A density-based metric learning approach to geometric inference

E. Borghini, **Ximena Fernández**, P. Groisman & G. Mindlin

Abstract

We address the problem of estimating intrinsic distances in a manifold from a finite sample. We prove that the metric space defined by the sample endowed with a computable metric known as sample Fermat distance converges a.s. in the sense of Gromov–Hausdorff. The limiting object is the manifold itself endowed with the population Fermat distance, an intrinsic metric that accounts for both the geometry of the manifold and the density that produces the sample. We show that this approach outperforms more standard methods based on Euclidean norm, with theoretical results and computational experiments.

Keywords: Manifold Learning, Distance Learning, Persistent Homology.

Fermat Distance

Let $\mathcal{M}$ be a smooth d-dimensional Riemannian manifold embedded in $\mathbb{R}^D$ with density $f : \mathcal{M} \to \mathbb{R}_{>0}$. Let $\mathbb{X}_n = \{x_1, x_2, \ldots, x_n\} \subseteq \mathcal{M}$ be a sample of n independent sample points in $\mathcal{M}$ with common density f.

Let $x, y \in \mathcal{M}$ and $p > 1$.

- The **population Fermat distance** is defined as

$$d_{f,p}(x,y) = \inf_{\gamma} \int_I \frac{1}{f(\gamma_t)^{(p-1)/d}} |\dot{\gamma}_t| dt.$$

where $|\cdot|$ denotes the Euclidean distance and the infimum is taken over all piecewise smooth curves $\gamma\colon I = [0,1] \to \mathcal{M}$, $\gamma(0) = x$, $\gamma(1) = y$.

- The **sample Fermat distance** is defined as

$$d_{\mathbb{X}_n,p}(x,y) = \inf_{\gamma} \sum_{i=0}^{r} |x_{i+1} - x_i|^p$$

where the infimum is taken over all paths $\gamma = (x_0, x_1, \ldots, x_{r+1})$ with $x_0 = x$, $x_{r+1} = y$ and $\{x_1, x_2, \ldots, x_r\} \subseteq \mathbb{X}_n$.

References

[1] E. Borghini, X. Fernández, P. Groisman, and G. Mindlin. Intrinsic persistent homology via density-based metric learning. *arXiv preprint arXiv:2012.07621*, 2020.

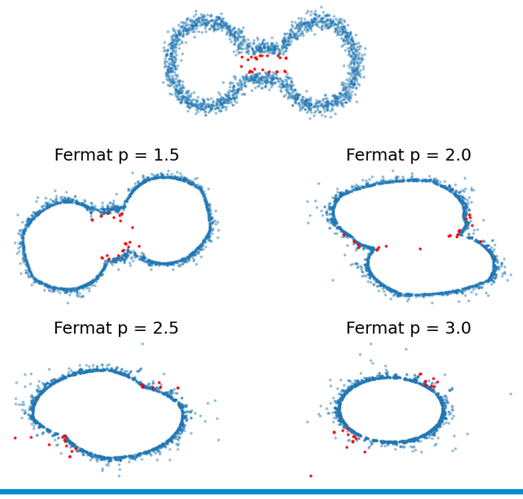
Distance learning

Theorem. *There exists a constant $C = C(n,p,d)$ such that for every $\lambda \in \big((p-1)/pd, 1/d\big)$ and $\varepsilon > 0$ there exist $\theta > 0$ satisfying*

$$\mathbb{P}\left(d_{GH}\left((\mathcal{M}, d_{f,p}), (\mathbb{X}_n, C d_{\mathbb{X}_n,p})\right) > \varepsilon\right) \xrightarrow[n\to\infty]{} 0$$

The choice of p

MDS projection of a point cloud endowed with sample Fermat distance for different choices of p.

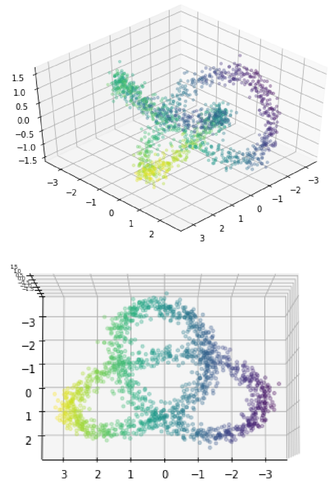


Contact Information

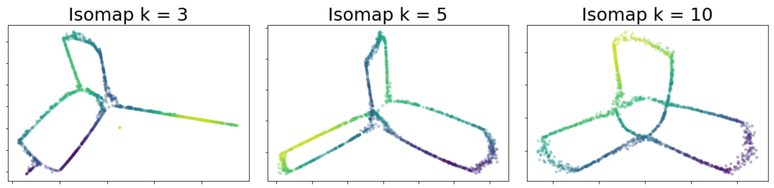
Web ximenafernandez.github.io/
Email x.l.fernandez@swansea.ac.uk

Dimensionality reduction

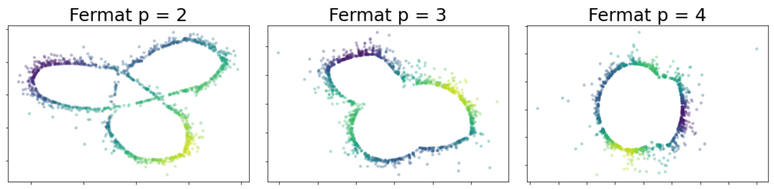
We couple the estimation of Fermat distance on input data with the Multidimensional Scaling method to achieve **dimensionality reduction**.



Isomap suffers from topological instability in presence of noise.

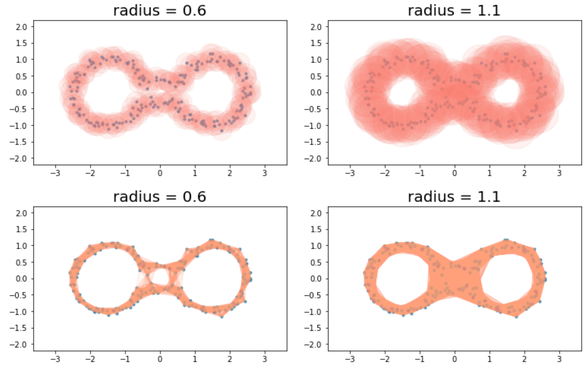


Fermat distance treats noisy points almost as not being part of the manifold, so they do not affect the inference of the right geometry.



Persistent Homology

Persistent homology pipeline. From a *point cloud* with a *metric* construct a filtration of simplicial complexes parametrized by real numbers representing the radius of a covering of balls. Then, compute the persistent generators of homology groups and summarize the information in a *persistence diagram*. Each point in the diagram represents the birth and the lifetime of a generator in homology. H_0 represents connected components (clusters) and H_1, one dimensional cycles.



The choice of the distance. Persistence diagrams strongly depend on the notion of distance defined in the input data. When data belongs to a manifold, the choice of (robust estimators of) intrinsic distances reflects more faithfully the topology of the manifold.