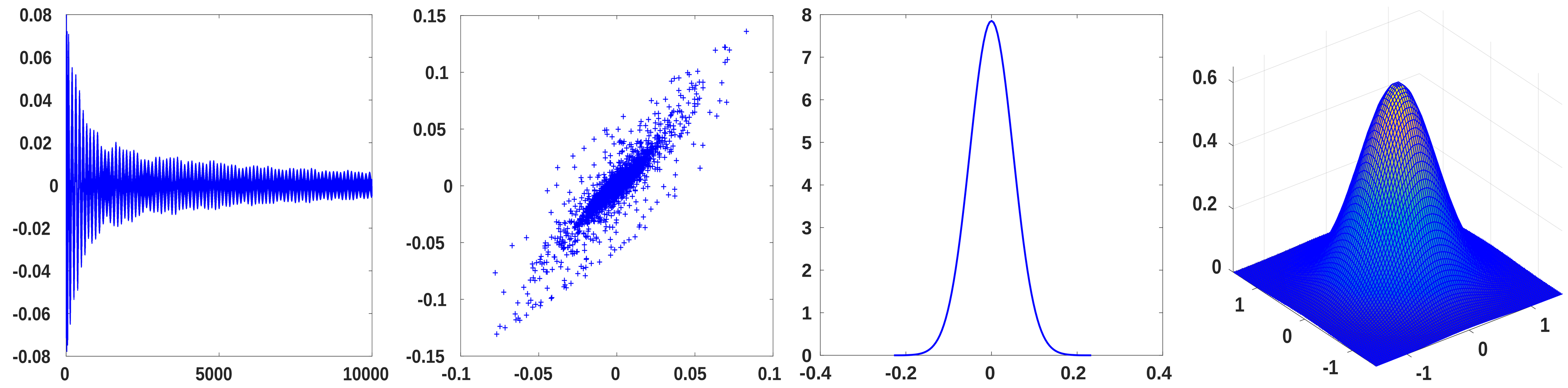
The reparametrization estimates (left) and the Fréchet means (right).

female (red) and male (blue)



The probability that $\mathbf{q}_i$ belongs to the first class (left) and the predicted cluster (right).

Illustration of spherical HMC sampling



Chain of a_1^1 — Chain of (a_1^1, a_2^1) — Posterior of a_1^1 — Posterior of (a_1^1, a_2^1)

Comparison with existing methods

Methods	ER	SP	SE
Euclidean-GMM	41.75%	58.07%	58.5%
Euclidean-kmeans	41.54%	58.44%	58.5%
Geodesic-kmeans	25.11%	74.4%	75.45%
Geodesic-kmedoids	10.85%	89.8%	88.41%
Proposed	4.26%	94%	97.73%